



**TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ**



**ÖNLEYİCİ KOLLUK FAALİYETLERİ
KAPSAMINDA SOSYAL MEDYADA VERİ
MADENCİLİĞİ: SUÇLARI ÖNLEMELİK İÇİN TWİTTER
ANALYTİC'İ KULLANMA**

Emre Cihan ATEŞ

**DİSİPLİNLERARASI ADLİ BİLİMLER ANABİLİM DALI
KRİMİNALİSTİK DOKTORA TEZİ**

DANIŞMAN

Doç. Dr. Gazi Erkan BOSTANCI

ANKARA

2021

**TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ**

**ÖNLEYİCİ KOLLUK FAALİYETLERİ
KAPSAMINDA SOSYAL MEDYADA VERİ
MADENCİLİĞİ: SUÇLARI ÖNLEMELİK İÇİN TWİTTER
ANALYTİC'İ KULLANMA**

Emre Cihan ATEŞ

**DİSİPLİNLERARASI ADLİ BİLİMLER ANABİLİM DALI
KRİMİNALİSTİK DOKTORA TEZİ**

**DANIŞMAN
Doç. Dr. Gazi Erkan BOSTANCI**

**ANKARA
2021**

ETİK BEYAN

Ankara Üniversitesi

Sağlık Bilimleri Enstitüsü Müdürlüğü'ne,

Doktora tezi olarak hazırlayıp sunduğum “Önleyici Kolluk Faaliyetleri Kapsamında Sosyal Medyada Veri Madenciliği: Suçları Önlemek İçin Twitter Analytic'i Kullanma” başlıklı tez; bilimsel ahlak ve değerlere uygun olarak tarafımdan yazılmıştır. Tezimin fikir/hipotezi tümüyle tez danışmanım ve bana aittir. Tezde yer alan deneysel çalışma/araştırma tarafımdan yapılmış olup, tüm cümleler, yorumlar bana aittir.

Öğrencinin Adı Soyadı: Emre Cihan ATEŞ

Tarih:03.01.2021

İmza:

İÇİNDEKİLER

Etik Beyan	ii
İçindekiler	iii
Önsöz	vii
Şekiller	lix
Simgeler ve Kısaltmalar	xxiv
Çizelgeler	xxvii

1. GİRİŞ	1
1.1. Suç Kavramının İncelenmesi	3
1.2. Suçla Mücadelede Reaktif ve Proaktif Yaklaşım	5
1.3. Proaktif Yaklaşım Olarak Öngörüşel Kolluk	8
1.4. Kolluk Modelleri	9
1.4.1. Toplum Destekli Güvenlik Hizmeti	10
1.4.2. Problem Odaklı Kolluk	10
1.4.3. İstihbarat Odaklı Kolluk	11
1.5. Sosyal Medya Kavramı	12
1.5.1. Sosyal Medyanın Suç ve Mağduriyet Üzerine Etkisi	12
1.5.2. Siber Zorbalık ve Sosyal Medya	15
1.5.3. Sosyal Ağ Analizi	16
1.5.4. Sosyal Medya ve Açık Kaynak İstihbaratı	17
1.5.5. Sosyal Medya, Halkla İlişkiler ve Güvenlik Güçleri	18
1.6. Büyük Veri	19
1.6.1. Veri, Enformasyon, Bilgi ve Bilgelik Kavramlarının İncelenmesi	20
1.6.2. Veri Bilimi	23
1.6.3. Büyük Veri	24
1.7. Veri Madenciliği	28
1.7.1. Veri Madenciliği Süreci	32
1.7.2. Veri Madenciliği Görev Türleri	36
1.7.3. Veri Analitiği Türleri	37
1.8. Yapay Zekâ	38
1.8.1. Yapay Zekâ ve Öğrenme	39
1.8.2. Yapay Zekânın Etkileri	40
1.9. Makine Öğrenmesi	41
1.9.1. Regresyon	43
1.9.2. Sınıflandırma	44
1.9.3. Toplanma (Topluluk Öğrenimi)	45
1.9.4. Kümeleme	46
1.9.5. Birliktelik Kuralları Yönetimi	46
1.9.6. Boyut Azaltma	47
1.10. Makine Öğrenmesi Yöntemleri	48
1.10.1. Denetimli Öğrenme	49

1.10.2. Denetimsiz Öğrenme	51
1.10.3. Yarı Denetimli Öğrenme	52
1.10.4. Pekiştirmeli Öğrenme	53
1.11. Makine Öğrenmesi Algoritmaları	53
1.11.1. Karar Ağaçları	56
1.11.2. Rastgele Orman	59
1.11.3. Naive Bayes	60
1.11.4. K-En Yakın Komşu (KNN)	61
1.11.5. K- Ortalamalar Yöntemi	63
1.11.6. Destek Vektör Makineleri (SVM)	64
1.11.7. Lineer Regresyon	65
1.11.8. Logistik Regresyon	69
1.11.9. Apriori Algoritması	70
1.11.10. Temel Bileşen Analizi (PCA)	71
1.11.11. Lineer Diskriminant Analiz (LDA)	73
1.11.12. Bellek Temelli Nedenleme	75
1.11.13. Torbalama	76
1.11.14. Artırma	76
1.11.15. Gradyan Artırma Makineleri	77
1.12. Derin Öğrenme	77
1.12.1. Yapay Sinir Ağları (ANN)	79
1.12.2. Evrişimsel Sinir Ağları (CNN/ConvNet)	81
1.12.3. Tekrarlayan Sinir Ağları (RNN)	82
1.13. Model Geçerleme ve Performans Metrikleri	83
1.13.1. Ortalama Mutlak Hata	84
1.13.2. Ortalama Kare Hatası	85
1.13.3. Kök Ortalama Kare Hatası	85
1.13.4. Normalleştirilmiş Hata	85
1.13.5. Doğruluk ve Hata	86
1.13.6. Kesinlik	87
1.13.7. Anma	87
1.13.8. Özgünlük	87
1.13.9. F-Ölçütü	88
1.13.10. Karışıklık Matrisi	88
1.13.11. Logaritmik Kayıp	89
1.13.12. Kappa	90
1.14. Metin Madenciliği	90
1.14.1. Doğal Dil	92
1.14.2. Makine Dili	94
1.15. Metin Verilerini İşleme ve Anlama	95
1.15.1. URL Etiketleri Kaldırma	95
1.15.2. Metni Bölme	96
1.15.3. Dizgeciklere Ayırma (Tokenleştirme)	97
1.15.4. Karakter Kodlama	98
1.15.4. Metinde Karakter Kodlamanın Tespiti	100
1.15.5. Aksanlı Harfleri Kaldırma / Değiştirme	101
1.15.6. Kullanılan Dilin Tespiti	101

1.15.7. Özel Karakterlerin Kaldırılması	102
1.15.8. Kısaltmaların Genişletilmesi	103
1.15.9. Büyük/Küçük Harf Dönüşümleri	104
1.15.10. Yinelenen Karakterleri Düzeltme	105
1.15.11. Yazım Düzeltme	108
1.15.12. Kök Bulma	110
1.15.13. Kök Çözümleme	112
1.15.14. Durağan Kelimelerin Kaldırılması	112
1.15.15. Sözcük Türü Etiketleme	114
1.15.16. İsimlendirilmiş Varlık Tanıma	119
1.15.17. Söz Dizimsel Analiz (Ayrıştırma)	120
1.16. Metin Temsil Yöntemleri	120
1.16.1. Kelime Torbası (Bag of Words - BoW)	121
1.16.2. N-Gram	122
1.16.3. Terim Frekansı-Ters Doküman Frekansı (TF-IDF)	123
1.16.4. Kelime Vektörleri (Word2Vec)	124
1.16.5. GloVe	127
1.16.6. FastText	127
1.17. Metin Benzerliği	127
1.18. Konu Modelleri	129
1.18.1. Gizli Anlamsal Analiz (LSA)	130
1.18.2. Olasılıksal Gizli Anlamsal Analiz (PLSA)	130
1.18.3. Gizli Dirichlet Tahsisi (LDA)	131
1.19. Tezin Amacı ve Kapsamı	132
2. GEREÇ VE YÖNTEM	136
2.1. Verilere Erişmede Kullanılan Programlama Dili	136
2.2. Veriler Üzerinde Yapılan İşlemlerde Kullanılan Kütüphaneler	138
2.3. Verileri Elde Etme ve Ön İşleme	139
2.3.1. Twitter	140
2.3.2. Veri Elde Etme	142
2.3.3. Veri Ön İşleme	147
2.4. Verilerin Sayısallaştırılması	149
2.5. Verilerin Analizi	149
2.5.1. TDGH Kapsamında Kolluk Kuvvetlerinin Paylaşımlarının İncelenmesi	150
2.5.2. Türkçe Sosyal Ağ Mesajlarında Siber Zorbalığın Tespiti	153
2.5.3. Büyük Veri Üzerinde Keşifsel Veri Analizi ve Açık Kaynak İstihbaratı	157
3. BULGULAR	161
3.1. Türkiye Özelinde Kolluk Teşkilatlarının Paylaşımlarının İncelenmesi	161
3.1.1. Jandarma Genel Komutanlığı'nın Paylaşımlarının İncelenmesi	161
3.1.2. Emniyet Genel Müdürlüğü'nün Paylaşımlarının İncelenmesi	172
3.1.3. Jandarma ve Emniyet Teşkilatlarının Mukayeseli Ortak Analizi	184
3.2. İtalya Özelinde Kolluk Teşkilatlarının Paylaşımlarının İncelenmesi	190
3.2.1. Arma dei Carabinieri'nin Paylaşımlarının İncelenmesi	191
3.2.2. Polizia di Stato'nun Paylaşımlarının İncelenmesi	201

3.2.3. Arma Dei Carabinieri ve Polizia Di Stato Teşkilatlarının Mukayeseli Ortak Analizi	211
3.3. Fransa Özelinde Kolluk Teşkilatlarının Paylaşımlarının İncelenmesi	217
3.3.1. Gendarmerie Nationale'nin Paylaşımlarının İncelenmesi	218
3.3.2. Police Nationale'nin Paylaşımlarının İncelenmesi	228
3.3.3. Gendarmerie Nationale ve Police Nationale Teşkilatlarının Mukayeseli Ortak Analizi	238
3.4. İspanya Özelinde Kolluk Teşkilatlarının Paylaşımlarının İncelenmesi	244
3.4.1. Guardia Civil'in Paylaşımlarının İncelenmesi	245
3.4.2. Policía Nacional'in Paylaşımlarının İncelenmesi	255
3.4.3. Guardia Civil ve Policía Nacional Teşkilatlarının Mukayeseli Ortak Analizi	265
3.5. Kolluk Paylaşımlarının Uluslararası Düzeyde Mukayeseli İncelenmesi	271
3.6. Yarı Denetimli Makine Öğrenimine Dayalı Siber Zorbalık Tespiti	276
3.6.1. Eğitim ve Test Veri Seti Üzerinde Keşifsel Veri Analizi	276
3.6.2. Eğitim Veri Seti Üzerinde Makine Öğrenimi Modelleri	280
3.6.3. Makine Öğrenimi Modellerinin Test Veri Seti Üzerine Uygulanması	307
3.7. Denetimsiz Öğrenme Kapsamında Büyük Veri Üzerinde Keşifsel Veri Analizi ve Açık Kaynak İstihbaratı	309
3.7.1. Keşifsel Veri Analizi	309
3.7.2. Duygu Analizi	312
3.7.3. Ağ Analizi	314
3.7.4. Veri Seti Üzerinde Konu Modelleme	319
3.7.5. Anahtar Kelime Analizi ve Veriyle İlişkisi	321
3.7.6. Sosyal Medya Gündemi ile Haber Ajansı Paylaşımlarının Kıyaslaması	323
4. TARTIŞMA	325
4.1. TDGH Kapsamında Kolluk Kuvvetlerinin Paylaşımlarının İncelenmesi	325
4.2. Türkçe Sosyal Ağ Mesajlarında Siber Zorbalığın Tespiti	331
4.3. Büyük Veri Üzerinde Keşifsel Veri Analizi Ve Açık Kaynak İstihbaratı	335
5. SONUÇ VE ÖNERİLER	341
ÖZET	344
SUMMARY	345
KAYNAKLAR	347
ÖZGEÇMİŞ	372

ÖNSÖZ

Suç evrenseldir, her daim var olacaktır ve doğası gereği insanlara çeşitli zararlar verecektir. Söz konusu zararların ne kadar azaltılabileceği kaygısıyla, bu çalışma suçla mücadelede kolluğun yapabileceklerine dair bazı yöntemlerin bütüncül yaklaşımını içermektedir. Bu çalışmada, dijitalleşmeye bağlı olarak, suç oranlarının da giderek arttığı günümüzde, suç ve suçla mücadele edebilmek adına Twitter üzerinden yapılan paylaşımlar veri madenciliği teknikleri ile analiz edilmiştir. Elde edilen verilerin makine öğrenimi yardımıyla, güvenlik güçleri tarafından suçu önlemeye yönelik yapılacak proaktif faaliyetlerde kullanılabilirlik durumunun belirlenmesi, içinde suç ihtiva eden paylaşımın önceden belirlenebilmesi ve ayrıca ulusal/uluslararası alanda faaliyet gösteren güvenlik güçlerinin paylaşımlarının toplum destekli güvenlik hizmeti açısından analizinin yapılması amaçlanmıştır.

Alandaki bilgi ve tecrübesinin yanı sıra, öğrencilerine her zaman daha fazla katkı sağlamak için büyük bir özveriyle çalışan, doktora eğitimim boyunca çalıştığım konularda bana kısıt koymaksızın özgürce çalışma imkânı tanıyan ve tez kapsamında desteklerini ve motivasyonunu hiç esirgemeyen değerli hocam Doç.Dr. Gazi Erkan BOSTANCI'ya,

Akademik ve alan bilgisi ile öğrencilerine rol model olmasının yanında, oldukça verimli ve keyifli dersler işleyerek, kriminalistik/adli tıp anlamında birçok konuyu bana öğreten ve ondan öğrendiklerimi dilimin döndüğünce kendi öğrencilerimde de anlatarak olumlu geri dönüşler almamı sağlayan kriminalistik/adli tıp alanında beni yetiştiren değerli hocam Prof.Dr. Nergis CANTÜRK'e,

Jandarma ve Sahil Güvenlik Akademisi (JSGA) Başkanlığındaki çalışma arkadaşlarıma ve komutanlarıma, Seneca'nın "Homines dum docent discunt" ("İnsanlar, öğretirken öğrenirler") dediği üzere, onlara birşeyler öğretirken öğrendiğim, önemli bir motivasyon kaynağım olan ve herbiri güzel ülkemizin dört bir yanında Jandarma personeli olarak görev yapan subay/astsubay öğrencilerime,

Beni suç kelimesiyle tanıştıran, fikirleriyle ve bilgisiyle suç olgusuna ve algısına bakış açımı değiştiren, mükemmel analiz yeteneğiyle hocam olarak atfedeceğim ve halihazırda Diyarbakır İl Jandarma K.lığı Siber Suçlarla Mücadele Şube Müdürü olan J.Asb.Kd.Bçvş. Engin CAN'a ve tezi bitirmem konusunda beni inanılmaz derecede motive ederek bazı olaylara farklı bakabilmeyi öğrendiğim Gazi Bnb.Dr. Mehmet Bedri ALUÇLU'ya,

2007 yılında sosyal medya hesabı kullandığım için aldığım (3) günlük oda hapsine müteakip, sosyal medya hesabı kullanmayı uzun süre bırakmışken, 2014-2017 yılları arasında sosyal medyanın önemini bana tekrar öğreten, Twitter ve teknik takibin eş zamanlı yürütülmesi, terörün finansman tespiti, Sur Hendek Operasyonları gibi bütün faaliyetlerde yanımda olan ve desteklerini hiç esirgemeyen Diyarbakır İl Jandarma Komutanlığı TEM Ş.Md.lüğündeki fedakar çalışma arkadaşlarıma,

İyi bir insan olmayı, fedakârlığı, ailenin öneminin ve desteğinin ne kadar önemli olduğunu bana öğreten ve desteklerini her zaman arkamda hissettiğim babam ve anneme, neşeli sohbetleriyle bana keyif veren, sırdaşlık yapan, kardeşliği öğreten ve onlara sahip olmaktan her zaman çok mutlu olduğum Buğra ve Berra'ya,

Hayatımın her aşamasında bana destek veren, bahanelerime katlanan, sürekli olarak öğrenmeyi bana sevdiren ve sevmeyi öğreten eşim Atiye Pınar ATEŞ'e, bize her türlü destek vererek fedakârlığı öğreten ve oğluma ikinci anne olan kayınvalideme,

Evimin neşesi, motivasyon kaynağım, oyun arkadaşım, bana gülmeyi ve tekrar çocuk olmayı öğreten biricik oğlum Emir Bilgen'e,

Hayatım boyunca bana birşeyler öğreten herkese çok teşekkür ederim. Tolstoy'un "İnsan ne ile yaşar?" kitabında dediği üzere; "Anladım ki insan içinde sevgiyle yaşar." Bu sebeple; her deneyim, karşılaşma, paylaşım ve duygu benim için çok kıymetli. İyi ki varsınız.

Emre Cihan ATEŞ
Ankara, 2021

viii

ŞEKİLLER

Şekil 1.1.	Suç olgusunun tarihi seyri	4
Şekil 1.2.	Suçla ilişkili bazı bilim dalları	4
Şekil 1.3.	Suçla mücadelede temel yaklaşımlar	6
Şekil 1.4.	Suç Üçgeni	6
Şekil 1.5.	İstihbarat odaklı kolluğun modellemesi	11
Şekil 1.6.	Sosyal ağlarda karşılaşılan suçlar	14
Şekil 1.7.	Verinin bilgelige dönüşümü	20
Şekil 1.8.	Verinin bilgelikle beklenen ilişkisi	22
Şekil 1.9.	Veri biliminde Conway'in venn diyagramı	24
Şekil 1.10.	2016-2018 yılları arası internetteki bazı etkinlikler (1 dakika)	25
Şekil 1.11.	Büyük veriyi tanımlayan 5V kuralı	27
Şekil 1.12.	Veri madenciliği ve ilişkili olduğu disiplinler	29
Şekil 1.13.	Veri tabanlarında bilgi keşfi (KDD)	31
Şekil 1.14.	Veri tabanlarında bilgi keşfinde CRISP-DM modeli	33
Şekil 1.15.	Veri tabanlarında bilgi keşfinde SEMMA modeli	34
Şekil 1.16.	Veri analitiği türleri	37
Şekil 1.17.	Makine öğreniminin genel çerçevesi	41
Şekil 1.18.	Regresyon metriklerinin formülasyonu	44
Şekil 1.19.	Topluluk öğrenmesinde model yapısı	45
Şekil 1.20.	Örnek sepet analizi	47
Şekil 1.21.	Makine öğrenmesi yöntemleri	48
Şekil 1.22.	Denetimli öğrenmede eğitim seti ve model	50
Şekil 1.23.	Denetimsiz öğrenmede girdi ve çıktı	51
Şekil 1.24.	Denetimsiz öğrenmede eğitim seti ve model	52
Şekil 1.25.	Yarı denetimli öğrenmede eğitim seti ve model	52
Şekil 1.26.	Makine öğrenmesinde hatalar	54
Şekil 1.27.	Karar ağaçlarının yapısı	56
Şekil 1.28.	Sınıf ve entropi formülasyonu	57
Şekil 1.29.	Karar ağacında girdi alanı	57

Şekil 1.30.	Farklı etiket yapılarını barındıran karar ağacı modellemesi	58
Şekil 1.31.	Naive Bayes formülasyonu	60
Şekil 1.32.	Öklid mesafesi formülasyonu	62
Şekil 1.33.	KNN’de öklit mesafesi	62
Şekil 1.34.	K-Ortalamlar algoritmasına göre sınıflandırma	64
Şekil 1.35.	Destek vektör makinelerine göre sınıflandırma	65
Şekil 1.36.	Lineer regresyon modellemesi	66
Şekil 1.37.	Lineer regresyon formülasyonu	66
Şekil 1.38.	Lineer regresyonda residual değerleri	67
Şekil 1.39.	Residuals formülasyonu	68
Şekil 1.40.	Polinomsal lineer regresyon	68
Şekil 1.41.	Polinomsal lineer regresyon formülasyonu	69
Şekil 1.42.	Logit model formülasyonu	69
Şekil 1.43.	Sigmoid eğrisi üzerinden olasılık	70
Şekil 1.44.	Google arama motorunda apriori örneği	71
Şekil 1.45.	Örnek PCA analizi	72
Şekil 1.46.	Gözlemlenen değerlerin iki ekseninde (X_1 / X_2) gösterimi	73
Şekil 1.47.	Gözlemlenen değerlerin yeni ekseninde toplanması	74
Şekil 1.48.	Gözlemlenen değerlerin yeni ekseninde gösterimi	74
Şekil 1.49.	Vektörel düzlemde örnek LDA analizi (k:10)	75
Şekil 1.50.	Yapay zekâ, makine ve derin öğrenme ilişkisi	78
Şekil 1.51.	Nöronlar yardımıyla öğrenme modellemesi	79
Şekil 1.52.	Nöronlar yardımıyla öğrenmenin matematiksel gösterimi	80
Şekil 1.53.	Üç katmanlı örnek yapay sinir ağı	81
Şekil 1.54.	Örnek RNN modeli	82
Şekil 1.55.	Eğitim ve test verisi kullanımı	83
Şekil 1.56.	K-kat çapraz doğrulama (Örnek k:4)	84
Şekil 1.57.	Ortalama mutlak hata formülasyonu	84
Şekil 1.58.	Ortalama kare hatası formülasyonu	85
Şekil 1.59.	Kök ortalama kare hatası formülasyonu	85
Şekil 1.60.	Gerçek ve tahmin değerleri matrisi (2*2)	86
Şekil 1.61.	Doğruluk ve hata formülasyonu	86

Şekil 1.62.	Kesinlik formülasyonu	87
Şekil 1.63.	Anma formülasyonu	87
Şekil 1.64.	Özgünlük formülasyonu	87
Şekil 1.65.	F ölçütü formülasyonu	88
Şekil 1.66.	Örnek gerçek ve tahmin değerleri matrisi (3*3)	88
Şekil 1.67.	Log Loss ve tahmin ilişkisi	89
Şekil 1.68.	Kappa formülasyonu	90
Şekil 1.69.	Metin madenciliğinde kullanılan diller	92
Şekil 1.70.	İletişimde sembolize etme ve sembolleri çözme	93
Şekil 1.71.	Doğal dilin makine diline dönüşümü	94
Şekil 1.72.	URL etiketi kaldırılması örneği	96
Şekil 1.73.	Metin bölme örneği	96
Şekil 1.74.	Tokenleştirme örneği	97
Şekil 1.75.	Karakter tespit ve çözümleme	100
Şekil 1.76.	Aksanlı karakterlere işlem	101
Şekil 1.77.	Kullanılan dilin tespiti	102
Şekil 1.78.	Türkçe örnek metinde özel karakterlerin kaldırılması	103
Şekil 1.79.	Kısaltma karakterlerin genişletilmesi	104
Şekil 1.80.	Büyük/küçük harf dönüşümleri	105
Şekil 1.81.	Tekrar eden karakterlerin doğru düzeltilmesi	106
Şekil 1.82.	Tekrar eden karakterlerin yanlış düzeltilmesi	106
Şekil 1.83.	Tekrar eden karakterlerin Wordnet üzerinden düzeltilmesi	107
Şekil 1.84.	Yazım düzeltme ve alternatif değerler	109
Şekil 1.85.	Türkçe cümle düzeltme	109
Şekil 1.86.	İngilizce cümle düzeltme	110
Şekil 1.87.	Türkçe kök bulma	111
Şekil 1.88.	İngilizce kök bulma	111
Şekil 1.89.	Kök çözümleme ve kök bulma	112
Şekil 1.90.	Türkçe durağan kelimeler	113
Şekil 1.91.	İngilizce durağan kelimeler	113
Şekil 1.92.	Durağan kelimelerin kaldırılması	114
Şekil 1.93.	Türkçe POS etiketleme	115

Şekil 1.94.	İngilizce POS etiketleme	116
Şekil 1.95.	Örnek cümlelerin ilişkisel analizi	116
Şekil 1.96.	Farklı yöntemlerle POS etiketleme ve doğruluk oranları	118
Şekil 1.97.	İsimlendirilmiş varlık tanıma	119
Şekil 1.98.	Kelime torbası gösterimi	121
Şekil 1.99.	Bigram gösterimine ilişkin örnek	122
Şekil 1.100.	TF-IDF formülasyonu	123
Şekil 1.101.	TF formülasyonu	123
Şekil 1.102.	IDF formülasyonu	124
Şekil 1.103.	TF-IDF gösterimine ilişkin örnek	124
Şekil 1.104.	Skip-Gram ve CBOW çalışma şekli	125
Şekil 1.105.	Kelimenin vektörel düzeyde gösterimi	125
Şekil 1.106.	En benzer kelimeler	126
Şekil 1.107.	En benzer kelimelerin grafik gösterimi	126
Şekil 1.108.	Metin benzerliği yöntemleri	128
Şekil 1.109.	LDA konu modellemesi ve token yapısı	131
Şekil 2.1.	Kaggle üzerinden makine öğreniminde kullanılan diller	136
Şekil 2.2.	2019 yılında veri bilimi ile ilgili kullanılan dil ve yazılımlar	137
Şekil 2.3.	Örnek kişisel twitter sayfası girişi	140
Şekil 2.4.	Örnek twitter paylaşımlarının genel görünümü	141
Şekil 2.5.	Örnek twitter konum bazlı popüler etiketler	142
Şekil 2.6.	Twitter Developer API sayfası genel görünümü	143
Şekil 2.7.	Twitter API token ve şifre edinimi	143
Şekil 2.8.	Twitter'dan çekilebilecek veriler	144
Şekil 2.9.	Örneklem twitter mesajında yer alan bilgiler	146
Şekil 2.10.	TDGH kapsamında kolluk kuvvetlerinin paylaşımlarının işlem prosesi	151
Şekil 2.11.	Makine öğrenimine dayalı siber zorbalık tespiti işlem prosesi	154
Şekil 2.12.	Büyük veri üzerinde keşifsel veri analizi işlem prosesi	158
Şekil 3.1.	Jandarma Genel Komutanlığının resmi Twitter sayfasının ekran resmi	162
Şekil 3.2.	@jandarma kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-I	162

Şekil 3.3.	@jandarma kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler–II	162
Şekil 3.4.	@jandarma kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler–III	163
Şekil 3.5.	@jandarma adlı kullanıcının veri paylaşımında kullandığı kaynaklar	163
Şekil 3.6.	Paylaşımların beğenilme ve retweet durumunun incelenmesi	164
Şekil 3.7.	@jandarma adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu	165
Şekil 3.8.	Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı	165
Şekil 3.9.	Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı	166
Şekil 3.10.	En fazla kullanılan sosyal medya etiketlerinin (Hashtag - #) kelime bulutu	166
Şekil 3.11.	En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)	167
Şekil 3.12.	En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)	168
Şekil 3.13.	Metnin duygusal durumunun radar grafiği	168
Şekil 3.14.	Metnin duygusal durumunun öznellikle analizi	169
Şekil 3.15.	Metnin duygusal durumunun saat bazlı incelenmesi	170
Şekil 3.16.	Paylaşımlar arasında sık kullanılan ifadelere ait kelime bulutu	170
Şekil 3.17.	Yapılan paylaşımlara ait kelime ağı analizi (100>Kelime Kombinasyonu)	171
Şekil 3.18.	Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması	171
Şekil 3.19.	Test verisinin duygusal matrisi	172
Şekil 3.20.	Emniyet Genel Müdürlüğünün resmi Twitter sayfasının ekran resmi	173
Şekil 3.21.	@EmniyetGM kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler–I	173
Şekil 3.22.	@EmniyetGM kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler–II	173
Şekil 3.23.	@EmniyetGM kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler–III	174
Şekil 3.24.	@EmniyetGM adlı kullanıcının veri paylaşımında kullandığı kaynaklar	174
Şekil 3.25.	Paylaşımların beğenilme ve retweet durumunun incelenmesi	175
Şekil 3.26.	@EmniyetGM adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu	176

Şekil 3.27.	Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı	176
Şekil 3.28.	Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı	177
Şekil 3.29.	En fazla kullanılan sosyal medya etiketlerinin (Hashtag - #) kelime bulutu	177
Şekil 3.30.	En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)	178
Şekil 3.31.	En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)	179
Şekil 3.32.	Metnin duygusal durumunun radar grafiği	179
Şekil 3.33.	Metnin duygusal durumunun öznellikle analizi	180
Şekil 3.34.	Metnin duygusal durumunun saat bazlı incelenmesi	181
Şekil 3.35.	Paylaşımlar arasında sık kullanılan ifadelere ait kelime bulutu	181
Şekil 3.36.	Yapılan paylaşımlara ait kelime ağı analizi (100>Kelime Kombinasyonu)	182
Şekil 3.37.	Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması	182
Şekil 3.38.	Test verisinin duygusal matriksi	183
Şekil 3.39.	Duygulara göre retweet ve beğenme durumu	184
Şekil 3.40.	Duygulara göre ortalama retweet ve beğenme durumu	185
Şekil 3.41.	Duygusal korelasyon matriksi	185
Şekil 3.42.	Jandarma ve Emniyet'in etki gücü	186
Şekil 3.43.	Yapılan paylaşımların benzerlik oranı (Kosinüs)	190
Şekil 3.44.	Arma dei Carabinieri resmi Twitter sayfasının ekran resmi	191
Şekil 3.45.	"@_Carabinieri_" kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-I	191
Şekil 3.46.	"@_Carabinieri_" kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-II	192
Şekil 3.47.	"@_Carabinieri_" kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-III	192
Şekil 3.48.	"@_Carabinieri_" adlı kullanıcının veri paylaşımında kullandığı kaynaklar	193
Şekil 3.49.	Paylaşımların beğenilme ve retweet durumunun incelenmesi	193
Şekil 3.50.	"@_Carabinieri_" adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu	194
Şekil 3.51.	Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı	195
Şekil 3.52.	Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı	195

Şekil 3.53.	En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)	196
Şekil 3.54.	En fazla kullanılan sosyal medya etiketlerinin (Hashtag - #) kelime bulutu	196
Şekil 3.55.	En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)	197
Şekil 3.56.	Metnin duygusal durumunun radar grafiği	198
Şekil 3.57.	Metnin duygusal durumunun öznellikle analizi	198
Şekil 3.58.	Metnin duygusal durumunun saat bazlı incelenmesi	199
Şekil 3.59.	Paylaşımlar arasında sık kullanılan ifadelerle ait kelime bulutu	199
Şekil 3.60.	Yapılan paylaşımlara ait kelime ağı analizi (50>Kelime Kombinasyonu)	200
Şekil 3.61.	Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması	200
Şekil 3.62.	Test verisinin duygusal matriksi	201
Şekil 3.63.	Polizia di Stato resmi Twitter sayfasının ekran resmi	202
Şekil 3.64.	@poliziadistato kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-I	202
Şekil 3.65.	@poliziadistato kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-II	202
Şekil 3.66.	@poliziadistato kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-III	203
Şekil 3.67.	@poliziadistato adlı kullanıcının veri paylaşımında kullandığı kaynaklar	203
Şekil 3.68.	Paylaşımların beğenilme ve retweet durumunun incelenmesi	204
Şekil 3.69.	@poliziadistato adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu	205
Şekil 3.70.	Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı	205
Şekil 3.71.	Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı	206
Şekil 3.72.	En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)	206
Şekil 3.73.	En fazla kullanılan sosyal medya etiketlerinin (Hashtag - #) kelime bulutu	207
Şekil 3.74.	En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)	207
Şekil 3.75.	Metnin duygusal durumunun radar grafiği	208
Şekil 3.76.	Metnin duygusal durumunun öznellikle analizi	208
Şekil 3.77.	Metnin duygusal durumunun saat bazlı incelenmesi	209
Şekil 3.78.	Paylaşımlar arasında sık kullanılan ifadelerle ait kelime bulutu	209

Şekil 3.79.	Yapılan paylaşımlara ait kelime ağı analizi (50>Kelime Kombinasyonu)	210
Şekil 3.80.	Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması	210
Şekil 3.81.	Test verisinin duygusal matriksi	211
Şekil 3.82.	Duygulara göre retweet ve beğenme durumu	212
Şekil 3.83.	Duygulara göre ortalama retweet ve beğenme durumu	212
Şekil 3.84.	Duygusal korelasyon matriksi	213
Şekil 3.85.	Carabinieri ve Polizia di Stato'nun etki gücü	214
Şekil 3.86.	Yapılan paylaşımların benzerlik oranı (Kosinüs)	217
Şekil 3.87.	Gendarmerie Nationale resmi Twitter sayfasının ekran resmi	218
Şekil 3.88.	@Gendarmerie kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-I	218
Şekil 3.89.	@Gendarmerie kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-II	219
Şekil 3.90.	@Gendarmerie kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-III	219
Şekil 3.91.	@Gendarmerie adlı kullanıcının veri paylaşımında kullandığı kaynaklar	220
Şekil 3.92.	Paylaşımların beğenilme ve retweet durumunun incelenmesi	220
Şekil 3.93.	@Gendarmerie adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu	221
Şekil 3.94.	Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı	222
Şekil 3.95.	Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı	222
Şekil 3.96.	En fazla kullanılan sosyal medya etiketlerinin (Hashtag - #) kelime bulutu	223
Şekil 3.97.	En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)	223
Şekil 3.98.	En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)	224
Şekil 3.99.	Metnin duygusal durumunun radar grafiği	225
Şekil 3.100.	Metnin duygusal durumunun öznellikle analizi	225
Şekil 3.101.	Metnin duygusal durumunun saat bazlı incelenmesi	226
Şekil 3.102.	Paylaşımlar arasında sık kullanılan ifadelerle ait kelime bulutu	226
Şekil 3.103.	Yapılan paylaşımlara ait kelime ağı analizi (25>Kelime Kombinasyonu)	227
Şekil 3.104.	Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması	227

Şekil 3.105.	Test verisinin duygusal matriksi	228
Şekil 3.106.	Police Nationale resmi Twitter sayfasının ekran resmi	229
Şekil 3.107.	@PoliceNationale kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler–I	229
Şekil 3.108.	@PoliceNationale kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler–II	229
Şekil 3.109.	@PoliceNationale kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler–III	230
Şekil 3.110.	@PoliceNationale adlı kullanıcının veri paylaşımında kullandığı kaynaklar	230
Şekil 3.111.	Paylaşımların beğenilme ve retweet durumunun incelenmesi	231
Şekil 3.112.	@PoliceNationale adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu	232
Şekil 3.113.	Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı	232
Şekil 3.114.	Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı	233
Şekil 3.115.	En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)	233
Şekil 3.116.	En fazla kullanılan sosyal medya etiketlerinin (Hashtag - #) kelime bulutu	234
Şekil 3.117.	En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)	234
Şekil 3.118.	Metnin duygusal durumunun radar grafiği	235
Şekil 3.119.	Metnin duygusal durumunun öznellikle analizi	235
Şekil 3.120.	Metnin duygusal durumunun saat bazlı incelenmesi	236
Şekil 3.121.	Paylaşımlar arasında sık kullanılan ifadelerle ait kelime bulutu	236
Şekil 3.122.	Yapılan paylaşımlara ait kelime ağı analizi (50>Kelime Kombinasyonu)	237
Şekil 3.123.	Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması	237
Şekil 3.124.	Test verisinin duygusal matriksi	238
Şekil 3.125.	Duygulara göre retweet ve beğenme durumu	239
Şekil 3.126.	Duygulara göre ortalama retweet ve beğenme durumu	239
Şekil 3.127.	Duygusal korelasyon matriksi	240
Şekil 3.128.	Gendarmerie ve Police'nin etki gücü	241
Şekil 3.129.	Yapılan paylaşımların benzerlik oranı (Kosinüs)	244
Şekil 3.130.	Guardia Civil resmi Twitter sayfasının ekran resmi	245
Şekil 3.131.	@guardiacivil kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler–I	245

Şekil 3.132.	@guardiacivil kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-II	245
Şekil 3.133.	@guardiacivil kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-III	246
Şekil 3.134.	@guardiacivil adlı kullanıcının veri paylaşımında kullandığı kaynaklar	246
Şekil 3.135.	Paylaşımların beğenilme ve retweet durumunun incelenmesi	247
Şekil 3.136.	@guardiacivil adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu	248
Şekil 3.137.	Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı	248
Şekil 3.138.	Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı	249
Şekil 3.139.	En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)	249
Şekil 3.140.	En fazla kullanılan sosyal medya etiketlerinin (Hashtag - #) kelime bulutu	250
Şekil 3.141.	En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)	251
Şekil 3.142.	Metnin duygusal durumunun radar grafiği	251
Şekil 3.143.	Metnin duygusal durumunun öznellikle analizi	252
Şekil 3.144.	Metnin duygusal durumunun saat bazlı incelenmesi	253
Şekil 3.145.	Paylaşımlar arasında sık kullanılan ifadelere ait kelime bulutu	253
Şekil 3.146.	Yapılan paylaşımlara ait kelime ağı analizi (50>Kelime Kombinasyonu)	254
Şekil 3.147.	Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması	254
Şekil 3.148.	Test verisinin duygusal matriksi	255
Şekil 3.149.	Policía Nacional resmi Twitter sayfasının ekran resmi	256
Şekil 3.150.	@policia kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-I	256
Şekil 3.151.	@policia kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-II	256
Şekil 3.152.	@policia kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-III	256
Şekil 3.153.	@policia adlı kullanıcının veri paylaşımında kullandığı kaynaklar	257
Şekil 3.154.	Paylaşımların beğenilme ve retweet durumunun incelenmesi	258
Şekil 3.155.	@policia adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu	259
Şekil 3.156.	Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı	259

Şekil 3.157.	Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı	260
Şekil 3.158.	En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)	260
Şekil 3.159.	En fazla kullanılan sosyal medya etiketlerinin (Hashtag - #) kelime bulutu	261
Şekil 3.160.	En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)	261
Şekil 3.161.	Metnin duygusal durumunun radar grafiği	262
Şekil 3.162.	Metnin duygusal durumunun öznellikle analizi	262
Şekil 3.163.	Metnin duygusal durumunun saat bazlı incelenmesi	263
Şekil 3.164.	Paylaşımlar arasında sık kullanılan ifadelerle ait kelime bulutu	263
Şekil 3.165.	Yapılan paylaşımlara ait kelime ağı analizi (25>Kelime Kombinasyonu)	264
Şekil 3.166.	Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması	264
Şekil 3.167.	Test verisinin duygusal matriksi	265
Şekil 3.168.	Duygulara göre retweet ve beğenme durumu	266
Şekil 3.169.	Duygulara göre ortalama retweet ve beğenme durumu	266
Şekil 3.170.	Duygusal korelasyon matriksi	267
Şekil 3.171.	Guardia Civil ve Policía Nacional'un etki gücü	268
Şekil 3.172.	Yapılan paylaşımların benzerlik oranı (Kosinüs)	271
Şekil 3.173.	Kolluk teşkilatlarının Twitter kullanım sıklığı	272
Şekil 3.174.	Nüfus ve takipçi sayılarının orantısı (*1000)	272
Şekil 3.175.	Ortalama duygu durum puanları (Kelime bazlı cümle skoru üzerinden)	273
Şekil 3.176.	Kolluk teşkilatlarının etki gücü (Takipçi sayısı orantılı)	273
Şekil 3.177.	Kolluk kuvvetleri paylaşım benzerlik matrisi (TF-IDF)	274
Şekil 3.178.	Tespitte kullanılan veri setlerinin etiket bazında dağılımı	276
Şekil 3.179.	Veri setinde yer alan boş değerler	277
Şekil 3.180.	Tweetler içindeki karakter sayıları	278
Şekil 3.181.	Tweetler içindeki kelime sayıları	278
Şekil 3.182.	Her bir tweet içindeki kelime uzunluğu	278
Şekil 3.183.	Tweetler içinde en sık kullanılan kelimeler (n-gram, n:1)	279
Şekil 3.184.	Tweetler içinde en sık kullanılan iki kelimeler (n-gram, n:2)	279
Şekil 3.185.	Eğitim setinin test ve eğitim olarak bölünmesi	281
Şekil 3.186.	Gaussian naive bayes ile elde edilen değerler	281

Şekil 3.187.	MultinomialNB ile elde edilen değerler	282
Şekil 3.188.	BernoulliNB ile elde edilen değerler	282
Şekil 3.189.	Karar ağacında max depths ile AUC skoru	283
Şekil 3.190.	Karar ağacında min samples split ile AUC skoru	283
Şekil 3.191.	Karar ağacında min samples leafs ile AUC skoru	283
Şekil 3.192.	Karar ağacında n_features ile AUC skoru	283
Şekil 3.193.	Karar ağacı ile elde edilen değerler	284
Şekil 3.194.	Ekstra ağaçlarda n_estimators ile AUC skoru	284
Şekil 3.195.	Ekstra ağaçlarda n_features ile AUC skoru	285
Şekil 3.196.	Ekstra ağaçlarda min samples split ile AUC skoru	285
Şekil 3.197.	Ekstra ağaçlar ile elde edilen değerler	285
Şekil 3.198.	Lineer diskriminant analizde components ile AUC skoru	286
Şekil 3.199.	Lineer diskriminant analiz ile elde edilen değerler	286
Şekil 3.200.	Karesel diskriminant analiz ile elde edilen değerler	287
Şekil 3.201.	Adaboost sınıflandırıcıda n_estimators ile AUC skoru	287
Şekil 3.202.	Adaboost sınıflandırıcıda öğrenme oranı (learning rate) ile AUC skoru	288
Şekil 3.203.	Adaboost ile elde edilen değerler	288
Şekil 3.204.	Gradyan artırmada öğrenme oranı (learning rate) ile AUC skoru	289
Şekil 3.205.	Gradyan artırmada n_estimators ile AUC skoru	289
Şekil 3.206.	Gradyan artırmada derinlik (depth) ile AUC skoru	289
Şekil 3.207.	Gradyan artırmada min samples split ile AUC skoru	289
Şekil 3.208.	Gradyan artırmada min samples leaf ile AUC skoru	290
Şekil 3.209.	Gradyan artırmada n_features ile AUC skoru	290
Şekil 3.210.	Gradyan artırma ile elde edilen değerler	290
Şekil 3.211.	Rastgele ormanda n_estimators ile AUC skoru	291
Şekil 3.212.	Rastgele ormanda derinlik (depth) ile AUC skoru	291
Şekil 3.213.	Rastgele ormanda min samples split ile AUC skoru	292
Şekil 3.214.	Rastgele ormanda min samples leaf ile AUC skoru	292
Şekil 3.215.	Rastgele orman ile elde edilen değerler	292
Şekil 3.216.	Lojistik regresyonda max iter ile AUC skoru	293
Şekil 3.217.	Lojistik regresyonda components ile AUC skoru	293

Şekil 3.218.	Lojistik regresyonda penalty ile AUC skoru	293
Şekil 3.219.	Lojistik regresyon ile elde edilen değerler	294
Şekil 3.220.	Perceptron modelinde max iter ile AUC skoru	294
Şekil 3.221.	Perceptron modelinde penalty ile AUC skoru	295
Şekil 3.222.	Perceptron modelinde alpha ile AUC skoru	295
Şekil 3.223.	Perceptron regresyon ile elde edilen değerler	295
Şekil 3.224.	Lineer SVC modelinde components ile AUC skoru	296
Şekil 3.225.	Lineer SVC modelinde max iter ile AUC skoru	296
Şekil 3.226.	Lineer SVC regresyon ile elde edilen değerler	296
Şekil 3.227.	XGBoost modelinde derinlik (max depth) ile AUC skoru	297
Şekil 3.228.	XGBoost modelinde öğrenme oranı (learning rate) ile AUC skoru	297
Şekil 3.229.	XGBoost modelinde n estimators ile AUC skoru	297
Şekil 3.230.	XGBoost ile elde edilen değerler	298
Şekil 3.231.	K-en yakın komşu modelinde n komşu (n neighbors) ile AUC skoru	298
Şekil 3.232.	K-en yakın komşu modelinde p değeri ile AUC skoru	299
Şekil 3.233.	K-en yakın komşu modelinde leaf size ile AUC skoru	299
Şekil 3.234.	K-en yakın komşu ile elde edilen değerler	299
Şekil 3.235.	SVM modelinde derece ile AUC skoru	300
Şekil 3.236.	SVM modelinde components ile AUC skoru	300
Şekil 3.237.	SVM modelinde max iter ile AUC skoru	300
Şekil 3.238.	SVM ile elde edilen değerler	301
Şekil 3.239.	Stokastik gradyan azaltma modelinde max iter ile AUC skoru	301
Şekil 3.240.	Stokastik gradyan azaltma modelinde alpha ile AUC skoru	302
Şekil 3.241.	Stokastik gradyan azaltma ile elde edilen değerler	302
Şekil 3.242.	LGBM modelinde n estimators ile AUC skoru	303
Şekil 3.243.	LGBM modelinde öğrenme oranı (learning rate) ile AUC skoru	303
Şekil 3.244.	LGBM modelinde max depth ile AUC skoru	303
Şekil 3.245.	LGBM modelinde min child samples ile AUC skoru	304
Şekil 3.246.	LGBM modelinde subsample ile AUC skoru	304
Şekil 3.247.	LGBM modelinde min child weight ile AUC skoru	304
Şekil 3.248.	LGBM modelinde num leaves ile AUC skoru	304

Şekil 3.249.	LGBM modelinde min split gain ile AUC skoru	305
Şekil 3.250.	LGBM ile elde edilen değerler	305
Şekil 3.251.	Voting ile elde edilen değerler	306
Şekil 3.252.	Light gradyan artırma modeli ile elde edilen ROC eğrisi	307
Şekil 3.253.	Veri çekilen zaman aralığı ve tweet miktarları	309
Şekil 3.254.	Çekilen verinin örnek görünümü ve boş değer miktarı	310
Şekil 3.255.	Tweetler içindeki karakter sayıları	310
Şekil 3.256.	Tweet içindeki kelime sayıları	311
Şekil 3.257.	Tweetler içindeki ortalama kelime uzunlukları	311
Şekil 3.258.	En yoğun verinin olduğu 30 gün ve veri miktarları	311
Şekil 3.259.	Veri seti üzerinde duygu durum analizi	312
Şekil 3.260.	Duygu korelasyon matrisi	312
Şekil 3.261.	Tespit edilen duygu durumlarının öznelliği	313
Şekil 3.262.	Tespit edilen duygu durumlarının ortalama saat bazlı analizi	313
Şekil 3.263.	Popüler konulara ilişkin kelime bulutu (Maks. 2000)	314
Şekil 3.264.	#sex etiketinden hesaplara veri trafiği	315
Şekil 3.265.	#porn etiketinden hesaplara veri trafiği	315
Şekil 3.266.	#teen etiketinden hesaplara veri trafiği	315
Şekil 3.267.	#tits etiketinden hesaplara veri trafiği	315
Şekil 3.268.	#anal etiketinden hesaplara veri trafiği	316
Şekil 3.269.	#sexy etiketinden hesaplara veri trafiği	316
Şekil 3.270.	#PKK etiketinden hesaplara veri trafiği	316
Şekil 3.271.	Popüler hesap etiketleri (@)	317
Şekil 3.272.	@Borisjohnson hesap etiketinden konu etiketlerine veri trafiği	318
Şekil 3.273.	@RTErdogan hesap etiketinden konu etiketlerine veri trafiği	318
Şekil 3.274.	#TurkeyisNotOurFriend konu etiketinden hesap etiketlerine veri trafiği	318
Şekil 3.275.	En popüler 200 kullanıcının hesap analizi	319
Şekil 3.276.	LDA konu modellemesi	319
Şekil 3.277.	t-SNE modellemesi üzerinden vektörel düzeyde kümeleme	321
Şekil 3.278.	“turkey” kelimesi üzerinden ilişki analizi	321
Şekil 3.279.	“turkey” kelimesiyle en ilişkili 10 kelime ile en ilgisiz 10 kelimenin t-SNE ile gösterimi	322

Şekil 3.280.	“pkk” kelimesi üzerinden ilişkisel analiz	322
Şekil 3.279.	“pkk” kelimesiyle en ilişkili 10 kelime ile en ilgisiz 10 kelimenin t-SNE ile gösterimi	323
Şekil 3.280.	İstihbarat temininde twitter benzerlik matrisi	324



SİMGELER VE KISALTMALAR

ADB/ AdaBoost	Uyumlu Artırma (Adaptive Boosting)
AI	Yapay Zekâ (Artificial Intelligence)
ANN	Yapay Sinir Ağları (Artificial Neural Network)
ASCII	Bilgi Alışverişi için Standart Amerikan Kodu (American Standard Code for Information Interchange)
AUC	Eğri Altında Kalan Alan (Area Under Curve)
BBC	British Broadcasting Corporation
BoW	Kelimeler Çantası (Bag of Words)
BPNN	Geriye Yayılım Sinir Ağı (Back Propagation Neural Network)
CART	Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees) algoritması
CBOW	Devam Eden Kelime Çantaları (Continuous Bag of Words)
CFG	İçerikten Bağımsız Gramer (Context-Free Grammar)
CNN	Cable News Network
CNN/ConvNet	Evrişimsel Sinir Ağları (Convolutional Neural Network)
CRF	Koşullu Rastgele Alan (Conditional Random Field)
CRISP-DM	Veri Madenciliği için Sektörler Arası Standart Süreç (Cross Industry Standard Process for Data Mining)
CSV	Virgülle Ayrılmış Değişkenler (Comma-Separated Variables)
DW	Veri Ambarı/Depolama (Data warehouse)
EB	Exabyte
EDA	Keşifsel Veri Analizi (Exploratory Data Analysis)
EDA	Keşifsel Veri Analizi (Exploratory Data Analysis)
FN	Yanlış Negatif (False Negative)
FP	Yanlış Pozitif (False Positive)
GB	Gigabyte
GBM	Gradyan Artırma Makineleri (Gradient Boosting Machines)
GloVe	Kelime Temsilinde Küresel Vektör (Global Vectors for Word Representation)

GPU	Grafik İşleme Ünitesi (Graphics Processing Unit)
HMM	Gizli Markov Modeli (Hidden Markov Model)
IoT	Nesnelerin interneti (Internet of Things)
IP	İnternet Protokolü (Internet Protocol)
JSON	JavaScript Nesnesi Gösterimi (JavaScript Object Notation)
KB	Kilobayt
KDD	Veri Tabanlarında Bilgi Keşfi (Knowledge Discovery in Databases)
KDT	Metinde Bilgi Keşfi (Knowledge-Discovery in Text)
KNN	K-En Yakın Komşu Modeli (K - Nearest Neighbors Model)
LDA	Gizli Dirichlet Tahsisi (Latent Dirichlet Allocation)
LGBM	Light Gradyan Artırma Makine (Light Gradient Boosting Machine)
LIFO	Son Giren İlk Çıkar (Last In First Out)
LSA/LSI	Gizli Anlamsal Analiz (Latent Semantic Analysis)
LSVA	Lineer Karar Destek Makine (Linear Support Vector Machine)
MAE	Ortalama Mutlak Hata (Mean Absolute Error)
MB	Megabayt
MBR	Bellek Temelli Nedenleme (Memory-Based Reasoning)
ML	Makine Öğrenmesi (Machine Learning)
MSE	Ortalama Kare Hatası (Mean Squared Error)
NB	Naive Bayes
NER	Varlık İsmi Tanıma (Named Entity Recognition)
NLP	Doğal Dil İşleme (Natural Language Processing)
NLTK	Doğal Dil Araçları (Natural Language Toolkit)
OSINT	Açık Kaynaklarda İstihbarat (Intelligence From Open-Sources)
PB	Petabyte
PCA	Temel Bileşen Analizi (Principal Component Analysis)
PCFG	Olasılıklı İçerikten Bağımsız Gramer (Probabilistic Context-Free Grammar)
pLSA	Olasılıksal Gizli Anlamsal Analiz (Probabilistic Latent Semantic Analysis)

POS	Sözcük Türü Etiketleme (Part-of-Speech Tagging)
QDA	Karesel Diskriminant Analizi (Quadratic Discriminant Analysis)
Re/Regex	Kurallı İfadeler (Regular Expression)
RF	Rastgele Orman (Random Forest)
RMSE	Kök Ortalama Kare Hatası (Root Mean Squared Error)
RNN	Tekrarlayan Sinir Ağları (Recurrent Neural Network)
ROC	Alıcı İşletim Karakteristiği (Receiver Operating Characteristic)
SEMMA	Örnekleme, Keşfetme, Değiştirme, Modelleme ve Değerlendirme (Sample, Explore, Modify, Model, and Assess)
SGD	Stokastik Gradyan Azaltma (Stochastic Gradient Descent)
SGD	Stokastik Gradyan Azaltma (Stochastic Gradient Descent)
SI	Uluslararası Birimler Sistemi (International System of Units)
SMOTE	Örnekleme Tekniği Üzerinde Sentetik Azınlık (Synthetic Minority Oversampling Technique)
SVD	Tekil Değer Ayrışımı (Singular Value Decomposition)
SVM	Destek Vektör Makineleri (Support Vector Machine)
TB	Terabayt
TDGH	Toplum Destekli Güvenlik Hizmeti
TF-IDF	Terim Frekansı-Ters Doküman Frekansı (Term Frequency - Inverse Document Frequency)
TM	Metin Madenciliği (Text Mining)
TN	Doğru Negatif (True Negative)
TP	Doğru Pozitif (True Positive)
t-SNE	t-Dağıtık [Stokastik] Olasılıklı Yakınlık Gömmesi (t-Distributed Stochastic Neighbor Embedding)
URL	Standart Kaynak Bulucu (Uniform Resource Locator)
UTF	Unicode Dönüşüm Biçimi (Unicode Transformation Format)
WCSS	Kümeler İçi Kareler Toplamı (Within Cluster Sum of Squares)
Word2Vec	Kelime Vektörleri
XGBoost	Ekstrem Gradyan Arttırma (Extreme Gradient Boosting)
ZB	Zettabyte

ÇİZELGELER

Çizelge 1.1.	Makine öğrenmesine dayalı farklı dillerden çeviri örneği	49
Çizelge 1.2.	Kappa değerlerinin yorumlanması	90
Çizelge 1.3.	ASCII karakter tablosu	98
Çizelge 1.4.	Treebank POS etiketleri	117
Çizelge 1.5.	Hipotezler ve araştırma soruları	133
Çizelge 3.1.	Jandarma ve Emniyet'in paylaşım yaptığı konuların LDA analizi ve konular arası mesafe haritalandırması	188
Çizelge 3.2.	Carabinieri ve Polizia di Stato'nun paylaşım yaptığı konuların LDA analizi ve konular arası mesafe haritalandırması	215
Çizelge 3.3.	Gendarmerie ve Police'nin paylaşım yaptığı konuların LDA analizi ve konular arası mesafe haritalandırması	242
Çizelge 3.4.	Guardia Civil ve Policía Nacional'un paylaşım yaptığı konuların LDA analizi ve konular arası mesafe haritalandırması	269
Çizelge 3.5.	Doğruluk değeri bazında modellerin karşılaştırılması	306
Çizelge 3.6.	Modellerin test verisi üzerindeki sonuçlarının karşılaştırılması	308
Çizelge 3.7.	# konu etiketinde LDA konu modelleri	320

1. GİRİŞ

Tarih boyunca, dünya genelinde yaşanan değişimler hayatımızın birçok alanını etkilemiştir. Yaşanan her değişim, insanı ve onun yaşadığı çevre üzerinde farklılaşmalar meydana getirmiş ve kimi zaman iyi, kimi zamansa kötü sonuçların ortaya çıkmasına sebebiyet vermiştir. Yakın tarihteki en önemli değişikliklerden biri bilgi devrimidir (Berman ve ark., 2020; Soares ve Simão, 2019). Bilgi devrimi, sosyal ağların iletişimde çığır aşması ve nesnelerin veri transfer gücünü kullanması (IoT – Nesnelerin interneti) gibi çalışmalarla giderek tüm topluma sirayet etmiş ve içinde insan kas gücünü minimal, beyin gücünü ise maksimize şekilde kullanmayı amaçlayan Endüstri 4.0'ın da barındığı bilgi toplumu haline dönüşülmektedir (Ateş ve ark., 2020; Gentner, 2016).

Mevcut düzeni bilgi toplumu olmaya iten sebeplerin en başında ise özellikle internet kullanımı yer almaktadır (Jiao, 2017). İnternet kullanımı ile birlikte, insanlık açısından zaman ve mekân kısıtlamalarının kalkması, mevcut iletişim yeteneklerini arttırmış ve devamında gelişen sosyal ağlarla da gerçek dünyada varolan ilişkilerin, mevcut yaşamdaki kısıtlamalardan muaf olacak şekilde, hızlı ve özgürce yapılabilmesine olanak sağlamıştır (Gross ve Harmon, 2016).

İnsanlık tarihi boyunca avcı – toplayıcı topluluklardan başlayarak, meydana gelen her değişimin sonuçlarının tek başına iyi ya da kötü olamayacağı gerçeği, bilgi toplumunda da kendine yer bulmuştur. İnternet ve sosyal ağların ulus sınırlarını yıkarak, uluslar ötesi düzeyde sağladığı bilgi ve anlık iletişim yeteneği, pek çok alanda faydalı sonuçlar ortaya koyduğu gibi suç kavramı başta olmak üzere diğer pek çok alanda da mevcut düzenin değişerek, toplumda yadsınamaz yeni oluşumlara sebebiyet vermiştir (Dwivedi, 2020). Özellikle bilgi toplumu ile birlikte insanların sahip olduğu sınırsız imkânların, sanal ortamda yeterince kontrol edilememesi veya toplum normalarına uygun olmayacak şekilde kimi zaman suç ihtiva etmesi sebebiyle, bilinçli veya bilinçsiz suç ihtiva eden davranışlar oluşmaya başlamıştır (Arora, 2016; Stalans ve Finn, 2016).

Ateş ve Tokay (2018) tarafından sosyal ağlar temelinde ülkemiz özelinde yapılan çalışmada, katılımcıların %50.3'ünün sosyal ağda suçla karşılaştığını ifade ettiği, söz konusu suçla karşılaşan katılımcıların da %12.2'sinin jandarma, polis veya savcılıklara başvuruda bulunduğu ve %59'unun ise ilgili sosyal ağ sitesine şikâyetini bildirdiği belirlenmiştir. Söz konusu veriler, sosyal ağda suç oluşturan eylemlerin çoğundan devletin geri bildirim alamadığını ortaya koymaktadır.

Suçun toplumsal anlamda istenmeyen bir olgu olması ve toplumlarda mevcut düzen içerisinde çeşitli olumsuzluklara sebebiyet verdiği bilinen bir gerçek olup, devletlerin toplumsal kalite yönetimi gereği suçla mücadele ederek, mevcut düzen başta olmak üzere, diğer insanların hak ve özgürlüklerini koruması en başlıca ödevidir (Saar, 2017). Bu kapsamda sosyal ağlardaki suç gibi olumsuz eylemleri teşvik edenlerin veya direkt suç işleyenlerin eylemlerinin kısıtlanması ve yapılabilecek sanal devriyelerle kısmen de olsa insanların suç mağduru olmasının engellenmesi ve yine insanların bildirmediği karanlık alanda kalan suç noktalarının tespit edilmesi devletlerin çalışması gereken esas alandır (Hagan ve Daigle, 2018). Bu alandaki en büyük zorluk ise, ulus yapılarından arınmış şekilde uluslararası faaliyet gösterip, suçun işlendiği ülkenin mevzuatının ilgili internet veya sosyal medya uygulaması tarafından tanınmama ihtimalidir (Aryono ve Prastyanti, 2020).

Mevcut istatistikler genel olarak, sosyal ağların kullanımına bağlı olarak, suç ve suçluluğun dünya üzerinde artış eğilimi içinde olduğunu göstermektedir (Asongu ve ark., 2019; Andrews ve ark., 2018). Bu yüzden de suç önleme maksatlı stratejilerinin belirlenmesinde sosyal ağdan elde edilecek veri ve veriden elde edilen bilgi önemli olmaktadır. Gelecek dönem hedef ve stratejilerinin, amaçlara uygun ve etkin olabilmesi güncel, doğru ve güvenilir veriye bağlıdır.

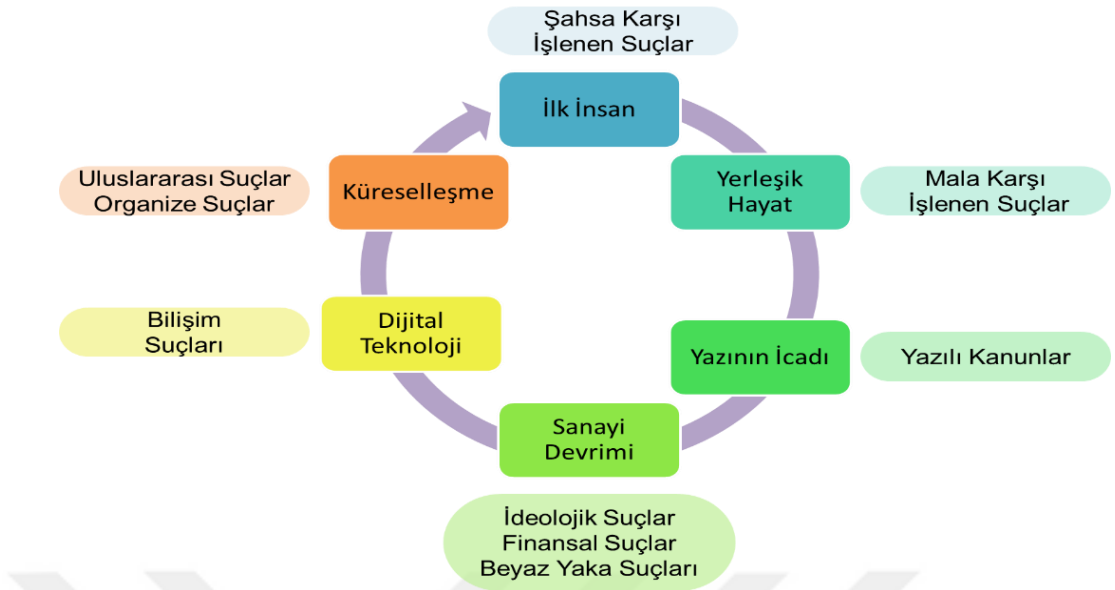
Verinin doğru, güvenilir ve güncel olmasının yanı sıra insanların her geçen gün artan oranlarda sosyal medya hesaplarını kullanması sonucu oluşan büyük veri, klasik istatistiksel yöntemlerle analize elverişli değildir (Arnaboldi ve ark., 2017; Emre ve Erol, 2017). Bu yüzden de sosyal ağda suçla mücadelede karşılaşılan veri yığınlarının veri madenciliği teknikleri kullanılarak işlenmesi, yapay zekâ temelinde çalışan makine ve

derin öğrenme algoritmalarıyla anlamlı bilgi haline getirilmesi ve elde edilen bilgilerin kolluğa hem önleyici hem de adli görevlerde destek vermesi temelinde tez çalışmasına yoğunlaşmıştır. Ayrıca tez kapsamında suç ve suçla mücadelede, toplum destekli güvenlik hizmeti kapsamında sosyal ağlardan vatandaşlarına destek olan Türkiye, Fransa, İtalya ve İspanya'nın Jandarma ve Polis teşkilatlarının paylaşımları makine öğrenimi temelinde incelenmiştir.

Tezin bulgular bölümünde, elde edilen tespitlere dayanak oluşturmak maksadıyla, bu bölümde alt maddeler halinde tezin kuramsal ve kavramsal çerçevesini oluşturan; suç kavramı, suçla mücadelede adli ve önleyici yaklaşım, sosyal medya, büyük veri, veri madenciliği, yapay zekâ, makine öğrenmesi, makine öğrenmesi yöntemleri, makine öğrenmesi algoritmaları, derin öğrenme, model geçerleme yöntemleri, model geçerleme ve performans metrikleri, metin madenciliği, metin verilerini işleme ve anlama, metin benzerliği ile konu modelleme kavramları açıklanarak, literatürde yer alan kullanım maksatları ele alınmıştır. Müteakibinde ise tezin amacı ve kapsamı ile araştırmanın önemi incelenmiştir.

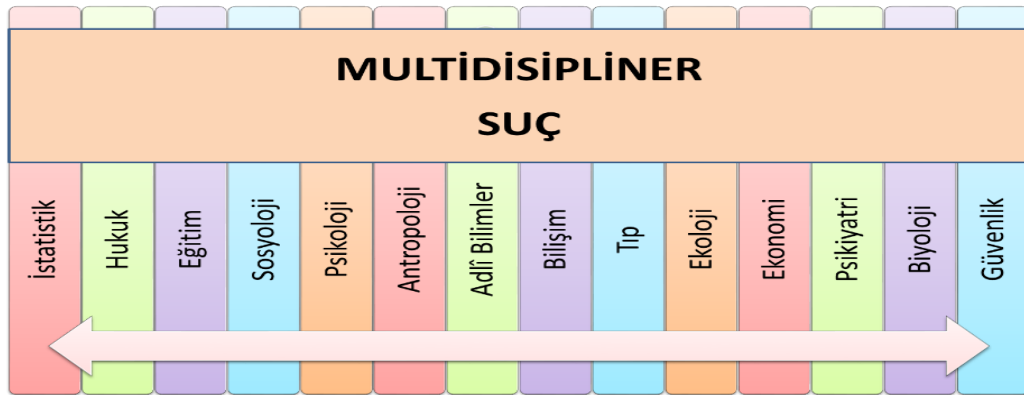
1.1. Suç Kavramının İncelenmesi

İnsanlığın varoluşundaki temel yaşam şekli tekildir, insanları tekil yaşamdan alıkoyarak çeşitli düzeylerde paradigma kaymalarına sebebiyet veren olgular ise genelde tarih boyunca karşılaşılan zorluklar olmuştur (Reiner, 2016). Tekil düzeyde bireyselci yaklaşımla özellikle bir dizi biyoloji ve güvenlik ihtiyacını karşılayamayan insanlar, pragmatik yaklaşımla birlikte bir arada yaşamaya başlamışlardır. Fakat bireysel düzeydeki olguların, toplu yaşamda özgürce yapılamaması zamanla bir takım kuralların mevcudiyetini zorunlu kılmıştır. Ortaya konulan kurallar çerçevesinde toplumda normlar oluşarak, çeşitli yasalarla istenmeyen davranış biçimleri tanımlanmıştır (Hagan ve Daigle, 2018). Yasalarla tanımlanan sapma türü davranış biçimleri ise toplumda çeşitli yaptırımları olan cezalarla korunmaya başlanmıştır.



Şekil 1.1. Suç olgusunun tarihi seyri

Şekil 1.1’de belirtildiği üzere, tarih boyunca suçlar zamana bağlı olarak çeşitli değişimlere uğramış ve ortaya çıkan her değişim suçun karmaşıklığını arttırarak, çözümünü zorlaştırmıştır (Göver ve Şenol, 2017). Zamana bağlı olarak gelişen suç kavramı, birey ve toplumların kendi aralarındaki ilişkisine bağlı olarak, zaman ve yer kavramları da dikkate alınarak değişim göstermiş ve kültür temelinde sürekli yeniden anlamlandırılmıştır (Levy ve Srebnick, 2017). Bu sebeple de insanlığın var olduğu ve henüz medeniyetlerin kurulmadığı ilk çağlardan beri suç, herkese aşina olan bir konudur.



Şekil 1.2. Suçla ilişkili bazı bilim dalları

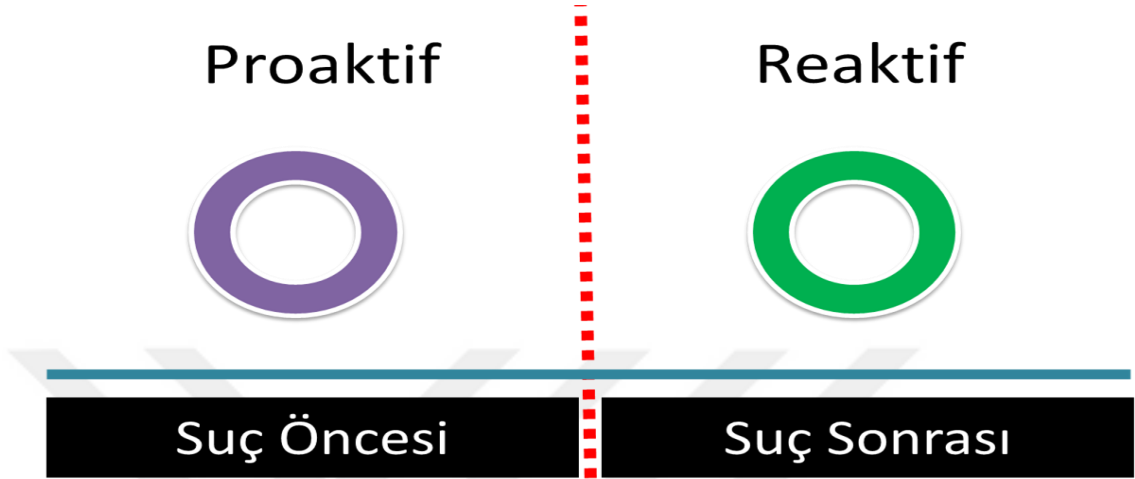
Karmaşası ve çözüm zorluğuyla problem haline gelen suç kavramı, aydınlanma dönemi ile başlayan süreçte bilimsellik temelinde anlama, tanımlama ve anlamlandırılmaya başlanmıştır (Newburn, 2017). Fakat suç ve suçluluk ile ilgili yapılan çalışmalarda, bilimsel manada net bir tanımlama ortaya koyulamamıştır. Bu durumun başlıca iki sebebi bulunmaktadır. Bunlardan ilki, şekil 1.2’de gösterildiği üzere, suç doğası itibariyle, çok sayıda disiplinle (inter-disciplinary) birlikte çalışma yapan bir alandır (Petherick ve ark., 2009; Taylor, 2015). Suç olayı ile ilgilenen her bir bilim dalı suç kavramını, kendi perspektifi ile açıklamaya çalışmakta ve diğer bilimsel disiplin dallarının alanına girememektedir. Suç ve suçluluk konusunda karşılaşılan diğer problem ise suç kavramının toplumlara, kültürlere ve zamana bağlı değişen tanımlamasıdır (Göver ve Şenol, 2017). Herhangi bir toplumda suç oluşturan fiil, farklı bir toplumda suç olmayabilir, ülkemizde zina örneğinde olduğu gibi aynı toplum yapısında da zamana bağlı olarak anlamlandırmada farklılıklar yaşanmaktadır.

1.2. Suçla Mücadelede Reaktif ve Proaktif Yaklaşım

Suç kavramının, kanun koyucu tarafından konulan kurallara muhalefet sonucu oluşması ve bu sebeple de genelde toplumda istenmeyen bir durum olması nedeniyle, devletlerce her zaman mücadele edilmesi gereken bir olgu olarak görüldüğü bilinmektedir. Suç işleyen bireylerin tespitinin yapılarak, yaptırım uygulanması aynı zamanda devlet otoritesinin de bir gereğidir. Yaptırımdaki amaç, kişinin ıslah edilmesi ve toplumdaki diğer suç işlemeyen bireyler gibi hareket etmesinin sağlanmasıdır (Sumachev, 2018).

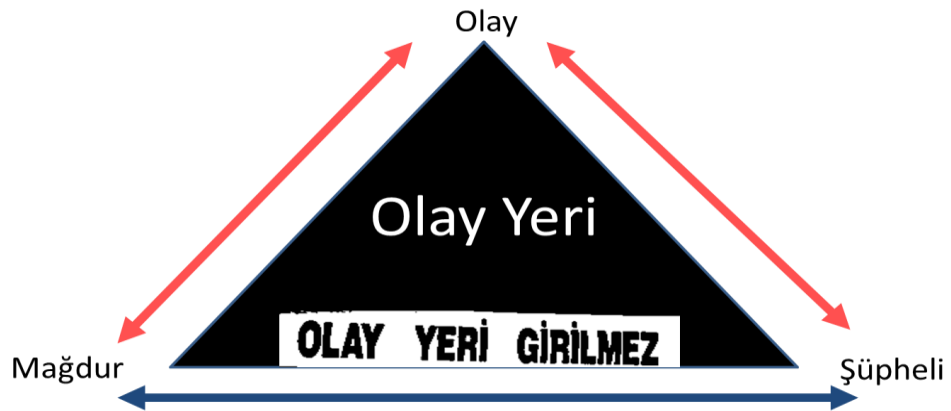
Suçla yaptırım uygulanmasının yanısıra suça dair olumsuz algının da etkisiyle “insanların neden suç işlediği” ve “suç işlendiğini ispatlamanın mümkün olup olmadığı” konuları da hep bir çalışma alanı olarak karşımıza çıkmıştır. İnsanların neden suç işlediğinin bilinmesi ve hangi motivasyonla suç işlediklerinin ortaya konulması suçun sebeplerini anlamamızı sağlarken, işlenen suçun ispatlanabilmesi ise toplumda kanun koyucuya olan güveni arttıracak ve haksız yere insanların ceza almalarının önüne geçilebilecektir. 19’uncu yüzyıl öncesi söz konusu suça yönelik

arařtırmalar pek yaygın olmamakla birlikte, bilhassa günümüzde adli bilimler temelinde oldukça sık karşılařtıđımız disiplinler arası bir alan olarak karřımıza çıkmaktadır (Houck ve Siegel, 2009).



Şekil 1.3. Suçla mücadelede temel yaklaşımlar

Adli bilimler disiplini altında suçla mücadele şekil 1.3’de gösterildiđi üzere reaktif ve proaktif olarak yapılabilmektedir (Brandl, 2019). Reaktif yaklaşım, kolluđun suçu ihbar, řikâyet veya doğrudan öğrenmesi ile başlayan adli suç soruřtırmalarını ifade etmekteyken, proaktif yaklaşım ise suçun oluşmadan önce mücadele edilmesine odaklanmaktadır.



Şekil 1.4. Suç üçgeni

Reaktif yaklaşımla suç olayını öğrenen kolluk, şekil 1.4’de görüldüğü üzere suç üçgeni içerisinde yer alan olay yerine giderek, ilk tedbirleri alır. İlk tedbirler kapsamında önce mevcut tehlikenin sonlandırılması ve hayati tehlikesi olanlarla ilgili sağlık kuruluşlarıyla koordinasyon gibi işlemlerin sağlanmasına müteakip, yetkili adli makamların talimatıyla suç olayının meydana geldiği olay yerindeki bulguları koruyarak, olay yerini incelemekle görevli uzman personelin yada acele hallerde bizzat kendisi, suça ilişkin kalıntı ve kayıtları toplayarak, gerektiğinde talimatla incelenmek üzere ekspertiz makamlarına gönderecektir (Brandl, 2019; Coupe ve ark., 2019). Suçun işlendiği mekânın sanal olması durumunda da yine aynı prosedür çerçevesinde yetkili adli makamların talimatıyla uzman personel tarafından olaya ilişkin kalıntı ve kayıtların toplanması amaçlanır (Holt ve ark., 2015)

Kalıntı ve kayıtların toplanmasındaki amaç suçun maddi gerçeğinin ortaya çıkarılması olup, yaklaşım tarzı ise özellikle Edmond Locard tarafından ortaya atılan ve Locard prensibi olarak da bilinen “her temas bir iz bırakır” şeklindedir (Mistek ve ark., 2018). Söz konusu yaklaşım çerçevesinde şekil 1.4’de görüldüğü üzere olayın tarafları arasındaki ilişki, olay-mağdur ve olay şüpheli ilişkisi belli bir zaman dilimi içerisinde mevcut kalıntı ve kayıtlarla mukayese edilerek, objektif şekilde gerçeğin açıklanmasına çalışılmaktadır (Siegel ve Mirakovits, 2015; Fish ve ark., 2013). Gerçeğin ortaya çıkması istenen esas sonuç olmakla birlikte, reaktif yaklaşımda olay çoğunlukla olup bitmiş ve mağdur ya da mağdurlar; ölüm, yaralanma, maddi ve manevi kayıp gibi çeşitli zararlar görmüştür.

Reaktif yaklaşım suç soruşturmalarında temel hareket tarzı olmakla birlikte, suçun mağdur insanlara verdiği zararın minimize edilmesi, suçluların çeşitli caydırıcılarla suç işlemekten alıkonulması ve gerektiğinde suça konu ortamda değişiklikler yapılarak, suçun işlenemez hale getirilmesi yani suç oluşmasının engellenmesi proaktif yaklaşım olarak karşımıza çıkmaktadır (Koper ve ark., 2020).

Proaktif yaklaşımda kolluk, suçun oluşmadan önlenmesi adına her türlü çabayı gösterir, kontrollerini arttırarak caydırıcı gücünü pekiştirir ve bununla birlikte genelde toplumun kendini daha güvende hissetmeleri sonucunu ortaya çıkarır. Proaktif

yaklaşım ile hareket ederken suçla karşılaşılması durumunda ise reaktif yaklaşım tarzındaki gereken adli işlemleri yapacaktır. Örneğin karayolu üzerinde herhangi bir ihbar veya şikâyet olmaksızın yol arama ve kontrol devriyemiz, yoldan geçen araçları kontrol etmesi proaktif bir yaklaşım iken, kontrol ettiği araçların birinde uyuşturucu bulmasıyla adli işlem tesis edilecektir. Söz konusu uyuşturucunun bulunması, uyuşturucunun götürülüp satıldığı kişilerin kullanımını engelleyerek, dolaylı olarak diğer insanların suç işlemlerinin önüne geçmektedir.

Proaktif yaklaşım, reaktif yaklaşıma göre kısmen daha yeni bir kavram olup, insanların suç mağduriyet yaşamalarının önüne geçilmesi için mücadele edilmesidir (Clarke, 2006). Suç mağduriyeti yaşamalarının önüne geçilmesi dışında, suç mağduriyetinin azalmasına paralel olarak kolluk ve adli makamlardaki iş yükünün azaltılması, toplumda suç korkusunun azaltılması ve toplum nezdinde oluşan güven ortamının etkisiyle vatandaşların kolluğa ve devlete olan güveni artacaktır. Söz konusu mücadele yukarıda yol kontrolü örneğinde verildiği üzere, sanal âlemde de olabilmekte olup, yapılan müdahale kanunen yetkili makamların da oluruyla yapılmalıdır.

1.3. Proaktif Yaklaşım Olarak Öngörüsöl Kolluk

Kolluğun reaktif ve proaktif yaklaşımı kapsamında hareket tarzı modellemeleri oldukça geniş yelpazede farklı öğeleri içermekle birlikte, son dönemde öngörüsöl kolluk modeli modern proaktif yaklaşımın önemli bir göstergesi konumundadır (Ferguson, 2016; Shapiro, 2017). Öngörüsöl kolluk ile farklı algoritmalarda analizler yapılarak, gelecekte olması muhtemel suçların tahmini, kullanılacak kolluk stratejisinin tespiti ve müdahalesine ilişkin modellemeler yapılabilmektedir (Cox ve ark., 2017).

Genel olarak tahmin ve öngörü her daim kolluk faaliyetlerinin bir parçası durumundadır. Bu kapsamda yine yol kontrol örneğine dönecek olursak, kontrol noktasında araç veya kişinin durdurulmasının içinde farklı olasılıklarda şüpheli

barındıran kollukça bir amacı olmalıdır, aksi takdirde rastgele uygulama yapmak zaman ve maliyet israfından başka bir şey olmayacaktır. Yapılan tahminlerin öznel olması, temelde bir felsefe ortaya koymayacak ve uygulamada farklılıklara sebebiyet verecektir. Bu sebeple de öngörüs el kolluk modeli, analize dayalı nesnel ve dinamik sonuçlar üretmektedir (Sappelli ve ark., 2017). Yapılan tahminlerde büyük verinin önemi oldukça fazladır (Shapiro, 2019). Makine öğrenmesi işlemi büyük suç verileri üzerinden gerçekleşmekte olup, suçun nerede ve ne zaman gerçekleşeceğini tahminleyebilen karmaşık algoritmalar kullanılmaktadır (Perry, 2013). Mohler ve ark. (2015) tarafından yapılan bir çalışmada, klasik olarak suç istihbaratını ve sıcak noktaları kullanma yöntemine göre, öngörüs el kolluk modellemelerinin 1,4 ile 2,2 kat daha fazla su u tahminleyebildiđi ve söz konusu tahminlemeleri kullanan kolluk devriyelerinin su  hacminde ortalama %7.4 azalma meydana geldiđi ortaya koyulmuştur.

Öngörüs el kolluk modeli ile klasik su  olayı, su lu ve mağdur tahminlemelerinin yanı sıra sosyal ađlar yardımıyla analiz yapılarak karar destek sistemleri tanımlanabilmekte ve yapılan analiz sonucu riskli unsurlar kolluđa bildirilebilmektedir (Cox ve ark., 2017).

1.4. Kolluk Modelleri

Kolluk kuvvetleri, toplum düzeni sağlamada devletlerin yegâne aracıdır. Bu anlamda Dünya’da oldukça geniz yelpazede olan kolluk modelleri mevcuttur; “askeri /devlet/ bürokratik kolluk”, “yasal kolluk”, “su la mücadele kolluđu”, “geniş kapsamlı kolluk”, “çok kurumlu kolluk”, “toplum odaklı kolluk”, “problem odaklı kolluk”, “teknolojik/bilgi/istihbarat odaklı kolluk”, “ kırık camlar modeli kolluk”, “sıfır tolerans kolluk”, “kamu- zel (böl nm ş) kolluk”, “konum temelli kolluk”, “ nleyici kolluk”, “toplum destekli güvenlik hizmeti kapsamında kolluk” gibi  ok fazla  eşide ayrılmaktadır (Ponsaers, 2001).  eşitlendirmedeki fazlalık,  lkelere göre yapılan farklı uygulamalardan kaynaklı olup, hepsi de kamu düzeninde istikrar sağlama görevini yür tmektedir. Modellemelerin her biri başı başına kesinlikle izlenilmesi

gereken bir yol olmamakla birlikte, genelde modellemelerin birçoğunun aynı anda olmasını ve kimi modellemelerin birbiriyle pozitif/negatif ilişkili sonuçlar ortaya koyabileceği sonucunun yansımasıdır. Bu kapsamda alt maddelerde söz konusu kolluk modelleri arasında daha yaygın kullanılan toplum destekli güvenlik hizmeti, problem odaklı kolluk ve istihbarat temelli kolluk modellerine değinilmektedir (Newburn, 2017).

1.4.1. Toplum Destekli Güvenlik Hizmeti

Topluluk destekli güvenlik hizmeti, suçun çok yönlü ele alınarak önleyiciliğin odak noktasında yer aldığı, suçla mücadelenin toplumun taleplerine göre şekillendirildiği ve mücadelede tüm kurumların sorumluluk aldığı, kolluk görevinin topluma hizmet olarak görüldüğü yaklaşımdır (Leventakis ve Habermeld, 2018).

Kolluğun toplumla iç içe güvenlikte hizmet sunması; kolluk ve toplum arasındaki mesafeyi yakınlaştıracak, halkla etkili iletişim sağlayacak, halkla yakınlıktan kaynaklı ihbar veya şikâyeteye dayalı bilgi bildirimlerinin kolluğa yapılmasını arttıracak ve sorunların toplum nezdinde tanımlanarak, mücadelenin birlikte yapılmasına imkân tanıyacaktır. Söz konusu yaklaşımın ortaya çıkardığı yapı, toplumundaki ana sıkıntıları gösterebilmesi açısından problem odaklı kolluğa ve toplumda olan bitenin kolluğun daha kolay haber edinimi açısından da istihbarat odaklı kolluğa destek vermektedir (Cox ve ark., 2017).

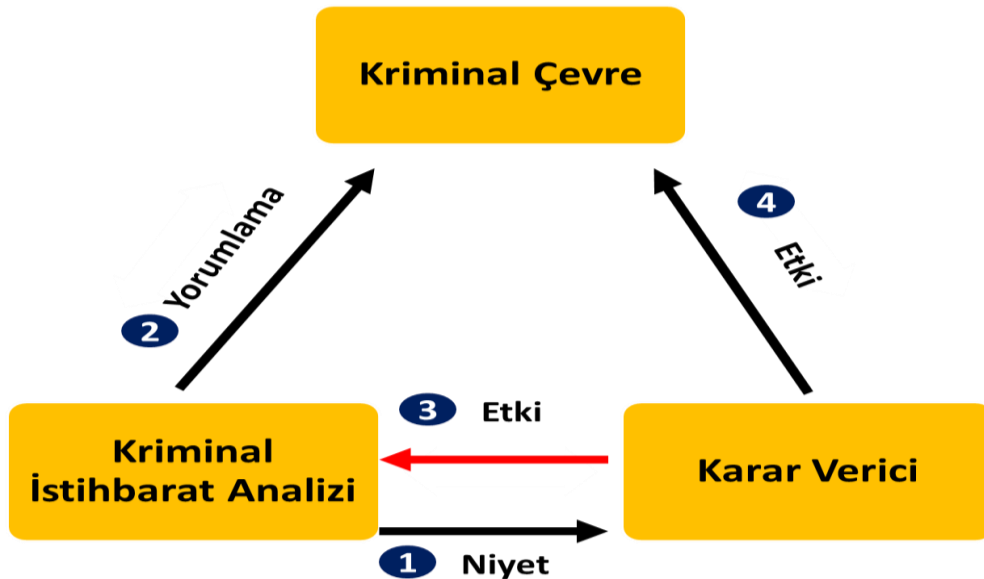
1.4.2. Problem Odaklı Kolluk

Kolluk kuvvetleri, her gün çok farklı zaman ve mekânlarda işlenen ya da işlenebilecek suç örgüsüyle karşılaşmaktadır. Suç ve suçu önlemede tüm sahada kolluğun hakim olması genel bir beklentiye, kaynakların sınırlı olduğu yaşamda imkânsızdır (Cox ve ark., 2017). Bu sebeple de önceliklerin belirlenmesi gerekliliktir. Problem odaklı kolluk ile reaktif modellemeden ziyade, geçmiş verilerden de

faýdalanılarak olası örüntülerin tespiti ve daha spesifik konu temelinde problemlere odaklanmak amaçtır (Scott ve ark., 2016). Özellikle mağdur kişilerin tekrar mağdur olma durumları, suç işleyen kişilerin tekrarlama durumları ve suçların sıcak noktalarının nereler olduğu temelinde çalışır. Bu kapsamda problemin ortaya net bir şekilde ortaya konulması; verinin, bilhassa büyük verinin (daha doğru modelleme için) ve verinin üzerinde çalışılacağı modelleme algoritmasının önemini ortaya koymaktadır (Sparrow, 2018).

1.4.3. İstihbarat Odaklı Kolluk

İstihbarat odaklı kolluk modeli, kolluk kaynaklarının verimli kullanımını istihbarat toplamaktan, farklı ifadeyle elde edilen verinin analitik süzgeçten geçirilmesi temelinde olmasıyla ortaya konabileceği yaklaşımdır (Cox ve ark., 2017; Burcher ve Whelan, 2019). Yapısı itibariyle oldukça karmaşıktır, hangi kaynaktan, hangi güvenilirlik seviyesinde, ne tür verilerin alınacağı ve söz konusu verilerin nasıl analitiğe dönüştürülerek, nasıl yorumlanacağı oldukça teknik olmasının yanı sıra görecelidir.



Şekil 1.5. İstihbarat odaklı kolluğun modellemesi

Ratcliffe (2016) tarafından yapılan 4i (intent, interpret, influence and impact) (niyet, yorumlama, etki (Diğer ifade modellemedeki görev farklılığından influence ve impact olarak belirtilmektedir.)) modellemesinde, istihbarat, çevre ve karar vericilerin arasındaki ilişkinin sonucu ortaya çıkan etkilerin ceza ortamını, yönetimi ve soruşturmaları etkilediğini ortaya koymuştur (Şekil 1.5). Bu kapsamda sahadan alınan istihbaratın, kolluğun öngörüsü üzerinde ve öngörüler üzerinden alınacak önlem stratejileri geliştirmesinde payı oldukça büyüktür (OSCE, 2017).

1.5. Sosyal Medya Kavramı

Günümüzde sosyal ağların hızlı gelişimi ile birlikte, internet kullanımının kapsamı oldukça genişlemiştir. Her geçen gün insanların duygularını ifade etme ve diğer insanlarla bir şeyler paylaşma gibi amaçlarla sosyal medyada sanal bir kişilik oluşturduğu dönemdeyiz (Mulyadi ve Fitriana, 2018). Genelde sosyal medya sitelerine üyelikle başlayan sanal kişilik oluşumu sonrasında, metin, ses, resim ve video gibi veri türlerinin transferi yapılabilmektedir. Veri transferinin yanı sıra sosyal medya, insanların birbirlerine mesaj, durum veya yorum gibi mesajları da atabildiği bir iletişim ortamıdır. Sosyal medyanın kullanımı iletişim başta olmak üzere pek çok disiplin için devrim niteliği taşımaktadır. Sosyal medya, insanlar arasındaki fiziksel iletişimi sanal boyuta taşıyan dijital ağ ortamıdır. Bu kapsamda konunun alt başlıklarında tez kapsamında sosyal medyanın; suç ve mağduriyete etkisi, açık kaynak istihbaratı açısından kullanımı, sosyal ağ analizi ve kolluğun halkla ilişkilerinde sosyal ağların önemine değinilmektedir.

1.5.1. Sosyal Medyanın Suç ve Mağduriyet Üzerine Etkisi

Suç kavramı, sosyal medyadan oldukça eski hatta ilk insanlardan beri var olan bir kavramdır. Fakat sosyal medyanın sağladığı zaman ve mekân sınırlamalarının ortadan kalktığı mevcut durum; suçun kavranması, işlenişi, yayılması ve etkileri üzerinde paradigma değişimi meydana getirmiştir.

Suçta paradigma değişimine sebep olan faktörler genel olarak;

- Sosyal ağların sanal kimlik kullanmaya imkân tanınması, gerçek kimliğin kullanılmamasının güvenlik güçleri tarafından yapılan işlemleri zorlaştırdığı ve bu durumun da suçluları oldukça fazla motive ettiği (Irshad ve Soomro, 2018),
- Sahte isimlerle tanınmış kişiler adına hesaplar açılarak toplumun manipülatif şekilde yönlendirilmesinin yapılabilmesi (Ferrari, 2018),
- İnternet ağ yapılarının IP gibi kimlik tespiti tanımlamalarının kolayca geliştirilen yazılımlarla farklılaştırmasının sağladığı rahatlık,
- Uygulamada sosyal ağların sanal-gerçek algı farklılıklarını ortadan kaldırmasına rağmen, kimi kullanıcıların sanal ortamda olduğu gerekçesiyle suç ihtiva eden davranış takınmaları,
- Konum etiketleme bilgilerinin maksat dışı kullanılarak rutin veya özel faaliyetlerin belirlenerek, fiziksel ortamda hırsızlık vb. suçlara sebebiyet verebilmesi (Omand ve ark., 2012),
- Sosyal ağ uygulamalarının uluslararası faaliyet gösteren birtakım şirketler tarafından yapılması ve ulus devlet anlayışında suç olarak belirlenen davranışların, şirket tarafından tanınmaması veyahut kısmen tanınmasının suçlular üzerindeki motivasyonu,
- Sosyal ağ uygulamalarının Cambridge Analytic davasında olduğu üzere, toplumun ilgi alanlarının belirlenerek manipülatif şekilde toplumun görüş ve düşüncelerinin değiştirilmeye çalışılması (Cadwalladr ve Graham-Harrison, 2018),
- İkinci kişilerle paylaşım yapılan kişisel veri ve bilgilerin, üçüncü kişiler tarafından ele geçirilerek istemli veya istem dışı kullanımının yarattığı veri güvenliği ve mahremiyeti sorunu (Wearesocial (2020) rakamlarına göre Türkiye’deki internet kullanıcılarının %58’i kişisel verilerinin yanlış kullanımından endişe duymaktadır.),
- Sosyal ağların iletişim kolaylığının organize örgütlerin faaliyetlerinde kullanılması ve organize olabilmelerindeki mekân etkisi (Cox ve ark., 2017),
- Radikal terörist grupların sosyal ağları propaganda ve eleman temini için kullanımı (Perliger ve Pedahzur, 2011),
- Sosyal ağlar üzerinden yalan haberlerin üretilmesiyle mevcut ulus ya da uluslararası kişi, grup veya devlete karşı itibarsızlaştırma ve karalama kampanyaları

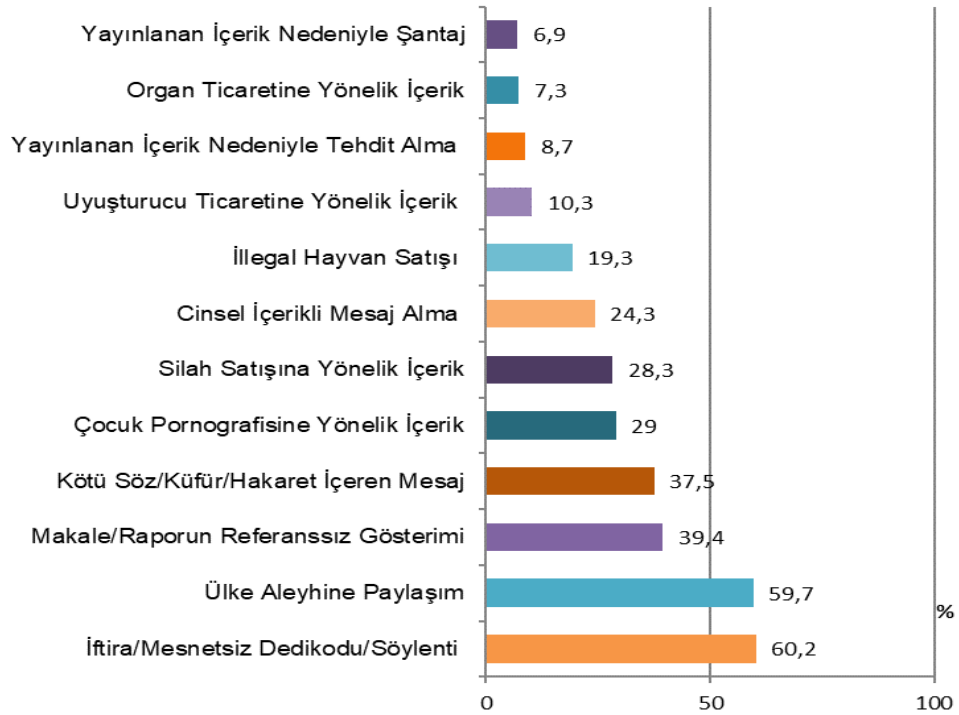
oluşturulması (Wearesocial (2020) rakamlarına göre Türkiye’deki internet kullanıcılarının %63’ü sahte haber ve yanlış bilgiye karşı endişe duymaktadır.),

– Mağdur insanların muhtemelen güvenlik güçlerinin suçluyu bulamayacağı düşüncesiyle karşılaştığı suçla ilgili kolluğa ihbarda bulunulmaması sonucu, gerçek suçun bilinmemesi ve karanlık alan boyutunun fiziksel ortamdan daha fazla olması (Ateş ve Tokay, 2018),

– Sosyal ağ şifrenin çalınması ile hesabın kontrolünün saldırgana geçmesi ve onay dışı bilgi paylaşımı,

– Kurum ve kuruluşlarının vatandaşlarla iletişimini arttırmak için sosyal medyayı aktif kullanmaya başlamaları (Tajudeen ve ark., 2018; Idris, 2018),

– İstihbarat temini ve elde edilen istihbaratın anlamlandırılması için sağladığı kolaylığın farklı gruplar tarafından yasal ya da yasal olmayan amaçlarla kullanılması şeklinde özetlenebilir (Paulo ve ark., 2013).



Şekil 1.6. Sosyal ağlarda karşılaşılan suçlar (Ateş ve Tokay, 2018)

Sosyal ağların suç açısından incelenmesini konu alan Ateş ve Tokay (2018) tarafından yapılan şekil 1.6’da görülen çalışmadaki değerler incelendiğinde, ortaya çıkan suç türü ve oranlarının hiç de azımsanmayacak boyutta olduğu görülmektedir.

1.5.2. Siber Zorbalık ve Sosyal Medya

Siber zorbalık, teknolojik aygıtların kullanılarak, arada güç dengesizliğinin olduğu başka bir kişi ya da gruba karşı saldırgan ve tekrarlayan olumsuz davranışlarda bulunma durumudur (Brailovskaia ve ark., 2018; Englander ve ark., 2017). Siber zorbalığın yol açtığı zarar, genelde fiziksel zorbalıktan daha fazladır. Zorbalığa konu öğeler, internet ve sosyal medya ortamında oldukça hızlı dağılabilmektedir. Ayrıca teknolojinin hızlı gelişimine paralel olarak, sanal platformlarda farklı metotlarda saldırgan profil kimliklerinin gizlenebilmesi ve bu sebeple de suçluların bulunmasında yaşanan zorluk, suçluları motive eden bir durum yaratmaktadır.

Günümüzde siber zorbalığın sosyal medya üzerinden işlenmesine ilişkin; tehdit edici ve kaba mesajlar gönderilmesi, özel kişisel bilgilerin ifşa edilmesi, çevrimiçi dedikodu yaratılması, tekrar eden küfür ve hakaret içerikli mesaj gönderimi ve herhangi bir kişi ya da grubun aşağılanması bazı siber zorbalık davranışları arasında yer almaktadır (Akça ve Sayımer, 2017). Siber zorbalık, her cinsiyette ve farklı yaş seviyelerinde ortaya çıkabilmekte olup, siber zorbalığa konu eylem de fiziksel, kültürel, ırksal ve hatta dini önyargılarla ilgili olabilmektedir (Ferreira ve ark., 2016; Muralidharan ve La Ferle, 2018).

Siber zorbalık davranışı sonrası, pek çok kişinin fiziksel ve zihinsel mağduriyet yaşadığı bilinmektedir (Gürhan, 2017). Söz konusu davranışlar aynı zamanda insanların sosyal medyaya ve çevreye olan güven eğilimini de giderek azaltmaktadır (Butt ve ark., 2019; Camp, 2016). Bu sebeple siber zorbalığın sosyal medya üzerinden tespit edilebilmesi bir ihtiyaçtan öte gerekliliktir. 2020 yılı Twitter örneklemini incelendiğinde 1 dakika içerisinde farklı dillerde ortalama 194.444 tweet atılıyor (Lewis ve Callahan, 2020). 24 saatlik bir zaman zarfı örnekleminde ise gönderilen ortalama tweet sayısı 279.999.360 rakamına ulaşmaktadır ve farklı dillerde böylesine büyük bir veri üzerinde fiziksel inceleme yapmak neredeyse imkânsızdır. Bu sebeple günümüzde yapılan çalışmalar, öncelikle doğal dil işleme teknikleri ile metnin normalizasyonuna ve müteakibinde de makine öğrenimine bağlı olarak sınıflandırılmasına odaklanmaktadır.

1.5.3. Sosyal Ağ Analizi

Sosyal ağlar üzerinde yapılacak analizle, sosyal ağı kullanan kişilerin ve paylaşımlarının (Resim, metin, konum, polarite vb.) ilişkilerin ortaya konulmasıdır. Sosyal ağ analizi, sosyal varlıkların örüntülerini tespit etmek ve yorumlamak için ağların yapısal yönlerine odaklanmaktadır (Tayebi ve Glässer, 2016). Çalışma şekli, genelde ağın ayırt edici özellikleri üzerinedir. Veri madenciliği teknikleri ile büyük veri tabanlarında gizlenmiş kalıp ve örüntüleri bulabilmektedir.

Genel olarak ağ analizi ile aşağıda belirtilen analizler yapılabilmektedir (Kempe ve ark., 2005; Tayebi ve Glässer, 2016);

- Merkezilik analizi, bir ağdaki kullanıcıların önemini ve ağdaki konum ve etkisini belirlemeyi amaçlamaktadır.
- Topluluk tespiti, herhangi bir ağ yapısı içerisinde ağın geri kalanına göre daha sık birbiriyle bağlantı kuran yapıları tanımlanmaya çalışmaktadır.
- Bilgi yayılımı, ağlar üzerinden bilgi akışını ve süreci inceler.
- Bağlantı tahmini, belirli bir sosyal ağ için yapının zaman içinde nasıl geliştiğini, daha farklı ifadeyle yeni bağlantıların nasıl oluşacağını tahmin etmeyi amaçlar.
- Üretken modeller, ağlarının topolojisini, zamansal dinamikleri ve gerçek dünya modellerini simüle eden olasılıklı yaklaşımdır.

Sosyal ağlardan elde edilen veriler üzerinde görselleştirme de yapılabilmekte olup, elde edilen çıktılarından suç modellemeleri yapılabilmektedir (Tayebi ve Glässer, 2016).

1.5.4. Sosyal Medya ve Açık Kaynak İstihbaratı

Kullanımı için özel bir kısıtlama olmaksızın, kişilerin herkese açık olarak yaptığı paylaşımların diğer bütün insanlar gibi kolluk kuvvetleri tarafından da kullanılması sosyal medyada açık kaynak istihbaratını oluşturmaktadır (Bartlett ve ark., 2013).

Wearesocial (2020) üzerinden ülkemiz özelinde elde edilen veriler incelendiğinde;

- Sosyal medya hesaplarını nüfusun %64'ünün kullandığı (Tüm yaşlar 0- ∞),
- Günde yaklaşık ortalama 7 saat 29 dakika internetin kullanıldığı (4 saati telefon, 3 saat 29 dakikası bilgisayar, tablet, laptop vb.) ve söz konusu sürenin ortalama 2 saat 51 dakikasında sosyal ağların kullanıldığı,
- Bazısı aktif şekilde kullanmasa da ülkemizde kişi başına düşen sosyal ağ hesabı sayısının 9.1 olduğu,
- Sosyal ağların kullananların %44'ünün iş amacı için bireysel kullanım olduğu,
- 37.000.000 Facebook kullanıcısı olduğu ve yapılan paylaşımların kişi başı ayda ortalama 3 yorum aldığı,
- 38.000.000 Instagram kullanıcısı olduğu,
- 11.800.000 Twitter kullanıcısı olduğu,
- 8.400.000 LinkedIn kullanıcısı olduğu,
- 7.700.000 Snapchat kullanıcısı olduğu belirlenmiştir.

Tüm yaş grubu ortalamasında %64 kullanım oranı, kişi başı 9.1 hesap ve günlük ortalama 2 saat 51 dakikanın sosyal medyada geçirildiği ülkemiz, açık kaynak olarak devasa bir veri tabanı özelliği taşımaktadır. Kolluk kuvvetleri genel olarak iki şekilde bilgi temininde sosyal ağları kullanabilmektedir; bunlardan ilki gizlilik ihtiva etmeyen suç verilerinin yayılımına fırsat sağlayarak, sosyal medyanın popülaritesi ve kolay erişimi nedeniyle potansiyel suç veya suça ait bulguları toplamayı amaçlar (Brandl, 2019). Fakat sosyal medyada bilgilerin paylaşımı sonrası, veri üzerinde kontrolün

kaybolması ve üzerine yapılması muhtemel sahte haberler risk unsurudur. İkincisi ise, kolluk kuvvetlerinin aradığı bir konuda elde ettiği bilgiyi ilgili siteden çekmesidir (Brandl, 2019). Bu yaklaşım herkese açık olarak yapılan paylaşımları ihtiva eder. Kullanıcın gizlilik ayarlarına bağlı olarak; ilgi alanı, mesleği, adresi, iletişim bilgileri, arkadaşları, mesajları, fotoğrafı, ses kaydı, yorumu, konumu ve videoları gibi bilgilere erişim sağlanabilmektedir. Organize grupların (çete, mafya, örgüt vb.) üyeleriyle yaptığı iletişimdeki örüntüler, kullanılan şifreli kelimeler, kelime ağı, anomali hareketler ve olası kullanılan özel kelime ve emojiler bulunabilmektedir (Cox ve ark., 2017). Mevcut kullarımdaki iki yöntemin ikincisine ilave olarak, gerektiğinde sosyal ağ sitelerinden kullanıcıların IP (Internet Protocol) adresleri, abone kayıtları ve hesap sahipliği gibi bilgiler de alınabilmektedir. Fakat IP adresi tekil bir şahsı işaret etmekten ziyade, internet aboneliğinin göstergesidir ve çok basit ara programlarla teknik olarak değiştirebilmektedir. Bu yüzden IP numarasından dolayı herhangi bir şahsı direkt şüpheli olarak görmek yanlış bir yaklaşımdır, tek başına yeterli bir delil olmayıp, ikame delillerle desteklenmelidir (Gedik, 2019).

Açık kaynak istihbaratı suç örgülerini çözmede yardımcı pozisyonda olup, sosyal medya vasıtalarıyla bilgi almak, soruşturmaya ilişkin ipuçları verebilmektedir (Dawson ve ark., 2018; Sampson, 2017). Aynı zamanda kolluk kuvvetlerinin sosyal ağ hesaplarının olması, kullanıcıların rahatlıkla kollukla iletişim kurabilmelerine olanak sağlayarak, TDGH (Toplum Destekli Güvenlik Hizmeti) kapsamında kolluğa erişimin daha kolay olmasını ve halkla bütünleşmeyi sağlamaktadır (Cox ve ark., 2017).

1.5.5. Sosyal Medya, Halkla İlişkiler ve Güvenlik Güçleri

Jandarma ve Polis kavramı kimileri için soğuk ve korkulacak bir ifade olarak tanımlanabilmektedir. Bunun en önemli örneği olarak, herhangi bir sebeple karakola ilk kez gelen vatandaşların kimi zaman stresli olduğu rahatlıkla gözlemlenebilmektedir. Söz konusu duruma sebep olan etmenlerin başında, dizi/film sektörü ile görsel ve yazılı medyanın, güvenlik güçlerini genelde aksiyon ve şiddet

içinde göstermesidir (Cox ve ark., 2017). Ayrıca kimi zaman örneklem olarak gösterilen üniforma içindeki personelin uygunsuz davranışı zaman zaman tüm kuruma mal edilmektedir. Hem basılı hem de görsel ve sesli yayın yapan medya, haber hikâyelerini saniyeler içinde tüm ulusa ve dünyaya teknolojik olarak iletme yeteneğine sahiptir (Hess ve ark., 2016).

Suç soruşturmalarında genelde kolluk kuvvetleri, mümkün olduğunca basın organlarından uzak durmaktadır, çünkü suç olayının ilk aşamasında bilgiler net değildir (Cox ve ark., 2017). Net olmayan bilgilerin ifşası, soruşturmanın gizliliğini tehlikeye düşürebilmektedir (Hess ve ark., 2016). Fakat olayı haber yapmak isteyen basın organları kolluktan alamadığı bilgiyi, şüpheli kaynaklardan elde ederek, zaman zaman kamuoyu üzerinde yanlış algıların oluşmasına sebebiyet vermektedir (Quandt ve ark., 2019).

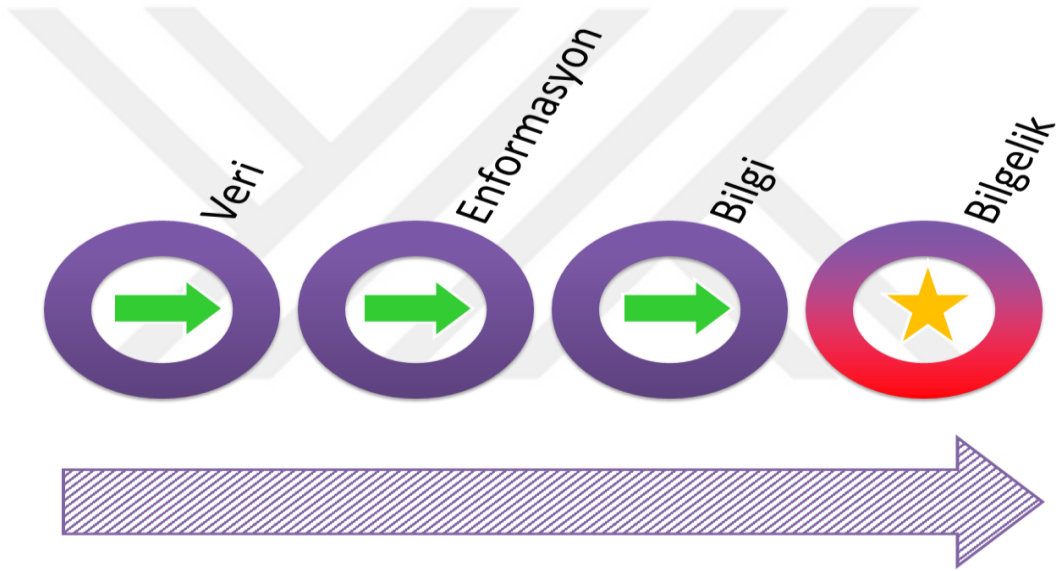
Bu sebeplerle son yıllarda güvenlik güçleri, toplumda oluşabilecek olumsuz algıların önüne geçmek, yaptıkları işi halka anlatmak ve yapılan paylaşımlarla halkla iç içe olunduğunun farklı bir ifadeyle suçla mücadele toplum destekli güvenlik hizmetinin gereği toplumla bütünleştiğini gösterebilmek adına aktif şekilde sosyal medya hesaplarını kullanmaya başlamıştır. Yapılan bu işlem özellikle halkla ilişki kurmada oldukça önemli olup, güvenlik güçlerinin halkla ilişkilerdeki amacı ise; suçla mücadele faaliyetlerinde halkın güven, destek ve işbirliğini sağlamaktır (Hamm ve ark., 2017).

1.6. Büyük Veri

Büyük veri kavramı, bilhassa son yıllarda en popüler araştırma konulardan biri haline gelmiştir. Özellikle işlem yapısı itibariyle bilgi teknolojilerinin tam merkezinde yer alması ve bilgi teknolojileri dünyasında yaşanan hızlı değişimin, verinin önemini daha da arttırması atfedilen önemi arttıran hususların başında gelmektedir. Bu kapsamda büyük veri kavramı; veri, enformasyon, bilgi ve bilgelik, veri bilimi ve büyük veri başlıkları altında incelenmiştir.

1.6.1. Veri, Enformasyon, Bilgi ve Bilgelik Kavramlarının İncelenmesi

Economist (2017) dergisi veri kavramına ilişkin yaptığı analizde; “Dünyanın en değerli kaynağı artık petrol değil, veridir.” (“The World’s most valuable resource is no longer oil, but Data”.) çıkarımında bulunmuştur. Doğru yer ve zamanda kullanılan veriler, yaptığımız işte başarıyı etkileyen temel faktörlerden biridir. Bu sebeple önemli olan konulardan biri de şekil 1.7’de gösterildiği üzere veri kavramının işlenerek enformasyona ve anlamlı bilginin yer aldığı bilgelik safhasında çıkarım yapmaya uygun nitelikte olmasıdır.



Şekil 1.7. Verinin bilgeliğe dönüşümü

Bu kapsamda “veri” kelimesi, gizlenmiş ancak henüz yorumlanmamış veya analiz edilmemiş bilgi (Zins, 2007), basit gözlem ve semboller, ham/organize edilmemiş gerçekler (Ahmed, 2020) ve belirli sorulara cevap vermeyen bir metin (Shao ve ark., 2017) şeklinde tanımlanmaktadır.

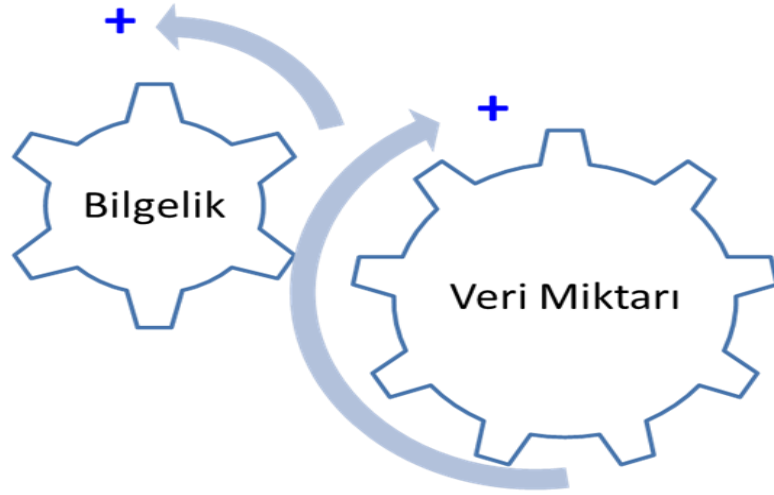
"Veri" kavramının günlük hayattaki kullanımında, genelde yapılandırılmış ya da yapılandırılmamış formatlardaki bilgi topluluğuna atıfta bulunuruz (Özdemir, 2016). Örneğin sosyal medyada her insanın kendi düşüncesini özgürce paylaştığı bir platformda yapılan paylaşımlar yapılandırılmamıştır. Muhtemelen paylaşımların

çoğunun konusu, amacı, dili, paylaşımı yapan kişi ya da grup öğeleri gibi değişkenlik taşıyabilecek kavramların farklılaştığını rahatlıkla görebiliriz. Söz konusu paylaşım şekli yapılandırılmamış bir veri türü olarak tanımlanırken, tanımlı sütunlar üzerinden yapılan örneğin cinsiyet, boy, yaş vb. gibi bilgileri ihtiva eden bir excel çizelgesi ise yapılandırılmış bir veri türü olarak isimlendirilmektedir.

Enformasyon ve bilgi kavramları incelendiğinde ise, söz konusu kavramlarının sınırlarının net olmadığı ve özellikle bu sözcüklerin ingilizcesi olan “knowledge” ve “information” kelimelerinin Türkçeye çevrilisinde, ikisinde “bilgi” kelimesi ile anılması sebebiyle sıkça karıştırıldığı belirlenmiştir (Yılmaz, 2009). Bu kapsamda veri kavramı oldukça genel bir kavramken, enformasyonda aklın süzgecinde geçmesi söz konusu olmakta, bilgi kavramında verinin anlam kazanmış hali karşımıza çıkarken, tüm bunların bileşiminden de bilgelik kavramı ortaya çıkmaktadır (Hey, 2004). Bilgelik kavramıyla birlikte, ilk önce ham olarak karşımıza çıkan veri, anlamlı hale gelerek, karar destek faaliyetlerimizi desteklemektedir (Pauleen ve ark., 2017).

Veri, enformasyon, bilge ve bilgelik kavramlarını bir örnekte açıklayacak olursak;

- Trafik ışığı levhasında yeşil renk görülmesi anlamsız bir veriyken,
- Söz konusu trafik ışığının 20 metre ilerde olan “A” kavşağında yanması enformasyon,
- Yanan trafik ışığının benim araç kullandığım yol istikametinde yer alması ve kırmızı ışığın “dur” anlamında olması bilgi,
- Kırmızı ışıktaki geçilmesi durumunda trafik cezası yenilebileceğinin öngörülmesi ise bilgelik olarak karşımıza çıkmaktadır.



Şekil 1.8. Verinin bilgelikle beklenen ilişkisi

Bilgelik evresi, veri bilimi için en üst basamak olup, makine öğrenimi ya da yapay zekâ temelinde ortaya çıkan metotların da yegâne amacı insanoğluna yapacağı faaliyetlerde karar desteği verebilmektir. Bu aşamada verinin niteliğinin yanı sıra aynı zamanda sayısal niceliğinin de fazla olması karar destek mekanizmalarını etkileyen bir husus olarak karşımıza çıkmaktadır (Sharma ve ark., 2016). Konuyu yine suç açısından bir örnekle açıklayacak olursak, A ilçesinde son bir yıl içerisinde sadece bir adet evden hırsızlık olayı meydana gelmiştir. Meydana gelen olayın zamanı ise gündüz saat 12:00-14:00 arası olduğu belirlenmiştir. Bu tespitten yola çıkarak, sonraki süreçte A ilçesinde meydana gelebilecek hırsızlık vakalarının saat 12:00-14:00 arasında olmasını beklemek oldukça yanlış olacaktır. Veriden çıkarım yapabilmek için şekil 1.8’de görüldüğü üzere verinin miktarının da fazla olması beklenir. Veri miktarının fazla olması, bilgelige giden yolda doğru cevaplara erişme kabiliyetini de arttıracaktır, bu anlamda aradaki ilişkinin doğru orantılı olduğunu söylemek yanlış olmayacaktır. Veri miktarının niteliğinin yanında nicelik olarak da fazla olması konusu büyük veri kavramının önemini ortaya koymaktadır.

1.6.2. Veri Bilimi

Veri bilimi kavramı tarihte ilk kez Peter Naur tarafından 1974 yılında yayınlanan “Concise Survey of Computer Methods” isimli kitapta kullanılmıştır (Cao, 2017). Özellikle internet kavramının çok daha yaygın kullanımı ile birlikte son yıllarda veri bilimi olarak isimlendirilen bu kavram, çok kullanılmaya başlanmıştır. Söz konusu kavram incelendiğinde, veri biliminin veri yoluyla bilgi edinme sanatı ya da bilimi olarak tanımlandığı anlaşılmaktadır (Cielen ve ark., 2016). Bu yüzden de veri bilimi, karar vermede, geçmişi ve bugünü anlamada, geleceği tahmin etmede ve yeni endüstri ve ürün gelişiminde aktif olarak kullanılmaktadır (Özdemir, 2016).

Uluslararası Birimler Sistemi’ne (SI - International System of Units) göre “Bit” bilinen en küçük veri birimi olarak bilinmekte olup, bir bitin 1 (açık) ve 0 (kapalı) olmak üzere iki olası değeri vardır. Bir “bayt” ise 8 bitten oluşmakta olup, olası değerler üzerinden hesaplama yapıldığında $2^8 = 256$ değeri temsil etmektedir. Bayt boyutundan daha büyük miktarlarda veriler için 1000’in katları kullanılmaktadır; 1 Kilobayt (KB) 1000 bayta eşittir, 1 Megabayt (MB) 1000 KB’ye eşittir, 1 Gigabayt (GB) 1000 MB’ye eşittir, 1 Terabayt (TB) 1000 GB’a eşittir, 1 Petabayt (PB) 1000 TB’a, 1 Exabayt (EB) 1000 PB’ye ve 1 Zettabayt (ZB) 1000 EB’ye eşittir. Mevcut veri bilimine göre en yüksek veri değeri olan 1 Zettabayt $10^{21} = 1.000.000.000.000.000.000$ bayt değerine eşit kabul edilmektedir.

Günümüzde zettabayt değerleri aktif olarak kullanılmaktadır. Teknolojik gelişmelere bağlı olarak, veri miktarlarının giderek büyük bir ivme ile arttığı ve verilerin çeşitli formlarda ve farklı kaynaklardan toplanması sebebiyle de, genelde yapılandırılmamış oldukları bilinmektedir. Bu durumda verilerin çözümlenebilmesi adına, veri biliminin farklı disiplinlerle etkileşim içine girmesi zorunluluk halini almıştır.



Şekil 1.9. Veri biliminde Conway(2010)'in venn diyagramı

Conway (2010) tarafından veri bilimcisinin sahip olması gereken özellikleri ihtiva eden şekil 1.9'da belirtilen veri bilimi diyagramı incelendiğinde;

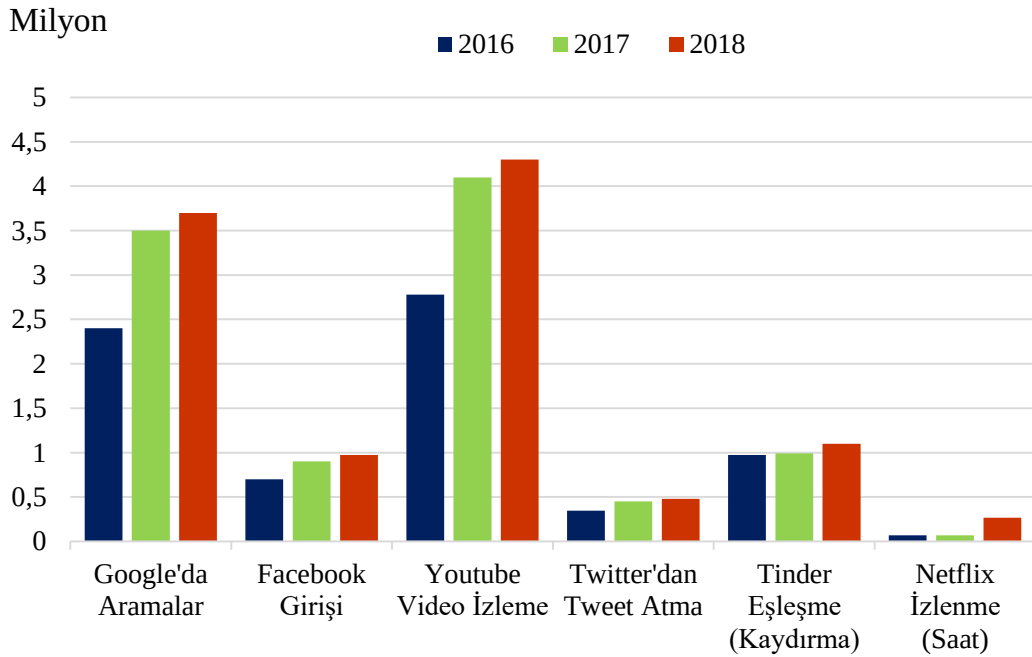
- Uzmanlık, matematik ve istatistik ile bilgisayar korsanlığı becerileri yeteneklerinin hepsinin veri bilimcisi için gerekliliği,
- Uzmanlık ile matematik ve istatistik bilgilerinin geleneksel araştırmayı,
- Bilgisayar korsanlığı becerileri ile matematik ve istatistik bilgilerinin makine öğrenmesini,
- Uzmanlık bilgisi ile bilgisayar korsanlığı becerilerinin olmasının ise tehlikeli bölge olarak adlandırılan alanı ihtiva ettiğini belirlenmiştir.

1.6.3. Büyük Veri

Geleneksel veri yönetimi tekniklerinin, verilerin boyutu ve karmaşıklığı nedeniyle yönetemediği veya işleyemediği büyük miktarda veriyi karakterize etmek

için ilk kez Roger Magoulas tarafından “Büyük Veri” kavramı kullanılmıştır (Halevi ve Moed, 2012). “Büyük veri” kavramı, çoğunlukla internet ve bilgisayar ağları gibi çeşitli kaynaklardan çok hızlı şekilde gelen büyük hacimli veri olarak (Villars ve ark., 2011), daha önce görülmemiş bir ölçekte depolanan soyut disiplinler arası kavram olarak (Chen ve ark., 2014) ve standart bilgi teknolojileri ile erişimin zor olduğu potansiyel büyük veri setleri üzerinde veri yakalama, işleme, toplama, görüntüleme ve analiz etme yöntemlerinin kullanılmasını kapsayan veri topluluğu (Vaidhyathan ve Bulock, 2014) olarak tanımlanmıştır.

Büyük veri kavramı, özellikle teknolojinin gelişmesine paralel olarak sürekli artmaktadır, mevcut yapı değerlendirildiğinde 2000’li yıllarda toplam hafızası 64 MB boyutunda yer alan MP3 çalar aletlerine ortalama 200’ü aşkın şarkı atabiliyorken, 2020 yılı itibariyle aynı belleğe sahip bir MP3 çalara ortalama 4 veya 5 adet şarkı atılabilmektedir. Söz konusu büyük veri artışının yılda ortalama % 40-60 oranında büyüdüğü değerlendirilmektedir (Wielki, 2013).



Şekil 1.10. 2016-2018 yılları arası internetteki bazı etkinlikler (1 dakika)

Şekil 1.10’da yer alan 2016, 2017 ve 2018 yıllarında bir dakika içinde internette neler olduğunu ortaya koyan veriler incelendiğinde (Lewis, 2016; Lewis, 2018);

- Google arama motorundaki arama sayısı 2,4 milyon iken 3,7 milyon (+ % 54,16),
- Facebook’da oturum açma sayısı 701.389 iken, 973.000 (+ % 38.72),
- Youtube platformu üzerinde video görüntüleme sayısı 2.78 milyon iken, 4.3 milyon (+ % 54.67),
- Twitter web sitesindeki yeni tweet sayısı 347.222 iken, 481.000 (+ % 38.52),
- Tinder web sitesindeki eşleştirme sayısı 972.222 iken, 1.1 milyon (+ % 13.14),
- Netflix platformunun izleme saati 69.444 saat iken, 266.000 (+ % 283.04) saat olduğu belirlenmiş olup, genel itibariyle bu değerler veri miktarlarının dünyadaki değişim hızını gözler önüne sermektedir.

Büyük veri kavramı genel itibariyle incelendiğinde, teknolojik bir yenilikten ziyade toplumda paradigma kayması yaratacağı değerlendirilmektedir (Antoncic, 2020). Bu sebeple de büyük veri kavramı mevcut düzende en popüler konulardan biri olarak karşımızda durmaktadır. Onu popüler yapan temel etmenlerden biri işletmeler için oluşan pazardaki rekabetin fazla olmasıdır. Rekabetin fazla olması, veriye atfedilen önemi arttırmakta ve bunun yanı sıra veri analizini çok daha önemli kılmaktadır (Singh, 2020). Bu sebeple de gelecekte de günümüzde olduğu gibi, büyük verinin öneminin giderek artması beklenmektedir.

Büyük veri kavramını daha da irdelemek adına literatürde 3V (Volume, Velocity, Variety – Hacim, Hız, Çeşitlilik), 4V (Volume, Velocity, Variety, Value – Hacim, Hız, Çeşitlilik, Değer) ve 5V (Volume, Velocity, Variety, Value, Veracity – Hacim, Hız, Çeşitlilik, Değer, Veri Doğrulama) gibi isimlerle adlandırılan ve büyük veriyi tanımlayan kuralları da incelemek daha tanımlayıcı olacaktır (Cielen ve ark. 2016; Wamba ve ark., 2015). Büyük veriyi tanımlayan bütün kurallarda ortak noktanın, hacim, hız ve çeşitlilik olduğu rahatlıkla görülmekte olup, değer ve gerçeklik

ise literatüre sonradan eklenmiş tanımlayıcı kurallardır. Fakat veri kaynakları ve çeşitliliğindeki artışlarla birlikte bu tanıma, bazı kaynaklarda Volatility (Oynaklık) ve Validity (Geçerlik) kavramlarının eklenerek 7V (Ramesh, 2017) yapıldığını, bazı kaynaklarda ise bunlara Vulnerability (Hassaslık), Variability (Değişkenlik) ve Visualization (Görselleştirme) kavramlarının ilave edilerek 10V (Manogaran ve ark., 2017) yapılsa da, şekil 2.5’de görüldüğü üzere esas itibariyle büyük veri 5V’de özetlenebilmektedir (Abdullah ve ark. 2015; Wamba ve ark., 2015).



Şekil 1.11. Büyük veriyi tanımlayan 5V kuralı

Şekil 1.11’de belirtildiği üzere büyük veri kavramını tanımlayan beş temel öğeyi inceleyecek olursak (Abdullah ve ark. 2015; Ateş ve ark., 2020; Wamba ve ark., 2015; Chen ve ark., 2014; Paryasto ve ark., 2014);

- Hacim, miktarı sürekli artan veriyi tanımlamak için kullanılmaktadır. Verilerin artmasıyla birlikte aynı zamanda bilgi yığını da oluşmaya başlamış ve hacimin artmasıyla birlikte de nitelikli bilgiye ulaşmanın zorlaşacağı,

- Çeşitlilik, üretilen verilerin çoğunluğu yapısal olmayan, sosyal ağ, cep telefonu, tablet vb. ortamlarda üretilmekte olup, veri kaynakları nispeten yeni olup, her geçen gün de yeni kaynaklar eklenmektedir. Bu durum da verinin kaynak sayısına artırarak, veri toplamada farklı format ve değişkenler altında verilerin çıkmasına olanak sağlayacak olup, analizi yapan kişinin ise kısmi oranda işini zorlaştıracaktır,
- Hız, verinin büyüme hızını ifade etmekte olup, verinin artma hızının yüksek oranlarda gerçekleşmesi aslında verinin büyük veri olduğunun da bir göstergesi olduğu,
- Veri doğrulama, verinin akışı sırasında güvenli olması önemli bir değişken olup, üst seviye güvenlik tedbirleri altında korunması, yetkisiz müdahalelerle değiştirilmemesi olduğu,
- Değer ise büyük verilerin bilgi yığını haline gelen verilerin değerlendirilmesini ve ham veriyi, enformasyon ve bilgiye dönüştürerek, karar destek birimlerine yardımcı olarak görev yaptığı ile ilgilidir.

1.7. Veri Madenciliği

Veri madenciliği, analiz yoluyla geçmiş verileri açıklayarak, ondan dersler çıkarma ve geleceğe ilişkin çıkarımlarda bulunma çabası olup, literatürde veri madenciliğinin tanımına ilişkin karşılaşılan bazı önemli tanımları aşağıda verilmiştir;

- İstatistikleri; kavramlar, araçlar, algoritmalar, makine öğrenimi ve çok büyük veri setlerinin analizi ile birleştirilerek anlama ve eyleme dönüştürülebilir bilginin elde edilmesini sağlamak (Williams, 2009).
- Önceden bilinmeyen verilerle ilgili potansiyel olarak yararlı bilgileri gizli olmayan bir şekilde ortaya çıkarmak (Srinivas ve ark., 2010),
- Büyük miktarda veriden bir bilgi alma veya madencilik çıkarma süreci olarak tanımlanabilen tüm bilgi keşif sürecinde bir adımdır ve özellikle belirli bir alandaki sorunları çözmek için gerekli olan bir bilgi keşfi biçimidir (Beniwal ve Arora, 2012; Olson ve Lauhoff, 2019).

Bu tanımlardan yola çıkarak, veri madenciliğinin diğer alanlarla ilişkisini de dikkate alarak şu tanımlı yapmak mümkündür: Veri madenciliği, geçmişte büyük veriden yararlanılarak, istatistik, makine öğrenimi, yapay zekâ gibi yöntemler kullanılarak bilgiyi elde etmeyi amaçlayan bir süreçtir ve genellikle bu bilgiyi gelecekle ilgili karar destek faaliyetlerinde kullanmaktır (Campbell ve Ying, 2011; Hand ve Adams, 2014).



Şekil 1.12. Veri madenciliği ve ilişkili olduğu disiplinler

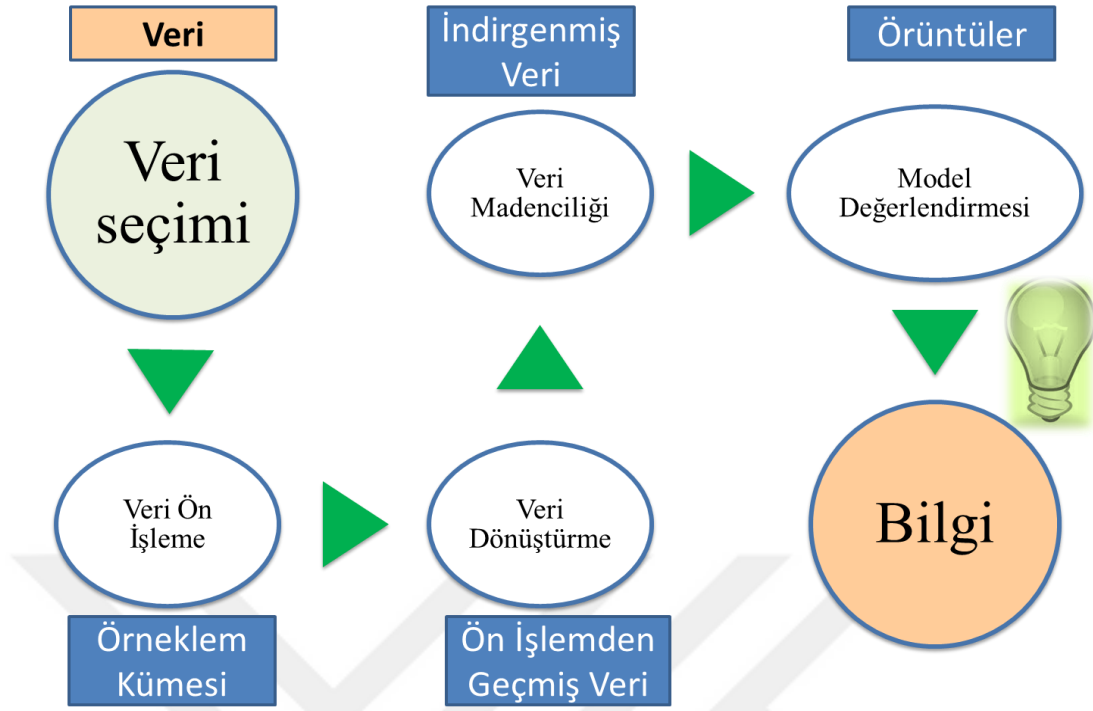
Veri madenciliğinin temel amacı, süreç verilerinden faydalı bilgiler elde ederek, sürecin işleyişini ve karar desteğini iyileştirmek için etkili bilgiyi ortaya koymaktır (Ge ve ark., 2017). Bu kapsamda da şekil 1.12’de görüldüğü üzere veri madenciliği; makine öğrenmesi, istatistik, veri tabanı ve yapay zekâ teknolojilerini bir araya getiren disiplinler arası bir alandır.

Veriler, taş yığınlarından çıkarılmayı bekleyen değerli madenler gibidir. Her gün çok fazla veri toplanan bir dünyada yaşamaktayız, teknolojinin gelişmesine paralel olarak, bilgisayar ve internet kavramları hayatımıza girerken veri boyutlarının ve türlerinin arttığı da bilinen bir gerçektir (Villars ve ark., 2011). Özellikle, veri

hacmindeki bu artış karmaşıklık sorununu beraberinde getirmektedir, bu yüzden doğru analiz edilmiş ve anlamlı verileri ortaya çıkarmak her geçen gün daha önemli hale gelmektedir.

Büyük veri içinde, verinin hacminin göreceli büyük olması sebebiyle, veriden nitelsiz olanları ayıklayarak, anlamlı veri elde etme işlemi öncelikle istatistik araştırmacılarını sınırlandırırken, sonrasında büyük boyutlu kümelerde mevcut istatistik yöntemlerinin yetersiz kalması sonucu veri madenciliği kavramı ortaya çıkmıştır (Özkan ve Boran, 2014). Farklı bir ifadeyle, büyük boyutlu veri kümesi, veri madenciliğini ortaya çıkarmakla birlikte, bu sonucu ortaya çıkartacak, içerisinde ilişki ve tahminlerin de yer aldığı sürecin temel kaynağıdır. Yöntemdeki önemli hususlardan biri de çok miktarda veri toplamaktan ziyade, mevcut verilerin analizidir. Veri madenciliği, genelde geçmiş örneklemden yararlanarak, gelecekte ortaya çıkacağı değerlendirilen olaylar hakkında tahmin modellerinin geliştirilmesini ve karar verme süreçlerine destek verilmesini amaçlanmaktadır. Veri madenciliği kavramı, literatürde; bilgi çıkarımı, veri madenciliği, bilgi keşfi, veri arkeolojisi, veri/desen analizi, bilgi madenciliği ve bilgi toplama gibi isimlerle de anılmaktadır (Khare ve Shrivasta, 2018; Koyuncugil ve Özgülbaş, 2009).

Veri madenciliği, ham veri yığınları arasında işlenebilir veriyi ortaya çıkartan ve bilgiyi elde etme amacıyla çeşitli modellemelerin de yapıldığı bir süreçtir. Öncelikle iş ve veri özellikleri detaylı analiz edilebilen her verinin, veri madenciliği işlemi mutlaka başarılı sonuçlar ortaya koyacaktır. Bu kapsamda özellikle veri tabanlarında bilginin keşfi, genel olarak KDD (Knowledge Discovery in Databases – Veri Tabanlarında Bilgi Keşfi) kısaltmasıyla da bilinmektedir (Koyuncugil ve Özgülbaş, 2009; Xu ve ark., 2014; Mariscal ve ark., 2010). KDD yöntemiyle bilgiye erişim, yinelemeli ve etkileşimli olmakla birlikte, gizli bilgileri veri tabanından ayıklama olarak da bilinmektedir (Shafique ve Qaiser, 2014).



Şekil 1.13. Veri tabanlarında bilgi keşfi (KDD)

Başarılı bir veri tabanlarında bilginin keşfi çalışmasında uygulanması gereken aşamalar şekil 1.13’de belirtildiği üzere aşağıda verilmiştir.

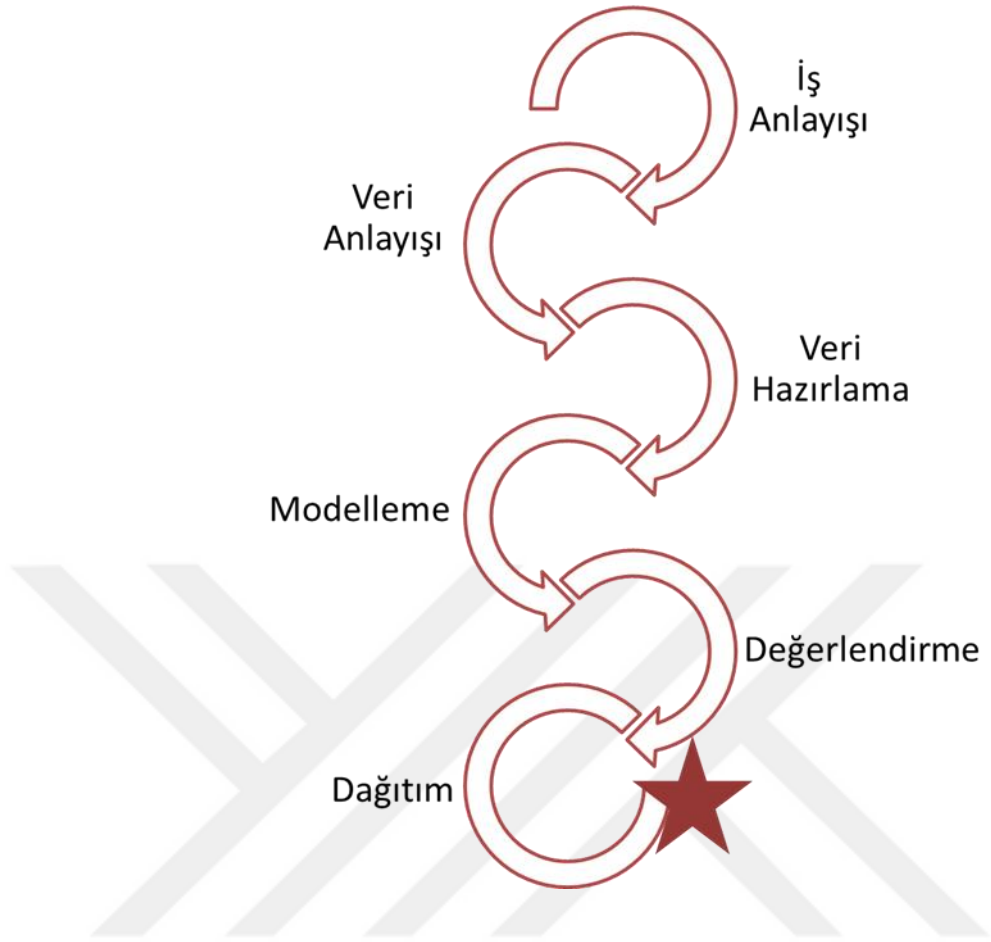
- Veri seçimi: Analiz ile ilgili verilerin seçimi yapılarak, veri toplama işlemine başlanır (Xu ve ark., 2014).
- Veri ön işleme: Veri yığını içerisindeki ilgisiz ve konuyla alakasız veriler dizinden kaldırılır (Deshpande ve Thakare, 2010). Genellikle heterojen olan birçok veri kaynağı, ortak bir kaynakta birleştirilir. Gereksiz verilerin temizlendiği safha olarak da bilinir (Koyuncugil ve Özgülbaş, 2009).
- Veri dönüştürme: Veri birleştirme olarak da bilinmekte olup, seçilen veriler madencilik prosedürüne uygun formata dönüştürülür (Chen ve ark., 2014).
- Veri madenciliği: Potansiyel olarak kullanışlı sonuçlar ortaya koyabilmek adına madencilik tekniklerinin uygulandığı önemli adımdır (Deshpande ve Thakare, 2010).

– Model değerdendirme: Bilgiyi temsil eden modeller belirli incelemelere dayanarak tanımlanır ve modelin işleyişı hakkında analizlerin yapıldığı adımdır (Koyuncugil ve Özgölbaş, 2009; Xu ve ark., 2014).

– Bilgi: Keşfedilen bilginin görsel olarak hedef kitleye sunulduğu son aşamadır. Özellikle kullanıcıların veri madenciliği sonuçlarını anlaması ve yorumlaması için görselleştirme tekniklerini kullanılabilir (Deshpande ve Thakare, 2010).

1.7.1. Veri Madenciliği Süreci

Veri madenciliği kavramında en önemli husus, ham veriden yola çıkarak, bilgiye ulaşımıdır. Bu işlem aynı zamanda bir süreci de beraberinde getirmektedir. Veri madenciliği süreci ile ilgili olarak, farklı kaynaklarda farklı tanımlamalar yapılmış olmakla birlikte, genel olarak CRISP-DM (Cross Industry Standard Process for Data Mining - Veri Madenciliği için Sektörler Arası Standart Süreç) (Keet ve ark., 2015) ve SEMMA (Sample, Explore, Modify, Model, and Asses - Örnekleme, Keşfetme, Değıştirme, Modelleme ve Değerdendirme) yöntemleri benimsenmektedir (Azevedo ve Santos, 2008; Shafique ve Qaiser, 2014; Wirth ve Hipp, 2000). SEMMA, ağırlıklı olarak modelin sonuçlarına odaklanırken, CRISP-DM modellemeyden önceki verinin anlaşılmasına, girdilere ve modellemeye daha çok odaklanmaktadır (Larose ve Larose, 2014). İlk bakışta CRISP-DM'nin SEMMA'dan daha eksiksiz olduğu sonucuna varmak mümkündür (Azevedo ve Santos, 2008). CRISP-DM metoduna göre, verinin bilgiye dönüşüm sürecinde toplamda altı aşama bulunmakta olup, aşamalar arası süreç çok katı olmamakla beraber, yapılan modellemenin geri bildirimi alınarak gerektiğinde geri ve ileri aşamaya geçilebilmektedir (Deshpande ve Thakare, 2010; Larose ve Larose, 2014).



Şekil 1.14. Veri tabanlarında bilgi keşfinde CRISP-DM modeli

Şekil 1.14’de belirtildiği üzere CRISP-DM veri madenciliğinin döngüsü altı ana aşamadan oluşmakta olup, söz konusu aşamalar incelendiğinde;

- İş Anlayışı: Öncelikle veri madenciliğinin hangi amaç ve ihtiyaçtan yapıldığının iyi anlaşılması, hangi bilginin veri madenciliği problemine konu olacağının belirlenerek, problemin tanımlanması için ön plana dönüştürülme işlemlerini konu alır (Huber ve ark., 2019).

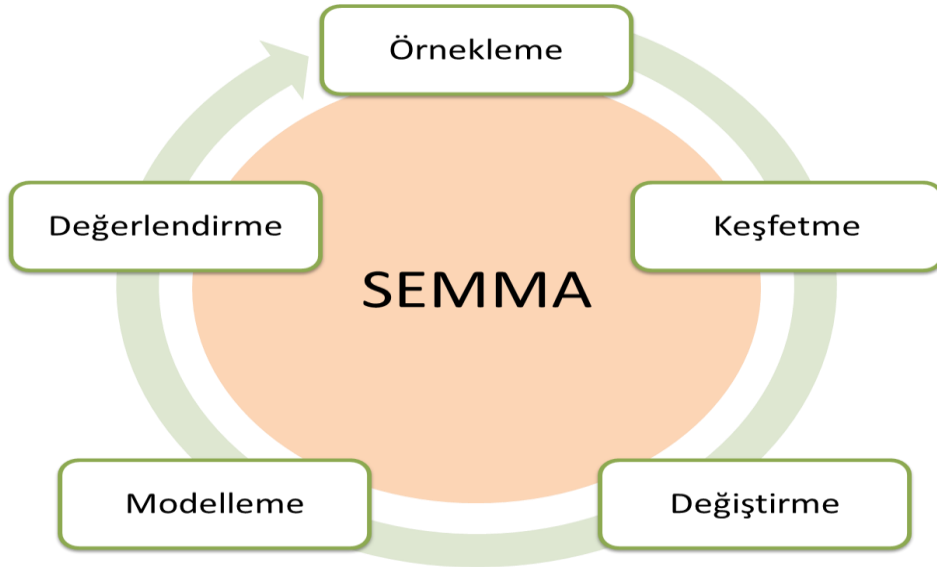
- Veri Anlayışı: Süreç, incelemeye konu verinin tanınması ve toplanması ile başlamakta olup, veri kalitesi sorunlarının belirlenmesi ve varsa veri toplama sırasında çıkan farklı tespitlerin de değerlendirilerek, hipotezler oluşturulmasını konu alır (Deshpande ve Thakare, 2010).

– Veri Hazırlama: İncelemeye konu olan ham veriden, gerekli elemelerin ve temizliğin yapılarak, incelemeye konu nihai veri kümesini oluşturmak için yapılan aşamaları kapsar (Khan ve ark., 2014).

– Modelleme: Çeşitli modelleme tekniklerinin analiz edildiği süreç olup, aynı zamanda farklı tipte modellerin farklı bilgi ihtiyaçları da olacağı değerlendirildiğinde sık sık veri hazırlama aşamasına dönüldüğü alandır (Wirth ve Hipp, 2000).

– Değerlendirme: Oluşturulan modelin ayrıntılı olarak incelenmesi gerekmekte olup, özellikle mevcut iş hedeflerini doğru bir şekilde karşılayıp karşılamadığı analiz edilme ve bu aşamanın sonunda veri madenciliğine dair olan bu sonuçların kullanılıp kullanılamayacağına dair de bir karar alınmalıdır (Deshpande ve Thakare, 2010; Mariscal ve ark., 2010).

– Dağıtım: Modelleme yapılarak üzerinde çalışılan veri madenciliği, mevcut bilgilerin kalitesini arttırmayı amaçlamakta olup, söz konusu çıktıların hedef kitlenin anlayabileceği raporlara dönüştürülmesi işlemi konu alır (Huber ve ark., 2019; Wirth ve Hipp, 2000).



Şekil 1.15. Veri tabanlarında bilgi keşfinde SEMMA modeli

Veri tabanlarında bilgi keşfi için kullanılan diğer modelleme ise SEMMA modelidir. Bu model, veri madenciliği projelerinin anlaşılmasını, organizasyonunun geliştirilmesini ve sorunlara çözüm üretmeye yardımcıdır (Shafique ve Qaiser, 2014). SEMMA modeli SAS Enstitüsü tarafından ortaya atılmış olup, şekil 1.15’de belirtildiği üzere beş aşamadan oluşmaktadır (Azevedo ve Santos, 2008; Shafique ve Qaiser, 2014).

- Örneklem aşamamasında, eldeki veri kümesi içerisinde veriyi temsil edecek pozisyonda yer alan bir örneklem belirlenir, ele alınan bu küçük örneklem aynı zamanda analizin de hızla yapılmasını sağlayacaktır (Azevedo ve Santos, 2008).

- Keşfetme aşamasında, veri kümesi içerisinde beklenmedik eğilimler ve anomaliler aranmaktadır (Shafique ve Qaiser, 2014),

- Değiştirme aşamasında, veri kümesi içinde yer alan değişkenler ve aykırı verilere odaklanılarak, bu verilere dönüşüm uygulanarak değiştirilmekte veya temizlenmektedir (Shafique ve Qaiser, 2014).

- Modelleme aşamasında, güvenilir bir şekilde tahmin yapılabilen modellemenin yapıldığı aşama olup, uygulamaya konu her modelin farklı teknik ve gücü ile birlikte spesifik güçlü kullanım alanları da mevcuttur (Azevedo ve Santos, 2008).

- Değerlendirme aşamasında ise elde edilen bulguların yararlılığı ve güvenilirliğini değerlendirilir ve ne kadar performans sağladığı tahmin edilmeye çalışılmaktadır (Shafique ve Qaiser, 2014).

Veri madenciliği alanında çalışan insanların çoğu, eksiksiz ve doğru olması sebebiyle KDD sürecini kullanmaktadır (Shafique ve Qaiser, 2014). Fakat CRISP-DM ve SEMMA da özellikle profesyonel çalışmalarda aktif olarak kullanılmaktadır. SEMMA süreci, veri madenciliği projelerinin organize şekilde geliştirilmesi ve sürdürülmesine olanak tanıyan, kolay bir süreci ihtiva etmektedir (Azevedo ve Santos, 2008). CRISP-DM, SEMMA’ya göre daha yaygın kullanıma sahip olmakla birlikte tüm bu süreç modelleri, insanlara veri madenciliğinin pratik senaryoları üzerinden nasıl uygulama yapıldığına yardımcı olmaktadır (Shafique ve Qaiser, 2014).

1.7.2. Veri Madenciliği Görev Türleri

Büyük veri yığınları, hayatımızın her alanında uğraştığımız ve aktif olarak bilgiye dönüşmediği sürece pragmatik açıdan bizim için hiçbir anlam ifade etmeyen yapılardır. Bu nedenle, bir çalışmanın amacını belirlemek önemlidir. Çalışmanın bir sonucu olarak, istenen sonuçlara göre, aşağıda belirtilen veri madenciliğinin beş farklı görev türü mevcuttur (Han ve ark., 2011; Larose ve Larose, 2014).

- Keşifsel veri analizi (EDA - Exploratory Data Analysis), aradığımız veya hedeflediğimiz sonuç hakkında hiçbir fikir edinmeden veriler üzerinde araştırma yapılmasını içerir. Bu nedenle, mevcut olağan yaklaşımları değiştirerek, çalışma yapısını ve modeli önyargıdan uzak şekilde ortaya çıkarmayı amaçlayan yaklaşımdır (Han ve ark., 2011).

- Tanımlayıcı modelleme (Descriptive Modeling), tüm verileri tanımlayan ve gerçek dünyadaki olaylar ile bunlarla ilişkili faktörler arasındaki ilişkileri matematiksel olarak tanımlayan bir süreçtir (Deshpande ve Thakare, 2010).

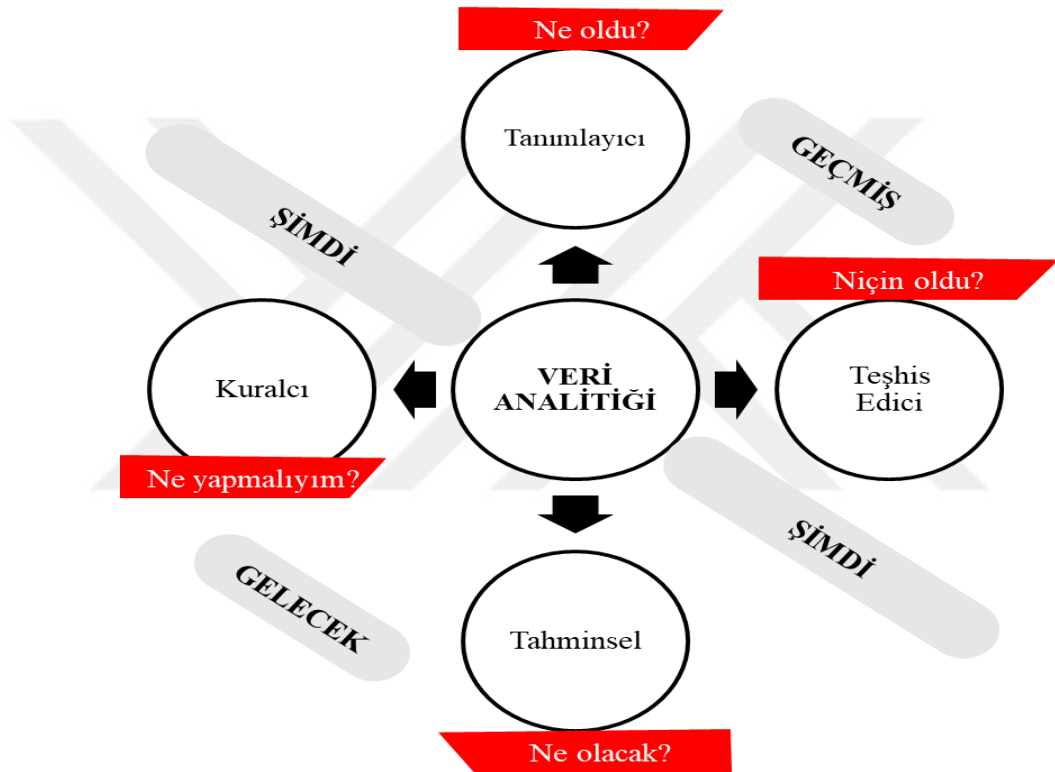
- Tahmini modelleme (Predictive Modeling), veri madenciliğinin mevcut verileri analiz ettiği ve gelecekteki sonuçları etkileyebilecek bir dizi değişkeni belirleyerek, tahmin sürecinde rol oynadığı modeldir. Mevcut bilinen öğelerin yardımıyla, gelecekteki verileri tahmin etmemize olanak sağlar (Han ve ark., 2011; Laxman ve Sastry, 2006).

- Kalıpları ve kuralları keşfetme (Discovering Patterns and Rules), verilerin diğer veri kümelerinden büyük ölçüde farklı olduğu süreç bölgelerinin tanımlanmasıdır, çünkü bu bölgelerde çalışmak yanlış modelleme yapılmasına neden olacaktır (Han ve ark., 2011; Pughazendi ve Punithavalli, 2011).

- Mevcut model yapısına benzer bir model bulmak için veri kümesinde “içeriğe göre alma” (Retrieval by Content) uygulanabilmekte ve söz konusu işlem, çoğunlukla metin ve görüntü veri kümelerinde kullanılmaktadır (Deshpande ve Thakare, 2010).

1.7.3. Veri Analitiği Türleri

Yüksek hacimli veriler arasından gereksinim duyulan bilgilerin sayısal yöntemler kullanılarak elde edilmesi veri analitiği olarak tanımlanmaktadır. Klasik analiz türlerinden en büyük farkı hacmi ve verilerin karmaşıklığıdır. Söz konusu farklılıklar, mevcut analizlerden daha farklı analitik çalışmaları da beraberinde getirmektedir.



Şekil 1.16. Veri analitiği türleri

Şekil 1.16’da gösterildiği üzere literatürde veri analitiğinin dört türü bulunmaktadır, bunlar incelendiğinde (Banumathi ve Aloysius, 2017);

- Tanımlayıcı (Descriptive) modelleme, geçmişte neler olduğunu odaklanır,
- Teşhis edici (Diagnostic) modelleme, tanımlayıcı modellemede yer alan geçmişte ne olduğu sorusuna cevap bularak, bunun yanında veri keşfi yapmakta, söz konusu durumun neden ve niçin olduğunun da ortaya koyulmasını amaçlar,

- Tahminsel (Predictive) modellemede, geçmişteki veriler üzerinden temel eğilimlerin belirlenmesi yani farklı ifadeyle mevcut verilerden gelecek tahmin edilmeye çalışılır,

- Kuralcı/Öngörüsöl (Prescriptive) modelleme ise, geleceğe ilişkin öngörülerimiz karşısında ne yapmamız gerektiğine odaklanır.

Dünyadaki verilerin çoğunun yapılandırılmamış olduğunu ifade edersek, en sık kullanılan yöntemlerin tanımlayıcı ve teşhis edici olduğunu kolayca söyleyebiliriz. Bu kapsamda tüm dünyada salgın boyutuna gelen koronavirüs örnekleminde veri analitiği türleri incelendiğinde;

- Geçmişteki vakalardan hastalığın belirtilerinin genel olarak bilinmesi tanımlayıcı,

- Hastalığın temas durumunda bulaşıcılığı tetiklemesi ve tedavisine yönelik hangi ilaçların iyi geleceğine yönelik çalışmalar teşhis edici,

- Geçmişteki ölüm ve iyileşen insan sayısı rakamlarına bakarak gelecekteki durumu ortaya koyulması tahminsel,

- Tüm mevcut bilgiler ışığında sosyal mesafenin artırılması, sokağa çıkma yasağı, maske kullanılması vb. uygulamalar ise kuralcı/öngörüsöl modellemeye girmektedir.

1.8. Yapay Zekâ

İnsan beyninin işleyişi oldukça karmaşık bir yapıda olup, günümüzde hala bütün detaylarıyla ortaya çıkarılabilmiş bir konu değildir (Namazi, 2018). Bu sebeple de insan zekâsına olan hayranlık, zekânın ne olduğu ve nasıl işlediği soruları temelinde kendine yer bulmuştur (Ertel, 2017). Yapay zekâ (AI) genel olarak insan düşünce ve eylemlerinin simülasyonu olarak tanımlanmakta olup, günümüzde otomasyona dayalı işlemlerin yürütülmesinde ana disiplin olarak anılmaktadır (Mainzer, 2020).

Alan Turing ise yapay zekâyı; “If there is a machine behind a curtain and a human is interacting with it (by whatever means, e.g. audio or via typing etc.) and if the human feels like he/she is interacting with another human, then the machine is artificially intelligent.” (“Perdenin arkasında bir makine varsa ve bir insan onunla etkileşime giriyorsa (örneğin ses veya yazım vb. yoluyla) ve insan başka bir insanla etkileşimde olduğunu hissediyorsa, o makine yapay zekâdır.”) ifadeleriyle açıklamıştır. (Akt. Joshi, 2020)

Yapay zekâ kavramı, bir sorunu mükemmel şekilde çözecek makine modellemesi yerine insan gibi çözecek modelleme yapmaktır (Joshi, 2020). Taklit etmek basittir, bu sebeple de istenen taklitten ötesidir. Fakat insanın yegâne iletişim kaynağı olan konuştuğu veya yazdığı dilin mükemmel şekilde anlaşılması ve buna istinaden çeşitli modellemeler yapılması kolay değildir. Bu durumun en büyük sebeplerinden biri de zaman zaman modellemede örnek alınan insanların bile birbirini yanlış anlaması gösterilebilir.

1.8.1. Yapay Zekâ ve Öğrenme

Yapay zekâ kavramıyla birlikte, canlılar haricindeki nesnelerden insan gibi davranışlar göstermesi beklenmektedir. Söz konusu beklentinin modellemesi ise öğrenme sürecini de beraberinde getirmektedir. İnsandaki öğrenme yapısına bakıldığında, sinir ağları üzerindeki birçok düşünce nöroloji biliminin ilgi alanına girmektedir. İnsan düşündüğü kavramlar hakkında çeşitli kararlar vermekte olup, söz konusu kavramların en uygun çözüm olduğu varsayılmaktadır (Ertel, 2017). Günlük hayatta her konuda olduğu gibi karar vermeden önce de çok farklı yol ve teknikler kullanmış olabiliriz. Bu sebeple de yapay zekânın sabit bir yöntemi kullanmaktan ziyade mümkün olduğunca farklı görev türlerini ihtiva edecek modellemeyi içinde barındırması önemlidir.

Öğrenme süreçleri sonucunda çıkan kararın bir dizi makine öğrenimi modellemesini de beraberinde getirmesi sebebiyle, makine öğrenmesi ve derin

öğrenme kavramları yapay zekâ altında yer almaktadır (Ertel, 2017). Yapılan işlem, her problemin söz konusu problem özelindeki verilerine göre modellenmesidir. Modellemede kullanılan algoritmalar, test kümesiyle sınanarak, en yüksek performans değerlerini bulmak amaçlanmaktadır.

1.8.2. Yapay Zekânın Etkileri

Yapay zekânın giderek yaygınlaştığı günümüz koşullarında, bu alandan etkilenmeyen verimli iş sahası bulmak yakın gelecekte neredeyse imkânsız olacaktır. Günümüzde dahi tekil görevlendirme bazında santraç gibi oyun sektöründe aktif olarak kullanılmakta olup, kullanım maksat ve çeşitliliğinin giderek arttığını söyleyebiliriz.

Bu kapsamda yapay zekâ kullanımının yaratacağı bazı etkileri inceleyecek olursak;

- Makinelerin insanlar gibi davranarak üretim yaptığı fakat insanlar gibi yorulmadığı, maaş almadığı, istirahat etmediği 7 gün 24 saat çalıştığı sistemlerin inşası endüstriyel alanda bir devrimdir. Söz konusu devrim Endüstri 4.0 olarak anılmakta olup, bu kavramın doğmasıyla birlikte siber fiziksel sistemlerin yaygınlaştığı bilinmektedir (Zanero, 2017).

- Siber fiziksel sistemlerin yaygın kullanımıyla üretim maliyetlerinin düşeceği fakat bunun yanı sıra yapay zekâ içeren makinelerin etkinliğinin insanların çalışma alanını kısıtlayacağı ve bu durumda işsizliği arttıracığı değerlendirilmektedir.

- Yapay zekâ temelinde yeni iş kolları türeyecek ve öğrenme için vazgeçilmez olan büyük veri önemini arttıracaktır (Wilson ve ark., 2017).

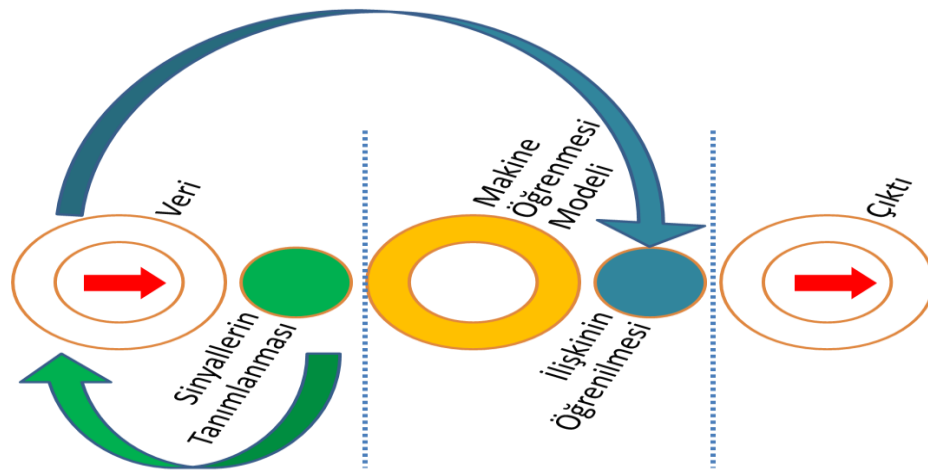
- Veri miktarlarındaki büyüklükler, genel desenin ortaya konularak, olağandışı yani anomali gösteren verilerin tespitini çok daha gerekli kılmaktadır (Ding ve ark., 2018).

– En iyi çözüm mantığında yapay zekânın sürekli olarak insan hayatında olması, insanlarda rasyonaliteyi arttıracak ve bunun yanı sıra kendileri daha az özgür hissetmelerine sebep olacaktır (Puaschunder, 2020).

– Yapay zekânın en büyük kısıtlılığı verilerden öğreniyor olmasıdır. Veri ne kadar büyük olursa olsun, üzerinde yapılacak manipülasyon kişilerin veya şirketlerin saldırılara maruz kalmasını arttıracaktır. Söz konusu durum da siber güvenliğin önemini arttıracaktır.

1.9. Makine Öğrenmesi

Dünyaya gözlerini açtığı ilk andan itibaren insanoğlunun temelde yaptığı davranış taklittir. Örneğin bir bebek düşünelim; bebeğin anneden, babadan ve çevresinden gördüğü kadarını yapmak istemesi taklittir, bu sebeple taklit bir öğrenme şekli olup, taklitte başlayan öğrenme zamanla özgünleşmeye ve genişlemeye başlayacaktır. Yaş ilerledikçe öğrenilen bilgilerin genelde niceliği ve niteliği de artacak, mümkün olduğunca en doğru karar verilmeye çalışılacaktır. Bu durum, günlük hayatta birikim ya da tecrübe olarak adlandırılabilir. Tüm bu öğrenme süreci temelde insanoğlunun farkında olmadan sahip olduğu öğrenme yeteneği sayesinde meydana gelmektedir.



Şekil 1.17. Makine öğreniminin genel çerçevesi

İnsanoğlunun sahip olduğu öğrenme yeteneği, yeryüzündeki diğer nesnelerden ayırt edici bir özellik olmakla birlikte, makine öğrenmesi kavramını da beraberinde getirmiştir. Makine öğrenimi kavramının hayatımıza girmesi, ilk olarak Alan Turing tarafından ortaya atılan “Makineler insan gibi düşünebilir mi? (Can machines think like human?)” önermesiyle başlamıştır (Turing, 1950). Söz konusu sorunun sorulması ile birlikte yapay sinir ağlarına dayalı geliştirilen ilk bilgisayar olma özelliğine de haiz olan SNARC (Stokastik Sinirsel Analog Güçlendirme Hesaplayıcısı) ve Samuel Arthur tarafından geliştirilen satranç oyunu makinelerin insan gibi düşünüp düşünemeyeceği konusunu sınımışlardır (Joshi, 2020). Makine öğrenmesi temelinde yapılan tüm gelişmeler, makinenin yaptığı davranışların, dışardan ilave insan yardımı olmaksızın insan tarafından da yapılabilmesi üzerine kuruludur (Sarkar, 2019). Farklı bir ifadeyle şekil 1.17’de belirtildiği üzere verilerde doğal hatalar olsa bile, belirli model ve kalıpları veriden çıkarım yapabilme yeteneği makine öğrenimidir, bu süreçte öncelikle veriler alınır ve sonrasında veriler içindeki ilişkiler bulunarak, modelin öğrendiklerini çıktı olarak vermesi amaçlanmaktadır (Özdemir, 2016).

Klasik algoritmik yaklaşımlarda, insan faktörünün en iyi cevabı nasıl bulacağını modele söylemesi koşuluyla, en iyi cevap normal bir insanın yapabileceği hesap hızından çok daha hızlı şekilde bulunabilmektedir. Makine öğrenimini diğer klasik algoritmik yaklaşımlardan farklı kılan konu ise modele en iyi cevabı belirtmek yerine örnekler sunarak, ondan uygun model geliştirerek hesap yapmasını istemektir. Klasik algoritmik yaklaşımlarla özellikle büyük veri üzerinden elle işlem yapmak, verinin boyutu da göz önüne alınarak zaman zaman imkânsız olabilmekte olup, bu durumun makine öğrenimi için model tasarlamayı zorunlu kıldığı bilinmektedir. Bu yüzden de makine öğreniminin mükemmel bir sistem olmadığı, fakat farklı öğrenme metotlarıyla insanoğlunun en büyük destekçisi konumunda olduğu ortadadır (Zhang, 2020).

Teknolojik hızla gelişimine paralel olarak, makine öğrenimi kavramı son yüzyılın en popüler konuları arasındadır. İnsanların yaptığı işlerin, çeşitli algoritmalar temelinde bazı sektörlerde makinelerce yapılmaya başlanması, çeşitli sektörlerde de çalışmaların devam etmesi aynı zamanda olayın toplumda sosyolojik ve psikolojik birçok değişimi de beraberinde getireceği açıktır. Hâlihazırda makine öğreniminin

endüstriyel üretim yapan birçok tesiste kullanılmasının yanı sıra, yakın gelecekte ordu yapısını robotların oluşturduğu asker tasarlama çalışmaları ise devam etmektedir. Geline nokta itibariyle Hong Kong merkezli bir şirket tarafından yaşlılara ve ziyaretçilere parklarda yardım etmesi için makine öğrenimi temelinde geliştirilen Sofia adlı robota, 2017 yılında Suudi Arabistan tarafından vatandaşlık verilmiştir. Makine öğrenimi temelinde bir değişimin yaşandığını göz önüne alacak olursak, bu bölümde öncelikle makine öğrenmesi kavramı tanımlanacak, makine öğreniminde literatürde yaygın olarak kullanılan; regresyon, kümeleme, sınıflandırma, kümeleme, ilişkilendirme ve boyut azaltma kavramları sırasıyla açıklanacaktır (Joshi, 2020).

1.9.1. Regresyon

Veri seti içinde yer alan gerçek sayısal değerler üzerinden analiz yapan modellemedir (Schroeder ve ark., 2016). Regresyon, değişkenler arasındaki ilişkiyi tahmin ederek geleceği tahmin etmeye yardımcı olmakta olup, sınıflandırmanın aksine, regresyon sürekli bir çıktı değişkeni döndürmektedir (Lee, 2019; Yang, 2019). Diğer modelleme türlerinden en büyük farkı tahminleme yapabilmesidir. Konuyu bir örnekle açıklayacak olursak, geçmiş yıllarda A ilinde işlenen cinayetlerin, hangi zaman aralığında (X1), hangi konumda (X2), ne kadar miktarda (X3) ve ne kadar ölü/yaralı (X4) ile sonuçlandığının analize dökülmesi sonucu, A iline ait yakın projeksiyonda suç tehdit durum haritası yapılmasına olanak sağlanılmaktadır. Buradaki girdiler olan X1, X2, X3 ve X4 veri yapısının bağımlı nitelikleri olup (Girdiler regresör olarak da bilinmektedir.), A iline ait suç durumunu (Y) açıklayan etmenlerdir. X temelinde yer alan girdilerin sonucunun, Y bağımlı değişkenini etkilemesi sonucu ortaya çıkan matematiksel ilişki regresyondur.

Regresyonun makine öğrenmesini kullanmada üç önemli metriği bulunmakta olup, bunlar; ortalama mutlak hata (MAE) (mean absolute error), ortalama kare hatası (MSE) (mean squared error) ve kök ortalama kare hatası (RMSE) (root mean squared error) olarak bilinmektedir. Bunların hepsi de tahminleme yapmada yardımcı olup, cinayet örnekleminde formülasyonu verildiği gibi ortalama mutlak hatayla yapılan

hataların ortalaması alınarak, yapılan işlemin ne kadar yanlış olduğu ortaya konulmakta, ortalama kare hatasıyla yapılan hataların karesi alınarak pozitif ve negatif ayrımı gözetmeksizin daha doğru yanlışlama bilgisi ortaya çıkmaktadır (Özdemir, 2016). Kök ortalama kare hatasıyla ise ortalama kare hatası kareköke alınmakta olup, yorumlaması çok daha kolay çıktılar elde edilmektedir (Şekil 1.18) (Özdemir, 2016).

$$\text{Ortalama Mutlak Hata (MAE)} = \frac{1}{n} \sum_{i=1}^n |Y_{tahmin} - Y_{gözlemlenen}|$$

$$\text{Ortalama Kare Hatası (MSE)} = \frac{1}{n} \sum_{i=1}^n (Y_{tahmin} - Y_{gözlemlenen})^2$$

$$\text{Kök Ortalama Kare Hatası (RMSE)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_{tahmin} - Y_{gözlemlenen})^2}$$

Şekil 1.18. Regresyon metriklerinin formülasyonu

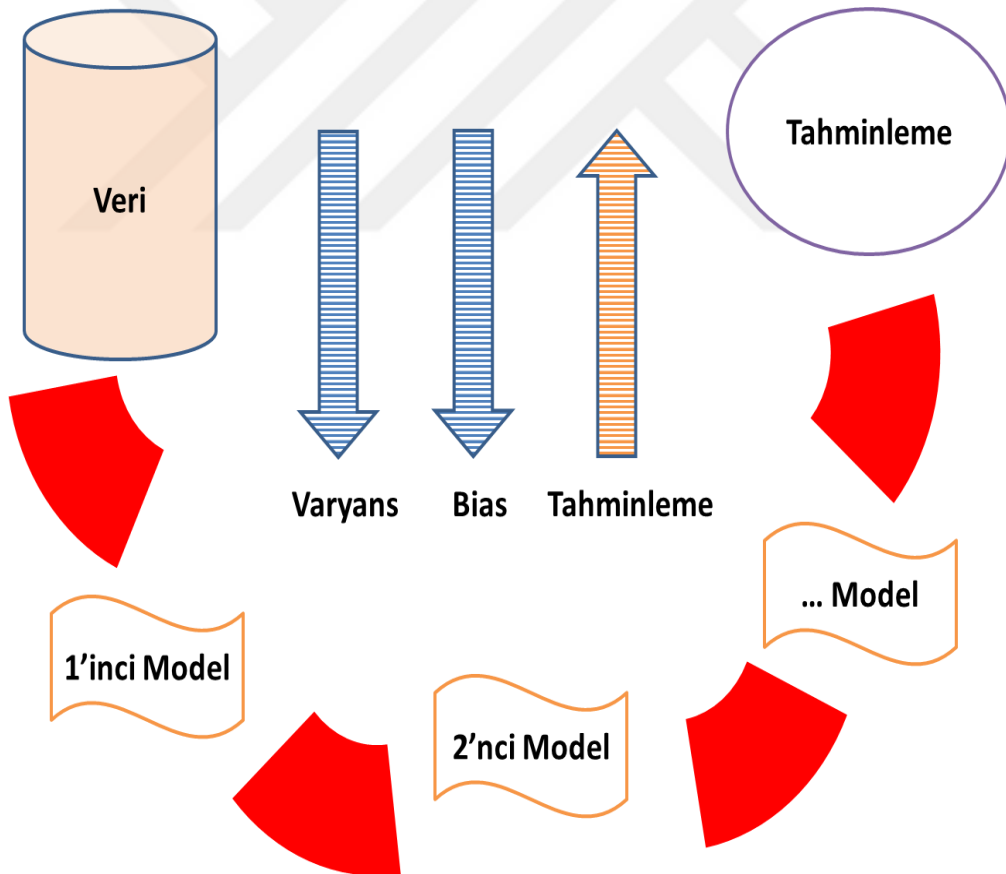
1.9.2. Sınıflandırma

Makine öğreniminde sınıflandırma, gözlemlenen kategorileri içeren eğitim verisi kümesine dayanarak yeni bir gözlemin hangi kategori kümesine ait olduğunu belirler (Lee, 2019). Regresyon sayısal olarak yer alan çıktı bağımlı değişkeni, sınıflandırmada kategorik düzeyde yer almaktadır. Sınıflandırma ile yapılmaya çalışılan işlem, bağımsız X değişkenlerinden yola çıkarak, “yaşlı”, “genç”, başarılı” veya “başarısız” gibi kategorik Y çıktı değişkeninin tahmin edilmeye çalışılmasıdır. Sınıflandırmanın amacı veri içinde yer alan herhangi bir ögenin, verinin alt kümesine ya da önceden belirlenmiş sınıflarla eşleştirmesini yapmaktır. Özellikle sınıflandırma kapsamındaki denetimli öğrenmelerde, verilerden örüntü keşfi yapılarak, yeni eklenen hususlarla ilgili sınıflandırma tahmini yapılabilir. Benzeri bir durum suç verileri özelinde de “bir şehrin ya da mahallenin “güvenli” ya da “tehlikeli” olarak sınıflandırmasında kullanılmaktadır. Genetik algoritmalar başta olmak üzere Naive Bayes, lojistik regresyon, destek vektör makineleri, bulanık mantık, kNN (k-en yakın

komşu); ses tanıma, çağrı yönlendirme ve metin sınıflandırma gibi birçok uygulamada sınıflandırma metodunu aktif olarak kullanmaktadır (Miller, 2017).

1.9.3. Toplanma (Topluluk Öğrenimi)

Toplanma (Topluluk öğrenimi), tek bir makine öğrenimi algoritması kullanmak yerine optimum sonucun elde edilmesi için birden fazla makine öğrenimi algoritmasının kullanıldığı denetimli bir öğrenmedir (Sagi ve Rokach, 2018). Bu yönüyle tek başına herhangi bir modelden ziyade, daha doğru bir model ortaya koymak için çoklu öngörücü modelleri birleştirme işlemidir (Laine ve Aila, 2016; Özdemir, 2016).



Şekil 1.19. Topluluk öğrenmesinde model yapısı

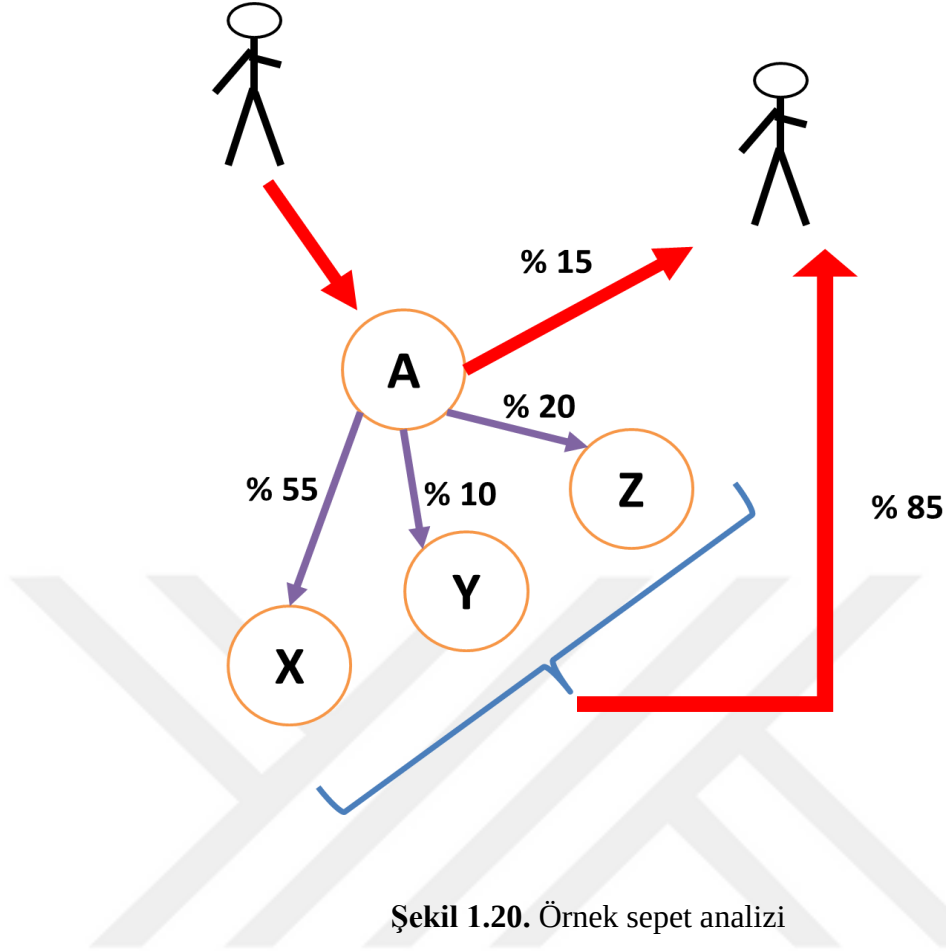
Şekil 1.19’da gösterildiği üzere farklı modellemelerin eğitiminden sistemi geçirerek, varyans (değişkenlik) ve bias (dizgesel hata) değerlerinin çok daha düşük olduğu güçlü tahminlemeler ortaya koymak amaçtır. Rastgelle orman, XGBoost, AdaBoost ve torbalama bu yöntemi temsil eden karma algoritmik örneklerdir.

1.9.4. Kümeleme

Kümeleme kavramı, verinin belirli yakınlık kriterleri baz alınarak, “küme” adını verilen gruplara ayrılmasıdır (Alelyani ve ark., 2018). Yapılan ayrıştırma sonrası aynı küme gruplamasında yer alan verilerin birbirine benzerlik göstermesi, farklı gruplarda olanların da çok benzerlik göstermemesi beklenmektedir (Özdemir, 2016). Benzerlikler mesafe ölçümüne göre yapılmakta olup, mesafe ölçümünde ise en çok öklit kullanılmaktadır. Denetimsiz öğrenme yöntemi olup, genelde etiketlenmeyen veri yığını içinde anlamlı bilgi çıkarımını yapmaktadır. Kümeleme benzer veri noktalarını sezgisel gruplara ayırmaya yardımcı olur (Lee, 2019). Ses ve görüntü işleme, konuşma tanıma, sık yapılan telefon arama, mesaj ve veri kullanımı, müşteri tasnifleme gibi birçok konunun yanı sıra özellikle sosyal ağ analizi maksatlı aktif kullanılmaktadır. Kohonen Yapay Sinir Ağları, Canopy, Mean Shift, kMeans, Bulanık C-ortalamlar, Latent Dirichlet Allocation, k-medoids ve MinHash kullanılan kümeleme algoritmalarına örnek olabilir (Mukhopadhyay ve ark., 2013).

1.9.5. Birliktelik Kuralları Yönetimi

Veri kümesi içerisinde yer alan değişkenlerin aralarında olan ilişkiden bulunan kuralların makine tarafından öğrenilmesi ilişkilendirme yaklaşımıdır (Abdel-basset ve ark., 2018; Ngai ve ark., 2009). Bu yöntem veri madenciliği için de sıkça kullanılmaktadır. Kümelemedeki benzerlik için olan mesafe ölçümünden farklıdır, sadece kural çıkarımı yapar.



Örnek şekil 1.20’de görüldüğü üzere B marketinin satış rakamları incelendiğinde A ürününü alan kişi; %55 oranında X ürününü, % 10 oranında Y ürünü, %20 oranında ise Z ürününü aldığı, % 15 oranında da başka bir ürün almadan marketten çıkmaktadır. Örnekte A ürününü alan kişilerin genelde X ürününü alması durumu kümelemede olduğu gibi iki ürünün benzer olduğunu değil, sadece birinin diğer ürün yada ürünlerle korelasyon kurduğunun göstergesidir. Özellikle sepet analizi, suçlu profillemeye gibi konular ilişkilendirmenin ilgi alanına girmektedir.

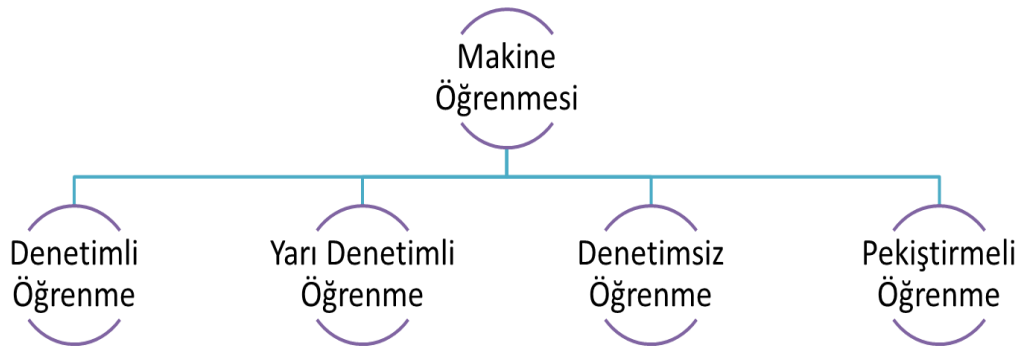
1.9.6. Boyut Azaltma

Son yıllarda veri kaynaklarına sürekli olarak yenileri eklenmiş ve veriler içindeki detaylar da artmıştır. Özellikle e-ticaret ile müşteriye ilişkin bilgilerin depolamasının sürekli artması ve resim boyutları başta olmak üzere temel veri düzenlerindeki kalitenin artmasına paralel olarak veri boyutları da artmıştır. Artan veri

miktariyla birlikte, yüksek boyutlu verilerden istenilen veriyi filtreleme ve bunları daha küçük boyuta indirme çok daha önemli hale gelmiştir. Veri setinin x boyutlarını, y boyutuna ($y < x$) düşürmek için y boyutunu doğrudan tanımlanabilir (filtrelenebilir), mevcut çoklu veri yapısını iyi temsil eden boyutlar (boyutların ağırlıklı ortalamaları) veya yeni veri boyutlarının bir kombinasyonu ortaya konulabilmektedir. Yüksek boyutlu verilerden gereksiz ve alakasız özelliklerin kaldırılması, madencilik performansının ve anlaşılabilirliğinin geliştirilmesine ve öğrenme doğruluğunun artırılmasına yardımcı olmaktadır (Jindal ve Kumar, 2017). Temel bileşen analizi (PCA) (Principal component analysis) de yapısı itibarıyla verileri temsil etmede fazlalığı azaltmayı amaçladığından aynı zamanda boyut azaltma içinde yer almaktadır (Yang, 2019). Boyut azaltma algoritmalarının en yaygın kullanımı görüntü işlemedir.

1.10. Makine Öğrenmesi Yöntemleri

Makine öğrenmesi algoritmaları denetimli öğrenme (supervised learning) ve denetimsiz öğrenme (unsupervised learning) başta olmak üzere genel olarak iki temel yaklaşımda öğrenim gerçekleştirirler (Sarkar, 2019). Söz konusu bu yaklaşımların yanı sıra literatürde pekiştirmeli öğrenme (reinforcement learning), yarı denetimli öğrenme (semi-supervised learning), çevrimiçi öğrenme (online learning), aktif öğrenme (active learning) ve transdüktif akıl yürütme (transductive inference) gibi diğer bazı yöntemlerde mevcuttur (Mohri ve ark., 2018).



Şekil 1.21. Makine öğrenmesi yöntemleri

Farklı makine öğrenim yöntemleri mevcut olsa literatürde en yaygın kullanılanları şekil 1.21’de gösterildiği üzere denetimli, denetimsiz, yarı denetimli ve pekiştireçli öğrenme olduğundan, bu yöntemler alt maddelerde açıklanmaktadır (Sarkar, 2019).

1.10.1. Denetimli Öğrenme

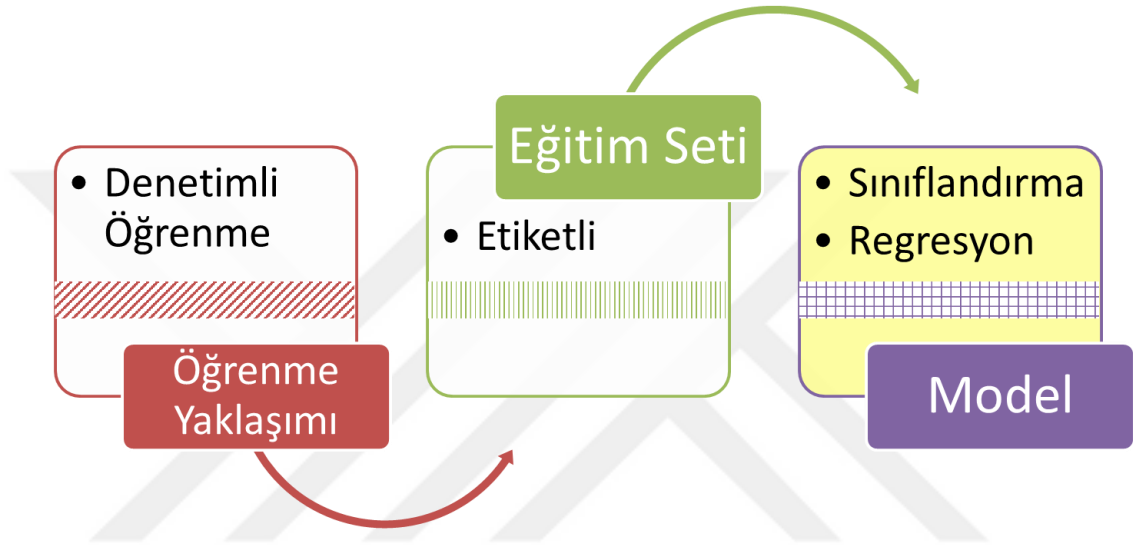
Girdi ve çıktıların sistemde tanımlı olduğu, modelde yer alan girdilere makine öğrenimiyle anlamlı çıktının verilebildiği öğrenme şekli denetimli öğrenme olarak tanımlanmaktadır (Özdemir, 2016). Bu kapsamda e-posta kutularına gelen spam maillerin tespiti denetimli öğrenmenin bir örneğidir (Mohri ve ark., 2018). Farklı bir örnekle ise denetimli öğrenme modelleri yardımıyla, bir kişinin ders çalışmasıyla ilgili hususlar (günlük ders çalışma saati, başlangıç başarı durumu, derse katılımı vb.) ile o kişinin sınavlarda başarılı olması (hedef değişken) arasındaki ilişki bulunmaya çalışabilir.

Çizelge 1.1. Makine öğrenmesine dayalı farklı dillerden çeviri örneği

İngilizce Cümle	Çevrildiği Dil	Çeviri Cümle
Big data is a very important issue.	Türkçe	Büyük veri çok önemli bir konudur.
	Rusça	Большие данные - очень важная проблема.
	Almanca	Big Data ist ein sehr wichtiges Thema.
	Fransızca	Les mégadonnées sont un problème très important.
	İtalyanca	I big data sono un problema molto importante.
	Çince	大數據是一個非常重要的問題。
	Arapça	البيانات الكبيرة هي قضية مهمة للغاية.

Denetimli öğrenmenin yapılabilmesi için ilk işlem olarak, girdi ve çıktıların bulunduğu bir eğitim setinin hazırlanması gerekir. Farklı bir ifadeyle girdilerin

etiketlenebilmesidir. Eğitim sistemi üzerindeki her veri, denetimli olarak makineye öğretilir. Makineye veriler arasındaki ilişki öğretildikten sonra daha önce sisteme girdi olarak tanımlanmayan girdiler hakkında çıkarımlarda bulunabilmesi beklenmektedir (Joshi, 2020). Bu kapsamda kelime tanımlaması ve aralarındaki ilişkiyi başarılı şekilde tanımlayan “Google Translate” uygulamasının örnek cümle için makine öğrenimine dayalı yaptığı çeviriler çizelge 1.1’de görülmektedir.

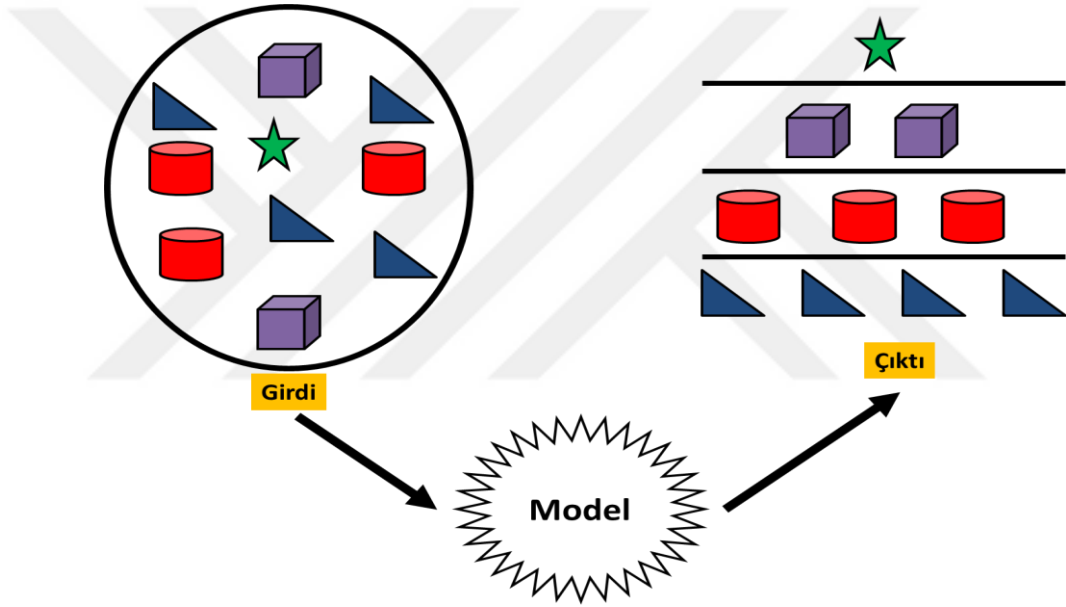


Şekil 1.22. Denetimli öğrenmede eğitim seti ve model

Denetimli öğrenmede, şekil 1.22’de gösterildiği akış ile geçmişe yönelik etiketli veriler üzerinden geleceğe yönelik çıkarımlarda bulunulmaya çalışılır. Sınıflandırma ve regresyon, denetimli öğrenmede en sık kullanılan öğrenme yaklaşımlarıdır. Denetimli öğrenmede etiketlenen veya denetlenen veri miktarı arttıkça, model de gittikçe daha doğru çıkarımlar yapabilmekte, farklı bir ifadeyle öğrenme algoritmaları veriyi daha iyi tanıyacağından tahminlemede yer alan başarı oranı yüksek olacaktır. Bu sebeple de etiketli veri miktarının hacmi denetimli öğrenmede oldukça önemli olup, bu durum aynı zamanda yapılan prosesi daha zor hale getiren dezavantaj unsuru olarak karşımıza çıkmaktadır.

1.10.2. Denetimsiz Öğrenme

Etiketsiz eğitim seti içerisinde örüntüler keşfederek, geleceği tahminden ziyade benzerlikleri ve farklılıkları ortaya koyarak, durumu anlamayı amaçlayan öğrenme biçimi denetimsiz öğrenmede olarak tanımlanmaktadır. Denetimsiz öğrenmenin, kümeleme ve boyut azaltma olmak üzere iki amacı vardır (Mohri ve ark., 2018). Kümeleme, benzer şekilde davranan gözlem gruplarını keşfederek, bunları gruplamayı; boyut azaltma ise verilerdeki örüntüleri kullanarak, verileri daha küçük bir formatta temsil edecek örneklem üzerinde çalışmaktadır (Özdemir, 2016).



Şekil 1.23. Denetimsiz öğrenmede girdi ve çıktı

Denetimsiz öğrenmede, eğitime ihtiyaç duyulmamasından dolayı etiketli veri ihtiyacı bulunmaması, yapılacak işlemi oldukça kolaylaştırmaktadır. Fakat denetimli öğrenmede sıkça kullanılan sınıflandırma ve regresyon ile tahminleme yapılabilmekte, bu tahminlerin başarı oranı hesaplanarak objektif ve somut bir çıktı ortaya konulabilmektedir. Denetimsiz öğrenmede ise şekil 1.23’de görüldüğü üzere, öğrenim işlemi sonrası benzerlik ve farklılıklar ortaya konulmakta olup, bu durum denetimsiz öğrenmeyi insan yorumlamasına yani öznelliğe açık hale getirmekte ve performansının nicel olarak değerlendirilmesini daha zorlaştırmaktadır.



Şekil 1.24. Denetimsiz öğrenmede eğitim seti ve model

Denetimsiz öğrenmede, şekil 1.24’de gösterildiği akış ile eğitim seti sadece etiketsiz girdi olarak belirtilen değişkenlerden oluşmakta olup, algoritmalar yardımıyla örüntüleri bulmak amaçtır. Modellemede net doğru ya da yanlış çıktı bulunmayıp, kendi içinde anlamlı bir bütün oluşturmasının yeterli olması veri madenciliğine oldukça benzemektedir (Bell, 2014).

1.10.3. Yarı Denetimli Öğrenme

Yarı denetimli öğrenmede genelde eğitim seti içerisinde yer alan verilerin büyük bir kısmı etiketsiz iken, etiketsizlere nazaran daha az bir kısmı ise etiketli olarak karşımıza çıkmaktadır (Sarkar, 2019).



Şekil 1.25. Yarı denetimli öğrenmede eğitim seti ve model

Yarı denetimli öğrenme, etiketli ve etiketsiz verilerin birleşiminde kaynaklı olarak, şekil 1.25’de görüldüğü üzere her iki öğrenme yöntemi olan denetimli ve denetimsiz öğrenmeyi birlikte aktif olarak kullanmaktadır (Mohri ve ark., 2018). Etiketli ve etiketsizi birlikte kullanmanın sebebi, verileri etiketlemenin maliyetli ve zahmetli olmasıdır. Regresyon, sınıflandırma ve sıralama yarı denetimli öğrenmede en fazla kullanılan öğrenme yaklaşımlarıdır (Mohri ve ark., 2018).

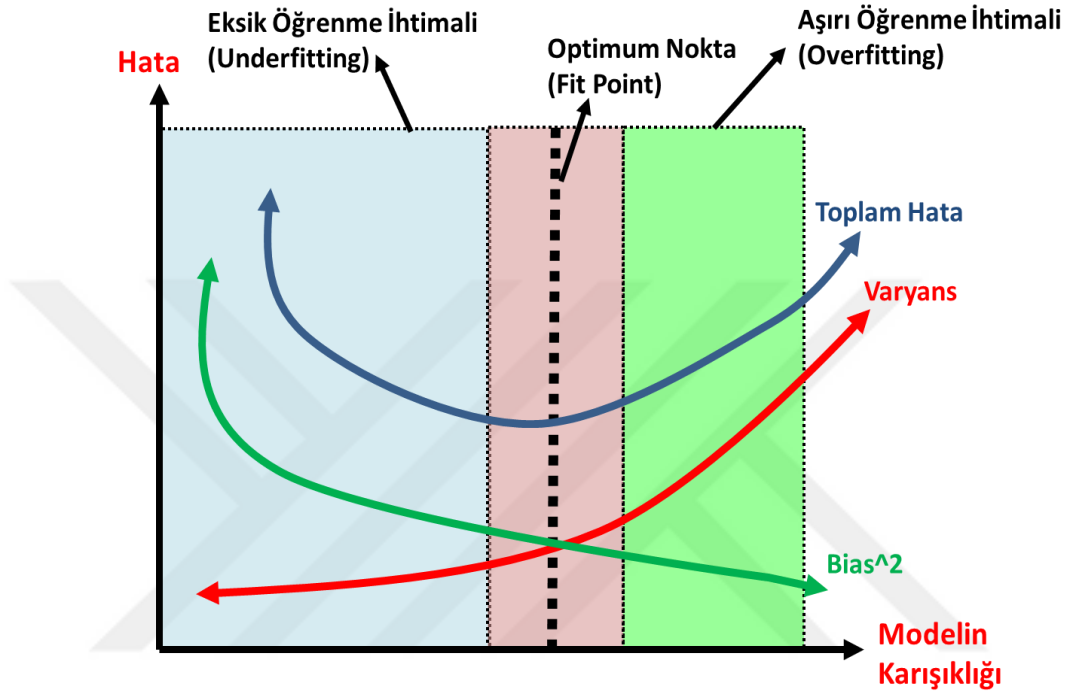
1.10.4. Pekiştirmeli Öğrenme

Pekiştirmeli öğrenme (takviyeli) yöntemi ile model, seçimde bulunarak yaptığı eylemin karşılığı olumlu ise ödüllendirilmektedir (Sutton ve Barto, 2018). Genel algoritma kısıtlamalar dahilinde maksimum ödül kazanma anlayışıdır. Bu sebeple de söz konusu yöntemde eğitim ve test aşamaları aynı süreçte meydana geldiğinden, denetçiye yani etiketlenmeye ihtiyaç vardır. İyi ya da kötü gibi etiketlemeler (Ödülün ikili sınıflandırma (binary classification) kapsamında olması gerekmemektedir.), sisteme istenilen modeli öğretmeye çalışmakta olduğundan, pekiştirmeli öğrenmede tek bir sonuçtan ziyade süreç önemlidir. Sonuç sürekli değişecektir, önemli olan sürecin öğrenilerek, en fazla ödülün alınmasıdır. Pekiştirmeli öğrenme, etiket kullanımından dolayı denetimli öğrenmeye benzemekte olup, gelecekte daha iyi kararlar alabilmek adına geçmişten ders almaktadır, fakat denetimli öğrenmeden ayrıldığı en önemli nokta ise ödülün yani öğrenmenin sürecin içinde olmasıdır (Mohri ve ark., 2018). Özellikle internet ve uydu bağlantılı navigasyon sistemleri ve sürücüsüz araç uygulamalarının çoğu pekiştirmeli öğrenme mantığında çalışmaktadır.

1.11. Makine Öğrenmesi Algoritmaları

Makinelerin örnekler ve deneyimlerden yararlanarak, karmaşık gerçek dünya üzerinde insanların uzun uğraşları sonucu yaptığı eylemleri öğrenme becerisini kazanması makine öğrenimidir. Makine öğrenimi işleminin arkasında büyük bir matematik yatmaktadır. Geliştirilen çeşitli algoritmalarla modele en iyi eğitimin

verilmesi amaçlanmaktadır. Söz konusu eğitimde, en büyük problem modelin sorunu öğrenme yerine ezberlemedir, bu genelleştirme hatasını ortaya çıkaracaktır. Genelleştirme hatası olarak tanımladığımız toplam hata; varyans (değişkenlik, yayılım), bias ve indirgenemez hatanın (irreducible error) toplamıdır (Özdemir, 2016).



Şekil 1.26. Makine öğrenmesinde hatalar

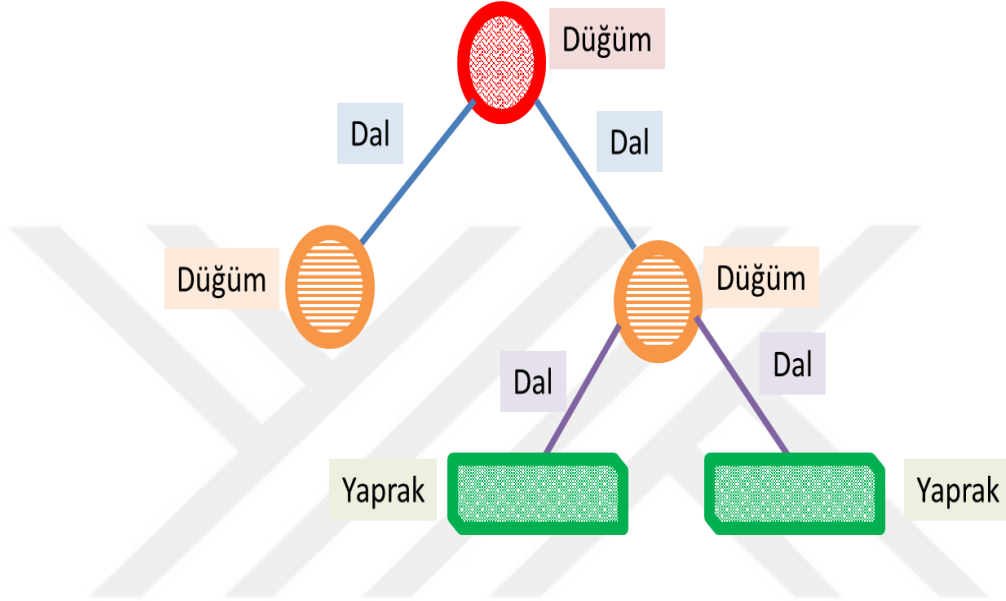
Bias ile ortaya konulan modelin ne kadar yanlış olduğu ölçülmekte, varyans ile modelin tahmin ettiği verilerin gerçek veri değerlerinin etrafında ne kadar dağılım gösterdiğine bakılmakta ve indirgenemez hata ile de verinin yeterince temizlenmediği ortaya konulmaktadır. İndirgenemez hataya veri analizine başlanması sonrası geriye dönme dışında pek yapılacak bir şey yoktur. Bias ile varyans ise analizde kontrol edilebilmekte olup, şekil 1.26'da görüldüğü üzere farklı eğilim yapılarında olması sebebiyle bias/varyans ikilemi (Bias/Variance tradeoff) sorununu beraberinde getirmektedir. Burada en önemli nokta bias ile varyansın, kesişimlerinin en düşük olduğu optimum noktanın bulunmasıdır. Diğer türlü modellemedeki öğrenme, aşırı uyum (overfitting) veya eksik uyum (underfitting) sorunuyla karşı karşıya kalabilmektedir. Aşırı uyumun önüne geçmek için, verinin test aşaması veriden

ayrılıp, karışıklı matrisi ile öğrenme test edilebilir veya kalan eğitim seti K-Fold çapraz doğrulama yöntemi ile k kadar bölünerek, k kadar doğruluk (accuracy) elde edilerek, bunların ortalaması alınabilmektedir (Özdemir, 2016). Grid araması yöntemi ile de K-Fold ile elde edilen değerlerin doğruluk oranı arttırılabilir. Bu sebeple, öğrenim işlemi birbirinden oldukça farklı belirli algoritmalar temelinde yapılmaktadır. Farklı algoritmik yaklaşımlarla hiperparametre optimizasyonu yapılmasıyla kayıplar (loss function) en aza indirilmek istenmekte ve farklı modellemelerle öğrenmenin en iyi gerçekleştiği şartlar ortaya konulmaya çalışılmaktadır. Söz konusu öğrenme yaklaşımları arasında en çok kullanılarak, alt maddelerde açıklananlar aşağıda listelenmiştir.

- Karar ağaçları,
- Rastgele orman,
- Naive Bayes,
- K-En yakın komşu,
- K-ortalamalar,
- Destek vektör makineleri,
- Lineer regresyon,
- Lojistik regresyon,
- Apriori,
- Temel bileşen analizi,
- Lineer diskriminant analiz,
- Bellek temelli nedenleme,
- Torbalama,
- Artırma,
- Gradyan artırma.

1.11.1. Karar Ağaçları

Karar ağaçları, regresyon veya sınıflandırma gerçekleştirebilen denetimli modellerdir (Özdemir, 2016). Bu sebeple CART (Sınıflandırma ve regresyon ağacı) olarak da bilinmektedir (Joseph ve ark., 2016).



Şekil 1.27. Karar ağaçlarının yapısı

Karar ağacı Şekil 1.27’de görüldüğü üzere; kök düğümü, dalları ve yaprak düğümlerini içeren bir yapı olup, veri üzerindeki değişkenleri kullanarak hedeflenen değeri tahminlemede kullanılan dallı ağaç görünümlü modeldir. Ağaçtaki en üstteki düğüm kök düğümdür. Söz konusu modellemede, veriler kök ile başlamakta, düğüm olarak isimlendirilen karar verme noktalarından geçerek dallara ayrılmakta ve sonrasında da yapraklara ulaşmaktadır. Regresyon ağaçları için algoritma düğüm noktalarında olası bölünmeleri yani dal ve yapraklara ayrılmaları inceler ve ağacın olası en düşük ortalama kare hatasına (MSE - mean squared error) sahip olacak şekilde bölünme gerçekleşir. Söz konusu bölünme noktasının en iyi şekilde tespit edilmesi önemlidir. Sınıflandırma ağaçları için, algoritma çok benzer olmakla birlikte, en büyük fark optimizasyon yaptığımız metriktir. MSE sadece regresyon problemleri için kullanılmakta olduğundan, sınıflandırma ağaçları regresyon ağaçlarından farklı olarak genelde gini indeksi veya entropi üzerinde optimize yapılmaya çalışılır (Özdemir,

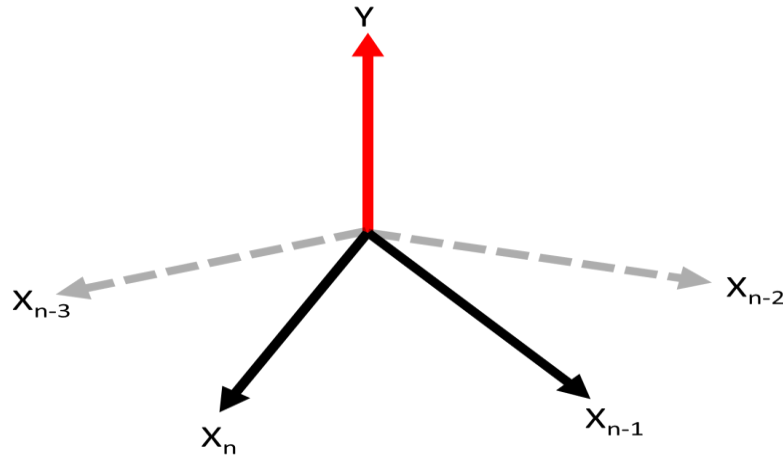
2016). Entropi düzensizliğin ölçülmesi olup, termodinamiğin ikinci yasası olarak kabul edilmektedir, Gini indeksi ise sıklık dağılımı içinde eşitsizliklerin genel dağılımını bulmak için kullanılan bir yöntemdir. Entropi formülasyonunun gösterimi kapsamında, bir veri setinin kaç kategoriden oluştuğu X ise, Y 'nin de sınıf verilerini tanımladığı varsayımıyla, herhangi bir sınıfın olasılığı (O_i) ve entropi (S) formülü aşağıda belirtilmiştir (Şekil 1.28).

$$O_i = X_i / |Y|$$

$$\text{Entropi (S)} = - \sum_{i=1}^n \log_2(O_i)$$

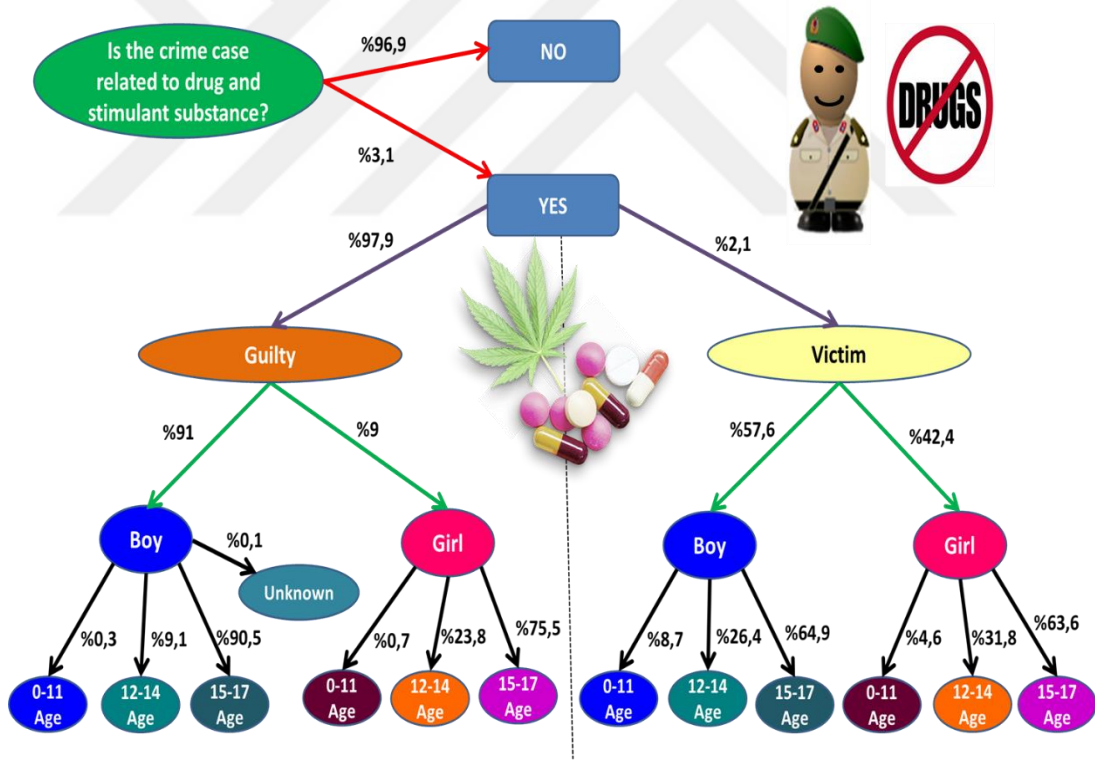
Şekil 1.28. Sınıf ve entropi formülasyonu

Karar ağacında her düğüm bir niteliği tasvir eder, her dal dikkate alınan bir testin sonucunu gösterir ve her bir yaprak da düğüm için sınıflandırılan bir etiketi içinde barındırmaktadır (Dubey ve ark., 2016). Dallar, düğüm sonucu oluşan kararların çeşitliliğine göre ayrılmakta olup, her dal ve yaprak sınıflandırılan ya da sayısal olarak atanan bir değere sahip olduğundan asla birbirine karışmaz (Bell, 2014).



Şekil 1.29. Karar ağacında girdi alanı

Şekil 1.29’da görüldüğü üzere karar ağacı, sınıflandırma için kullanılan tahmin modeli olup, girdi alanını her hücrenin bir sınıfa ait olduğu hücrelere ayırmaktadır. Bölümlemesi, $Y = X (x_1, x_2, \dots, x_n)$ gibi bir test dizisi olarak temsil edilir; burada Y , belirli bir test sınıfının ve $x_1, x_2, \dots, x_n$ ’nin sonucudur (Dubey ve ark., 2016). Farklı bir ifadeyle bağımsız X parametreleri, bağımlı Y değerini ortaya koymaktadır. X parametresi, tahminsel bir işlev sağlayacak kümülatif dağılım veya olasılıksal girdiler olmalıdır. Ağaç, kök düğümünden yaprak düğümüne kadar orijinal kümenin belirli bir sınıfına ait niteliklere göre alt kümelerle bölünmesiyle alt dal ve yapraklara ayrılır. Bu süreç “Karar ağaçlarının yukarıdan aşağıya indüksiyonu” olarak Greedy Algoritması şeklinde bilinmektedir (Dubey ve ark., 2016). Söz konusu yapının yukarıdan aşağıya doğru olmasından dolayı “böl ve fethet” (“divide and conquer”) olarak da bilinir (Mohri ve ark., 2018).



Şekil 1.30. Farklı etiket yapılarını barındıran karar ağacı modellemesi (Ateş ve Bostancı, 2018)

Karar ağaçlarının, eğitilmesi ve değerlendirilmesi hızlıyken ile yorumlanması ve anlaşılması da diğer algoritmalara kıyasla daha kolay olmaktadır. Fakat veriler

içindeki karmaşıklık arttıkça ortaya çıkan fazla dal ve düğümlerden kaynaklı anlaşılabilirlik ve yorumlanabilirlik seviyesi düşecektir. Ayrıca karar ağacı modeli, şekil 1.30'da olduğu üzere sayısal verilerden ziyade, etiketlenmiş kategorik verilerde kullanıma imkân sunmaktadır. Karar ağacının CART, ID3, C4.5, SLIQ, SPRINT, CHAID ve QUEST gibi birçok çeşidi bulunmakta olup, bunlarla klasik istatistiksel yöntemlerden daha verimli sonuçlar elde etmek mümkündür (Danjuma, 2015; Hemalatha ve Megala, 2011).

1.11.2. Rastgele Orman

Rastgele orman algoritması, birden çok makine öğrenimi algoritmasının biraraya getirilerek, yeni bir algoritmanın elde edildiği topluluk öğrenme modeli olarak ifade edilen modellemelerinden biridir (Qiu ve ark., 2017). Genel olarak amacı ise, birden çok karar ağacı modellemesi ile daha dayanıklı ve doğru modellemenin elde edilmesi olup, varyansı azaltıp, potansiyel aşırı uyumun önlenmesi ile topluluk temelli karar ağacı olarak doğruluğun artırılmasıdır (Yang, 2019).

Çalışma mantığı olarak, öncelikle eğitim verisi belirlenen ölçüde parçalara ayrılır. Bu veri setleri üzerinde karar ağacı oluşturacak olanlar “Out of Bag” olarak tanımlananlardır. Geri kalan veri seti ise hataları test edebilmek için kullanılmaktadır. Karar ağacı oluşturma amacıyla belirlenen eğitim verisi üzerinden rastgele n tane örneklem alınarak, yeni bir veri seti üzerinden karar ağacı modellemesi çıkarılır. Sonrasında k kadar aynı işlem tekrar edilir. Bunların sonucunda ise k kadar karar ağacının oluşması beklenir. Öğrenim gerçekleştikten sonra test verisi üzerinden yapılacak sınıflama veya tahminleme k kadar olan karar ağacı üzerinden yapılarak ortaya konulacaktır (Özdemir, 2016). Her biri farklı eğitim verisi üzerinden öğrenim yapan bir toplu modellemedir. İçerisinde barındırdığı her bir karar ağacı kendi kararını vermekte olup, sonuçlar diğer karar ağaçlarından farklı olabilir. Birden çok algoritma kullanması sebebiyle toplu öğrenme sağlamakta ve ortaya çıkan sonuç genelde daha hızlı ve yüksek doğrulukta olmaktadır (Joshi, 2020). Klasik karar ağaçları modellemesinde kullanıldığı üzere, bunda da genelde CART algoritması

kullanılmaktadır. Karar ağaçlarında ortaya çıkacak sonuçlar, optimum sonucu elde etme mantığıyla olacak, en iyi düğüm ve dal noktasının bulunmasında klasik karar ağacı modellemesinde bahsedilen yöntemler kullanılacaktır.

Klasik karar ağacı yapılanmasına göre toplu öğrenme sağladığından dolayı güvenilirlik geçerliliğinin çok daha fazla olması ve kolay işlem yapılabilmesi avantajken, bunun yanı sıra örneklemin tesadüfi seçilmesi ve farklı örnek yeni veri setlerine uyumda yaşanan zorluklar ise dezavantaj olarak yansımaktadır.

1.11.3. Naive Bayes

Naive Bayes teoremine göre çıktıyı tanımlayan girdilerin birbirinden bağımsız ve eşit öneme sahip olduğu varsayımı mevcuttur (Yang, 2019). Bu açıdan bakıldığında oldukça basit gibi gözüксе de gerçek hayata uyarlanan birçok örnekte başarılı ve hızlı sonuçlar alınmaktadır. Girdi verilerinin bağımsız olması, her niteliğin belli sayıda girdiden etkilendiği ve aralarında önem derecesi olmadığı anlamına gelmekte olup, bu durum normalde uygulamada imkânsız yakındır. Naive Bayes algoritması, tanımlayıcı ve tahmin edici sınıfta yer almakta olup, olasılık ve gözlemlere ihtiyaç durmaktadır (Mohammed ve ark., 2016). Verinin oluşumuna ilişkin varsayımlarda bulunmakta ve ortaya koyduğu varsayımlar üzerinden olasılık modellerini belirleyerek, etiketli eğitim setini de kullanarak modelin parametrelerinin tahminini yapmaktadır. Basit anlamda yapılan bayesian modellerinde, modelin öğrenimi aşamasında her sonucun çıkma olasılığı hesaplanır ve sonrasında yeni gelen değerler için de olması gereken sınıf tahmin edilmeye çalışılmaktadır. Sınıflandırma olduğu için de denetimli öğrenmede aktif kullanılmaktadır (Özdemir, 2016).

$$P(X|Y) = \frac{P(Y|X)(X)}{(Y)}$$

Şekil 1.31. Naive Bayes formülasyonu

Şekil 1.31'deki Naive Bayes teoreminin formülasyonunu inceleyecek olursak;

$P(X | Y)$; Y'nin gerçekleşmesi durumunda X'in meydana gelme olasılığı,

$P(Y | X)$; X'in gerçekleşmesi durumunda Y'nin meydana gelme olasılığı,

$P(X)$; veri toplanmadan, gözlemlenmeden X'in olasılığı (prior probability),

$P(Y)$; Normalleştirme sabitindeki (Similarity range) hipotezde olan verilerin olasılığıdır (Marginal likelihood).

1.11.4. K-En Yakın Komşu (KNN)

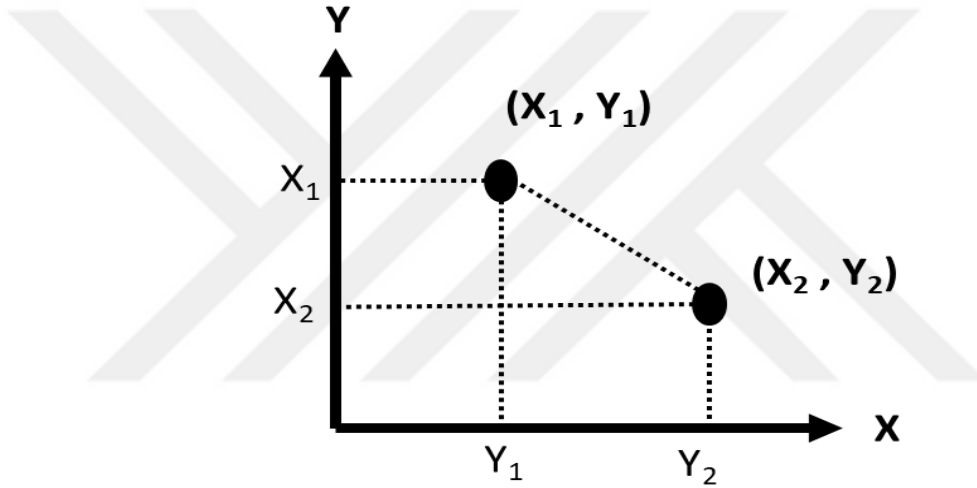
K - en yakın komşu algoritması, tüm makine öğrenme algoritmalarının en basitleri arasında olup, sınıflandırma veya tahminleme yapılacak nesne, komşularının çoğunluk oyuyla sınıflandırılmakta ve en yakın k komşusu arasındaki en yaygın sınıfa atanmaktadır (Joseph ve ark., 2016; Yang, 2019). K-en yakın komşu (KNN) algoritması etiketleme yani sınıflandırılma bilgisine ihtiyaç duyduğundan denetimli makine öğrenimi içerisinde yer almaktadır. Söz konusu bilgiler çerçevesinde eğitim seti üzerinden model eğitime çalışılmaktadır. Verilerdeki benzerlik ve örüntüleri kullanarak sınıflandırma ve hatta geçmişteki benzer veri noktalarına dayalı olarak tahminleme yapabilmektedir (Özdemir, 2016).

KNN algoritmasında kullanıcı bazlı k değeri tespit edilir, model öğrendiği eğitim setindeki yeni gözlem verisi hakkında fikir yürütmek için gözlem verisinin k tane en yakın komşusuna olan uzaklığına göre sınıflandırma yapmaktadır (Joseph ve ark., 2016). Bu sebeple de k değerinin belirlenmesi sınıflandırmanın da nasıl yapılacağını etkileyen hususlardan biridir. K değeri dışında, varyans, gürültü (noisy) ve bias (dizgesel hata) da çıktıyı etkileyen önemli unsurlardandır. Büyük bir k değeri seçmek gürültü verilerinin varyans değerlerini düşüreceğinden, bias değerini arttıracak ve küçük fakat önemli birçok verinin gözden kaçırılmasına sebep olmaktadır (James ve ark., 2013). Örneğin içerisinde 100 örneklem bulunan veri yapısında, bilinmeyen bir X değeri k=1 ise en yakın komşusu üzerinden, k=10 ise en yakın 10 komşusu

üzerinden, $k=100$ ise de tüm veriler üzerinden tahmin edilmeye çalışılacaktır. Bu sebeple k değeri genelde 3 ile 10 arasında ya da eğitim veri sayısının karekökü şeklinde alınabilmektedir fakat k değerinin yanısıra verilerin özellikleri arasında normalizasyon yapmak da incelemedeki en önemli etmendir (Lantz, 2013). Modelin eğitim seti ile çok uyumlu (overfitting) ya da az uyumlu (underfitting) olması durumunda bias ve varyans arasındaki denge kurmaya (bias-variance tradeoff) çalışılmaktadır.

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Şekil 1.32. Öklid mesafesi formülasyonu



Şekil 1.33. KNN’de öklid mesafesi

KNN algoritmasında k değeri kapsamındaki en yakın komşular üzerinden mesafe hesabı yapılarak öğrenme gerçekleşmektedir. Söz konusu mesafeler, benzerlik ve örüntü modelleri konusunda fikir vermektedir. Etiketli veriler ile gözlem verisi arasındaki mesafenin hesaplanmasında en fazla öklid kullanılmaktadır olup, Manhattan, Simple Matching, Hamming ve Jaccard gibi yöntemler de mevcuttur (Yang, 2019). Öklid iki nokta arası doğrusal uzaklık olup, formülasyonu (Şekil 1.32) ve iki boyutlu düzeyde gösterimi (Şekil 1.33) aşağıda belirtilmektedir.

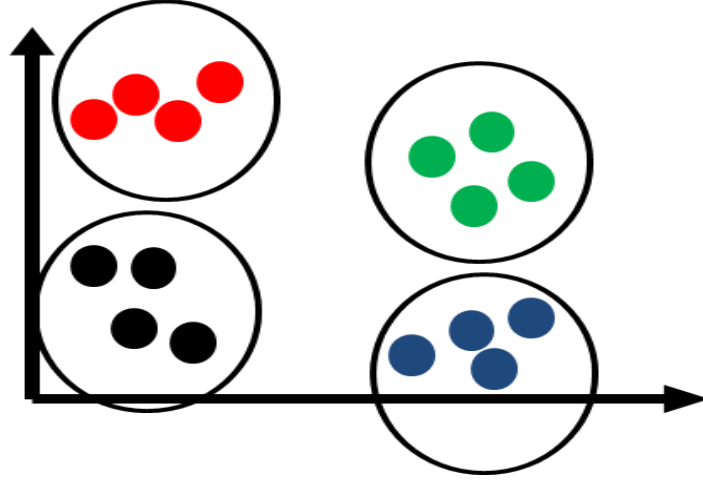
KNN algoritmasının basit ve verimli olmasından dolayı eğitim sürecinin hızla gerçekleşmesi ile genel veri dağılımı hakkında varsayım yapmaması avantajken,

yüksek depolama alanı gerektirmesi, sınıflandırma işleminin diğer algoritmalara kıyasla daha yavaş gerçekleşmesi ve kayıp değer (missing value) olarak tanımlanan veriseti üzerindeki kayıp verilere ilişkin işlem yükünü arttırması ise dezavantaj olarak yansımaktadır (Amra ve Maghari, 2017).

1.11.5. K- Ortalamalar Yöntemi

K-ortalamalar yöntemi, kendi içinde kategorize edilmemiş verileri sınıflandırmak için kullanılan denetimsiz öğrenme metodudur (Yang, 2019). Veri seti üzerinde yer alan verileri k tane gruba bölerek, kümeleme işlemi sağlamaktadır. Bu sebeple de verinin kaç gruptan oluştuğunun yaklaşık olarak bilinmesinde fayda bulunmaktadır.

K noktası kadar rastgele centroid noktalar veriye atanır, diğer veri setleri mesafelerine göre kümelenir, sonrasında yeni kümelemeye göre yeni bir centroid noktası seçilerek kümeleme işlemine devam edilir. Centroid kavramı, kümenin merkezini ifade etmektedir (Özdemir, 2016). Centroid noktaların yeri değişmeye ve bütün veri noktalarına bir kümeye atanana kadar bu işleme devam edilir. İşlem tekrarlandıkça centroid noktasının yerinin sabitleşmeye başladığı görülecektir. Buradaki ana amaç, fonksiyon içindeki yapıların, küme merkezinde yer alan centroid noktasına en az mesafede olmasıdır. Burada önemli olan diğer bir nokta k sayısı, veri miktarı kadar olduğunda uzaklık 0 çıkacak, fakat istenen gruplama yapılmayacaktır. Bu sebeple en iyi sonuç yerine, optimum sonucun alındığı k değerinin tercih edilmesi önem arz etmektedir. Centroid noktalara olan mesafede öklit mesafesi kullanılmaktadır (Joshi, 2020).



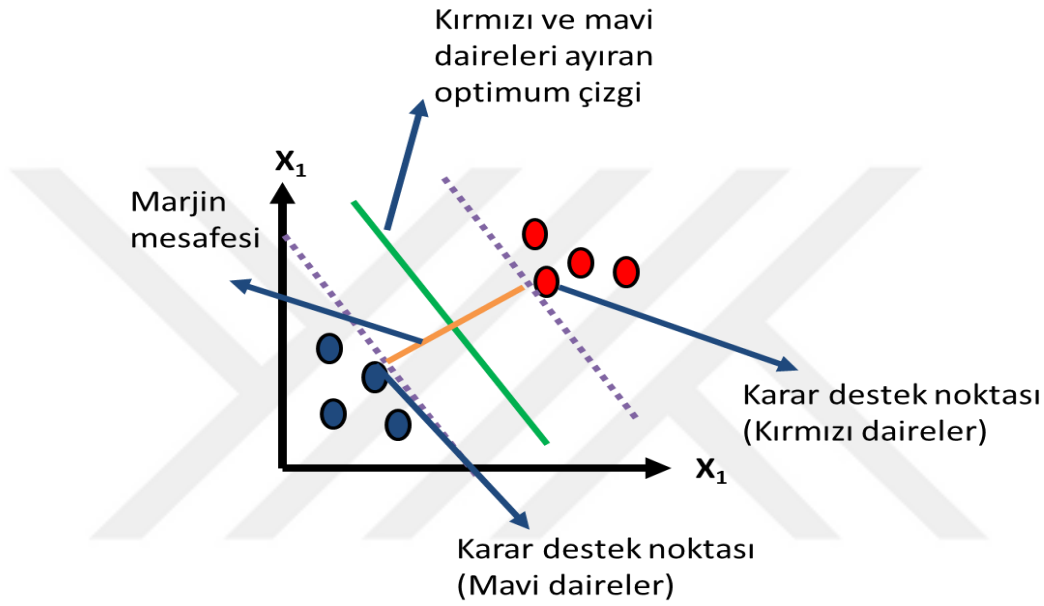
Şekil 1.34. K-Ortalamalar algoritmasına göre sınıflandırma

Şekil 1.34’de belirtildiği üzere kümeleme işlemi sonrası, aynı grup içinde yer alan verilerin birbirlerine benzerlikleri fazlayken, farklı gruplar arası ise benzerliklerin daha az olması beklenmektedir. Şekil 1.34 üzerinde örnekle ifade etmek gerekirse, grafiğin sol üst tarafında yer alan kırmızı veri grubu ile grafiğin sağ üst tarafında yer alan yeşil grubun birbirine benzer olmadığı, fakat renk gruplarının kendi içerisinde benzer olduğu kümeleme yöntemidir. K-ortalamalar yöntemi; uygulanması basit, büyük veriler için hızla sonuç alınabilen ve ölçeklenebilir bir algoritmadır (Borlea ve ark., 2017). K değerinin belirlenmesi için ise WCSS (Within Cluster Sum of Squares) formülasyonu kullanılmaktadır (Mohamad ve Usman, 2013).

1.11.6. Destek Vektör Makineleri (SVM)

Destek vektör makineleri özellikle sınıflandırma temelinde çalışarak, genelleme performansının oldukça yüksek olmasından dolayı kullanılan denetimli algoritmadır (Joshi, 2020). Öğrenme süresi uzun olmakla birlikte, öğrenmedeki işlem sayısını azaltması, ezber tarzı öğrenmeyi daha fazla engellemesi ve büyük hacimli veri setleri üzerinde de başarılı olması sebebiyle daha optimize sonuçlar elde edilebilmektedir. Kendisini oluşturan özelliklerin kendi aralarındaki ilişkiden etkilenmemesi sebebiyle, genelleştirme yöntemleri olarak da bilinen makine öğrenimi yöntemlerinden biridir.

Destek vektör makineleri, hem doğrusal hem de doğrusal olmayan veri kümelerini sınıflayabilmekte olup, özellikle veri gruplarını birbirinden ayırmada en uygun fonksiyonun ortaya konulması konusunda en başarılı sonuçları sağlayan algoritmadır (Dey, 2016). Destek vektör makineleri iki boyutlu yapıda doğrusal şekilde, üç boyutlu yapıda düzlemsel ve çok boyutlu yapılarda ise hiperdüzlem şeklinde veri setini sınıflandırabilmektedir.



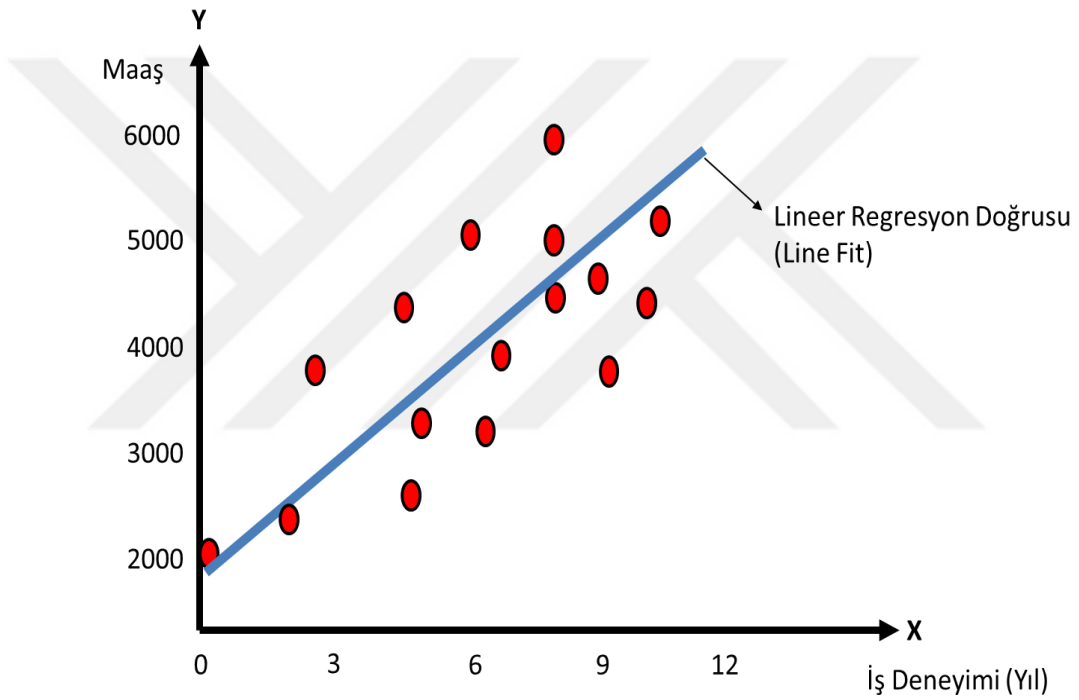
Şekil 1.35. Destek vektör makinelerine göre sınıflandırma

Sınıflandırmada şekil 1.35’de görüldüğü üzere veri seti içerisinde yer alan iki farklı veri grubu, karar destek noktaları üzerinden birbirinden ayrılmıştır. Karar destek doğrusuna en yakın noktalar, karar destek noktaları olarak isimlendirilmektedir. Amaç, karar destek noktaları arası mesafeyi maksimuma çıkarmaktır (Marjin mesafesi) (Lee, 2019).

1.11.7. Lineer Regresyon

Genel olarak bir bağımlı değişkenin, bir veya daha fazla bağımsız değişkenle ilişkisini istatistiksel olarak ortaya koyarak çıkarımlar yapma regresyon analizi olarak

tanımlanmaktadır (Joseph ve ark., 2016). Lineer regresyon basit açıdan bakıldığında iki nitel veri arasında ilişkiyi istatistiksel olarak ortaya koyan ve aralarındaki ilişkiden yola çıkarak tahminleme yapabilen makine öğrenimi modelidir, sayısal verilerle sınırlı olup, öğrendiklerini açıklayamaz (Joseph ve ark., 2016; Lee, 2019). Nitel veri sayısının ikiden fazla olması durumunda çoklu lineer regresyon (multiple linear regression) olarak isimlendirilmektedir. Gerçek hayatta karşılaştığımız çoğu olayda birden fazla değişkenin bulunması sebebiyle, çoklu lineer regresyon daha sık kullanılmaktadır. Bağımlı değişkenlerin genelde normal dağılım yapması beklenir.



Şekil 1.36. Lineer regresyon modellemesi

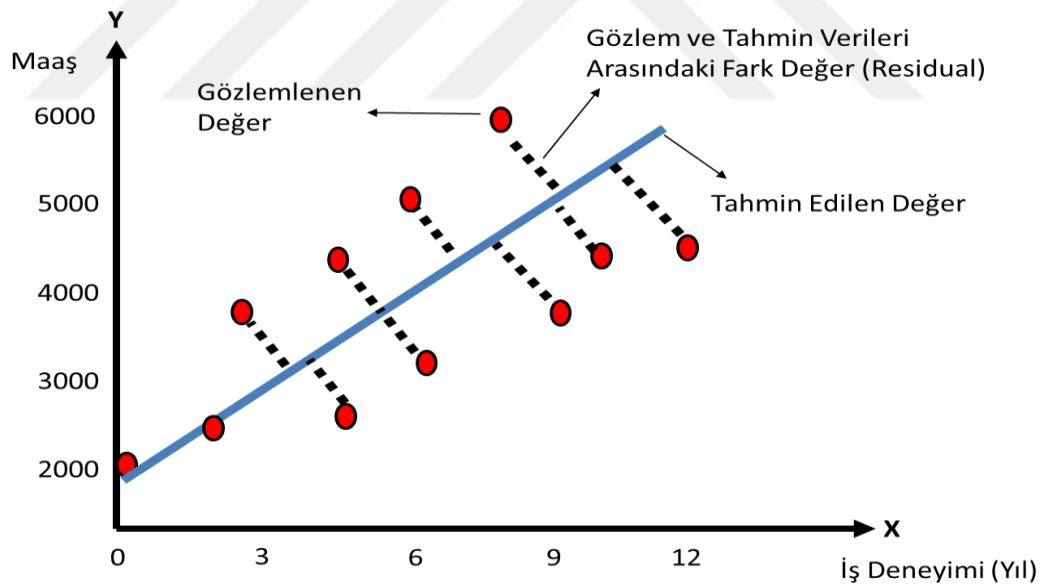
Şekil 1.36’da yer alan iş deneyimi ile alınan maaş arasındaki ilişki örneğinde görüldüğü üzere, lineer regresyon doğrusu üzerinden tahminleme yapmak mümkün olmaktadır.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

Şekil 1.37. Lineer regresyon formülasyonu

Bu kapsamda şekil 1.37’de belirtilen lineer regresyon doğrusunun matematiksel olarak nasıl ifade edildiğini inceleyecek olursak;

- y ; y ekseninde yer alan bağımlı çıktı değişkeni (Örnekteki maaş)
- x ; x ekseninde yer alan bağımsız girdi değişkeni (Örnekte iş deneyimi)
- X_n ; X ekseninde n ’inci girdi değişkeni
- β_0 ; sabit değer (Constant) (Bias) (Örnekte y eksenini kesme noktası)
- β_n ; X_n terimi için katsayıdır (Coeff) (Eğim)



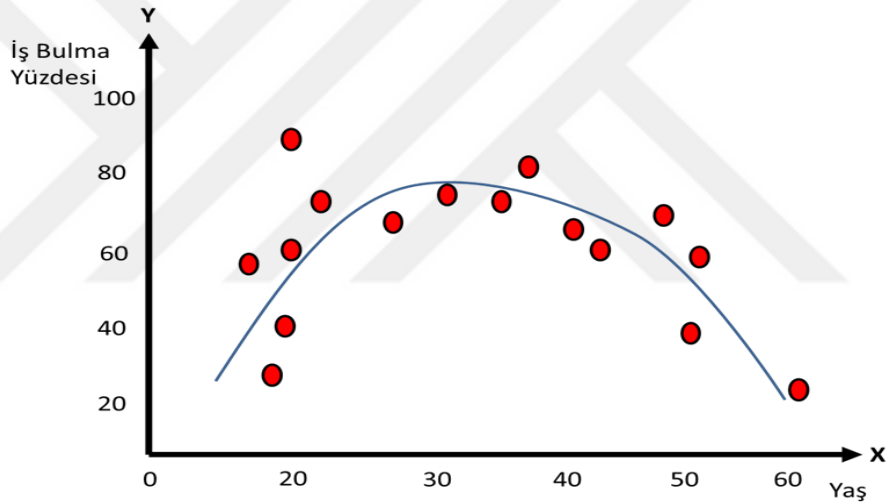
Şekil 1.38. Lineer regresyonda residual değerleri

Şekil 1.38’de görüldüğü üzere iş deneyimi bağımsız değişkenine bağlı olarak, maaş bağımlı değişkeni tahminlemesinde gözlem ve tahmin verileri arasında farklar oluştuğu, ortaya konulan modelin her zaman doğru tahminleme yapamadığı ortadadır

(Lee, 2019; Özdemir, 2016). Mümkün olan en iyi lineer doğru, tahmin ve gözlem verilerinin farkının karelerinin (Söz konusu verilerin fakı bazen pozitif bazen de negatif çıkacağı için, hatayı tam olarak görebilmek adına karesi alınmaktadır.) toplamının alabileceği minimum değerin bulunduğu doğru yapısıdır. Bu kapsamda aşağıda matematiksel formülasyonda belirtildiği üzere her y değeri için fark hesaplanarak, toplanmaktadır (n: gözlem sayısı, $Y_{\text{tahmin}}(i)$ = i verisi için tahmin edilen değer, Y_i = i verisi için gözlemlenen değer) (Şekil 1.39).

$$SS_{\text{residuals}} = \sum_{i=1}^n (Y_{\text{tahmin}}(i) - Y_i)^2$$

Şekil 1.39. Residuals formülasyonu



Şekil 1.40. Polinomsal lineer regresyon

Çoklu lineer regresyon dışında, polinomsal lineer regresyon (polynomial linear regression) ile parabolik eğriler elde etmek mümkündür. Örneğin şekil 1.40'da belirtilen yaş ile iş bulma imkânını inceleyecek olursak, sürekli olarak yaşın artmasıyla iş bulma oranının artmadığı görünmektedir. Grafiğin bu şekilde olmasının sebebi de genç yaşlarda tecrübe eksikliği sebebiyle, ilerleyen yaşlarda da yaşlılık sebebiyle işverenlerin iş vermek istemesinden kaynaklı grafiğin parabol olması beklenmektedir. Bu kapsamda parabol şeklinde formülasyonları elde etmek için yapılan polinomsal lineer regresyonun formülasyonu aşağıda belirtilmiştir (Şekil 1.41).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \dots + \beta_n x_n^n$$

Şekil 1.41. Polinomsal lineer regresyon formülasyonu

1.11.8. Lojistik Regresyon

Lojistik regresyon, tıpkı lineer regresyonda olduğu üzere bağımlı değişkenin, bağımsız değişkenler ile ilişkisini incelemek suretiyle makine öğrenimi yapan bir algoritma olup, klasik lineer regresyondan en büyük farkı verileri sınıflamasıdır (Hilbe, 2016). Regresyonun özel bir şekli olup, algoritmanın ismi regresyon olarak geçse de, tahminlemeden ziyade kategorize etme yani sınıflandırma amaçlı kullanılmaktadır (Yang, 2019). Bu sebeple, lojistik regresyon yöntemi ikili sınıflandırma (binary classification) olarak da bilinmektedir. Lojistik regresyon basit işlem yapısından dolayı düşük varyansa sahiptir ve bu nedenle aşırı uyuma (overfitting) daha az eğilimlidir (Joseph ve ark., 2016). Doğrusal regresyonda bağımlı değişkenin alacağı değer tahmin edilmeye çalışılırken, Lojistik regresyonda bağımlı değişkenin alacağı değerler kategorize olup, bunların olma olasılıkları hesaplanmaktadır. Bu sebeple de bağımlı değişkeni kategorize bir veriden oluşmaktadır. Logit modeli olarak bilinen matematiksel yapıyla, tahmin yapılan bağımlı değişken ile gözlem değerleri arasındaki benzerliği maksimum yapmaya çalışılmaktadır. Lineer regresyonda olduğu gibi, bağımsız değişkenlerin normal dağılımda olması beklenmez. Matematiksel formülü aşağıda belirtilmiştir (Şekil 1.42).

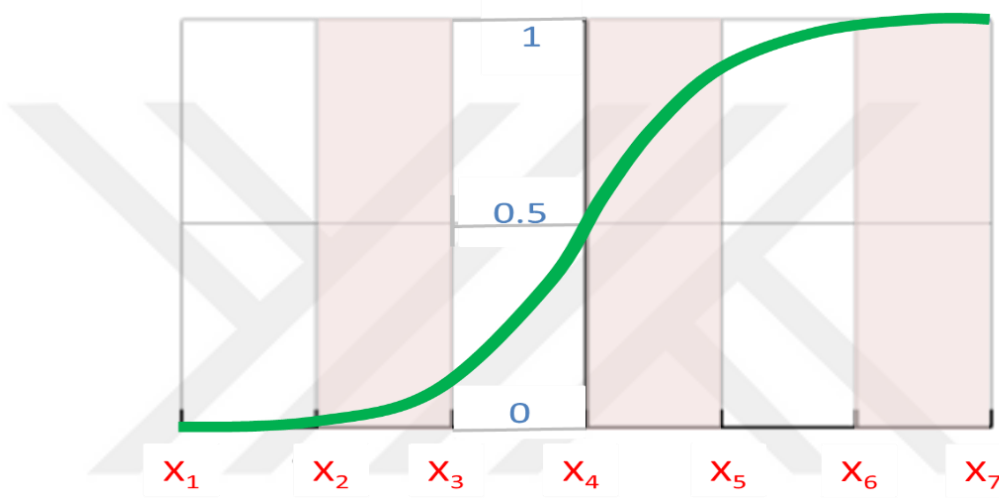
$$\pi = \text{Olasılık}(y = 1 | x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Şekil 1.42. Logit model formülasyonu

Lojistik regresyon formülasyonunun, lineer regresyondan farklı olmasının sebebi; lineer regresyonda bağımlı değişkenin sürekli olduğunu farzedilir, (Şekil 1.36’da belirtildiği üzere mantıksız olsa bile 1000 yıllık iş deneyimi olan birinin maaşı hesaplamak mümkündür.) fakat lojistik regresyonda y sadece meydana gelen bir

olayın olma olasılığını ifade eder. Bu sebeple de olasılık sadece 0 ile 1 arasında sürekli bir aralığı temsil etmektedir.

Bir sonucun ortaya çıkma olasılığı (Odds), sonucun gerçekleşme yollarının sayısının, olası tüm sonuçlara değil, diğer olası sonuçlara bölünmesiyle elde edilen orandır (Özdemir, 2016). Bu sebeple bir olayın gerçekleşme olasılığı O_i ise, gerçekleşmeme olasılığı da $1 - O_i$ 'dir. Odds oranı ise $O_i / (1 - O_i)$ olarak bulunmaktadır.



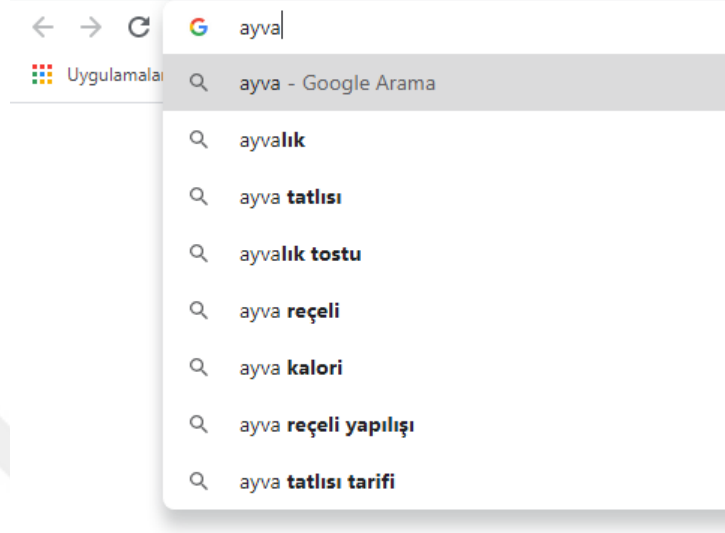
Şekil 1.43. Sigmoid eğrisi üzerinden olasılık

Şekil 1.43'de sigmoid eğrisi üzerinde görüldüğü gibi olasılık değerleri 0 ile 1 arasında olduğundan, odds oranıyla söz konusu değer 0 ile $+\infty$ arasında ifade edilecek ve logitin alınması ile birlikte de değişken $-\infty$ ile $+\infty$ arasında tanımlanmış olacaktır (Joshi, 2020). Bu sebeple de tahmin edilen olasılık, odds değerinin logaritmasının alınması ile oluşturulmaktadır.

1.11.9. Apriori Algoritması

Apriori algoritması, kendi veri kümesi içerisinde sık kullanılan öğeleri tündengelim yöntemi ile bularak, bunları kendi arasında ilişkilendirmek için kullanılmaktadır (Yuan, 2017). İlişkilendirme sonrası elde edilecek kombinasyonlar

temelinde kural çıkarımı da yapabilmektedir (Wang ve ark., 2019). Aynı zamanda bu ilişki verinin keşfi içinde oldukça önemlidir.



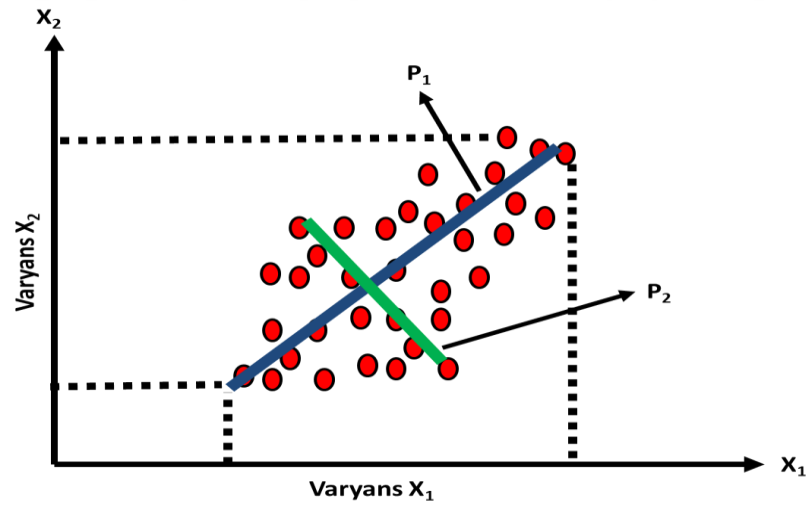
Şekil 1.44. Google arama motorunda apriori örneği

Yaklaşımı örnekle inceleyecek olursak, market alışverişinde makarna ve mayonez alan insanlar, ketçap da alıyorsa, bu ilişki “{makarna, mayonez} > ketçap” şeklinde özetlenebilmekte ve bu da söz konusu ilişkinin belli bir güven eşiğinin üzerinde olduğunu ortaya koymaktadır. Şekil 1.44’de görüldüğü üzere farklı bir örneklemede ise google arama motoruna “ayva” kelimesi yazıldığında, google bu kelimeden sonra yazılan bütün ilişkili kelimeleri veritabanından arayacaktır. Ortaya çıkan sonuçta, ilgili kelimenin aramalarda en sık kullanıldığı hali olarak; “ayva”, “ayvalık”, “ayva tatlısı”, “ayvalık tost”, “ayva reçeli”, “ayva kalori”, “ayva reçeli yapılışı” ve “ayva tatlısı tarifi” sonuçları ilişkilendirme sırasına göre sunulmaktadır.

1.11.10. Temel Bileşen Analizi (PCA)

Temel bileşen analizi (PCA) esasen bir dizi veri noktasını (örneğin ilişkilendirilebilen m değişkenleri için), ilişkisiz değişkenlerin bir dizi dönüştürülmüş noktası olan $k < p$ temel bileşen kümesine dönüştürmeyi amaçlamakta olup, bileşenlerin

en yüksek varyasyonla tanımlanması önemlidir (Jindal ve Kumar, 2017; Yang, 2019). Temel bileşen analizi (PCA), veri içerisindeki değişken sayısının azaltılmasını sağlayan yöntemdir (Özdemir, 2016). Değişken sayısının azaltılması hesaplamaların daha kolay ve hızlı olmasını sağlayacak ve veri boyutunu da kısmen küçültecektir (Dey, 2016). Özellikle veri görselleştirme yapılabilmesi için boyut indirgemesi oldukça önemlidir (Joshi, 2020). Çünkü örneğin 20 değişkeni (yaş, cinsiyet, aile yapısı, ailenin ekonomik durumu, kaçınıcı çocuk olduğu, baba mesleği, anne mesleği, çocuğun kendi odası olup olmadığı, cep telefonu olup olmadığı, tableti olup olmadığı, bilgisayarı olup olmadığı vb.) olan öğrenci başarısının incelendiği bir örneklemi, 20 boyutta görselleştirmek imkânsızdır. İdeal olarak, k, p'den çok daha az olmalıdır, bu nedenle PCA, boyutsallığın azaltılması için bir teknik olarak düşünülebilmekte olup, bu durumda da hiperparametre olan k'nın seçimi oldukça zor olabilmektedir (Yang, 2019). Hiperparametre olan k değeri çok küçükse, çok fazla bilgi kaybolduğu, p'ye yakınsa da bilgilerin korunduğuna işaret eder. Dolayısıyla k seçimi büyük ölçüde ihtiyaç duyulan temsillerin ve bilginin kalitesine bağlı olarak, sezgisel bir yaklaşımla, k bileşenlerinde tutulabilecek varyans yüzdesine dayanır (Yang, 2019).



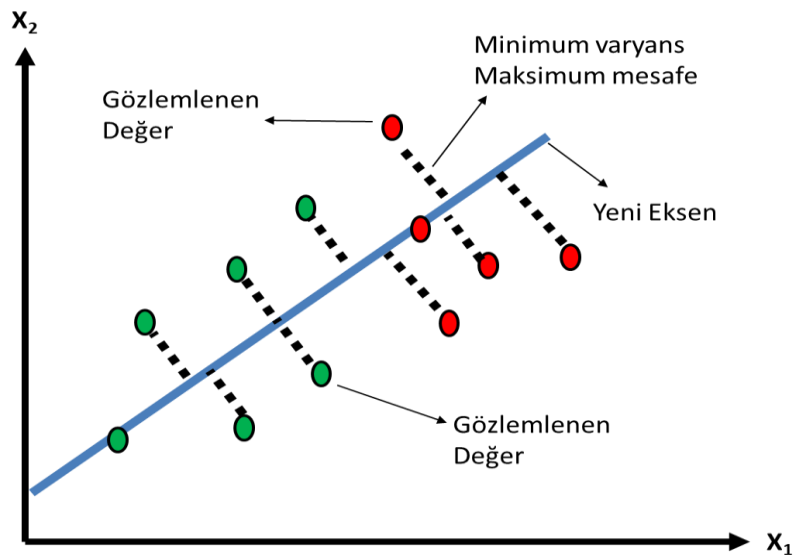
Şekil 1.45. Örnek PCA analizi

Çok yüksek boyutlarda görselleştirme olmasa da bunun yerine şekil 1.45’de gösterildiği üzere varyans değerleri maksimum tutularak, yeni bir ana bileşen elde

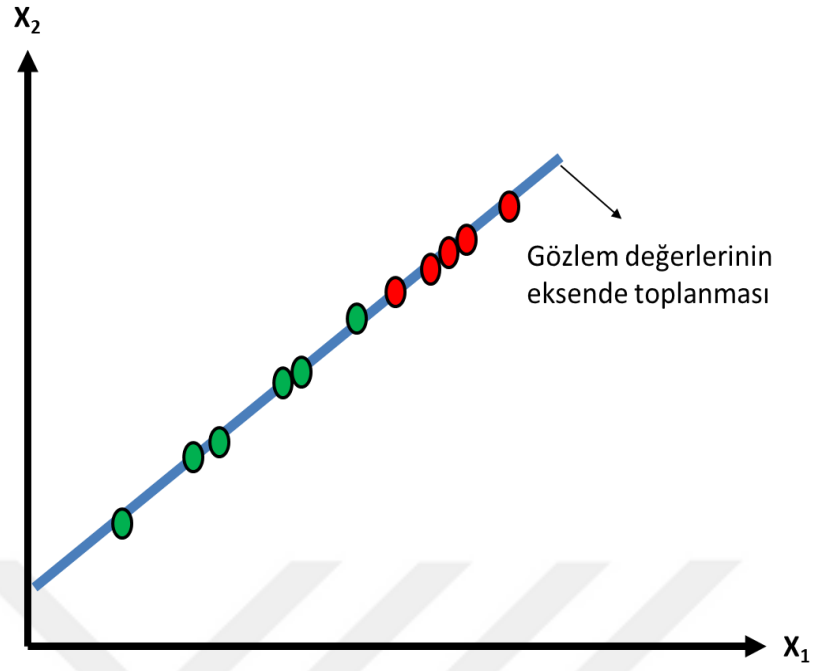
edilmeye çalışılır. Buradaki en önemli nokta PCA işlemi yapılırken verinin önemli özelliklerini koruması fakat boyutunun da düşük olması istenmektedir. Elde edilmeye çalışılan yeni ana bileşen değeri, boyut değeri azaltılacak bileşenlerin eksen değerlerinin maksimum varyansı elde ettiği yer olması sebebiyle, her PCA birbirine diktir. PCA'ların kendi arasındaki diklik ilişkisi korelasyonun sıfır olduğunun da göstergesidir. Bunun dışında tekil değer ayrıştırması analizi de (Singular value decomposition) (SVD), PCA ile oldukça yakındır.

1.11.11. Lineer Diskriminant Analiz (LDA)

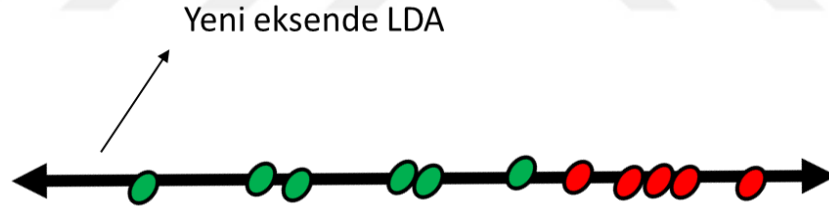
Lineer diskriminant analizin ana fikri, farklı sınıfların ayrılabilirliğini en üst düzeye çıkaran bir dönüşüm sistemini ortaya koymaktır (Yang, 2019). Temelde bir boyut azaltma tekniğidir, veri kümesi içindeki değişkenler azaltılırken, bilgi mümkün olduğu kadar çok bırakılmaya çalışılmaktadır (Joshi, 2020). Yapılan işlemle sınıf içi varyanslar en aza indirilirken, farklı sınıflardan araçların farkı da en üst düzeye çıkarılmakta olup, bu da temel varsayım olarak verinin tek modlu Gauss dağılımına uyması anlamına gelmektedir. (Yang, 2019).



Şekil 1.46. Gözlemlenen değerlerin iki ekseninde (X_1 / X_2) gösterimi

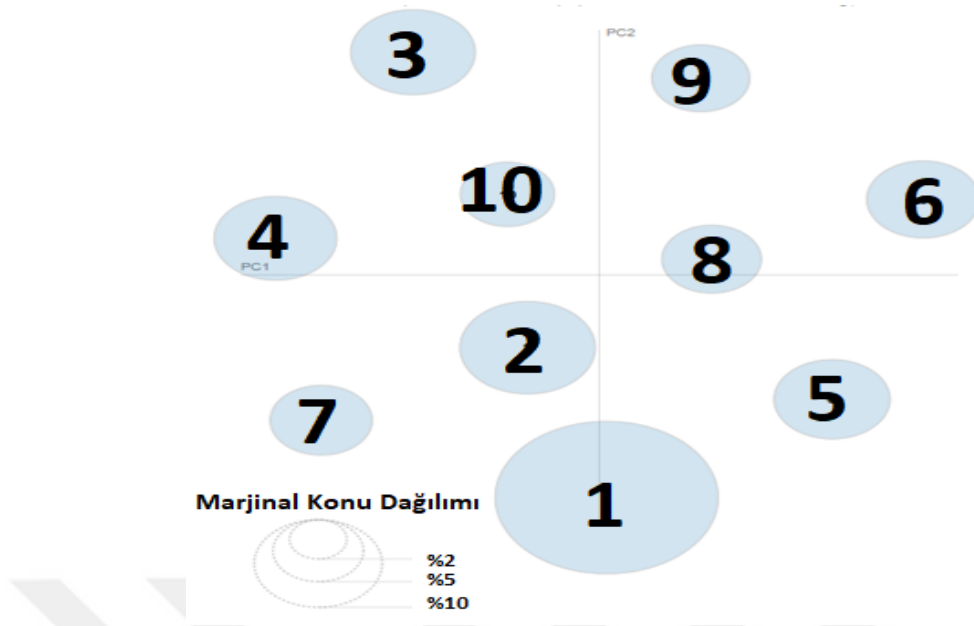


Şekil 1.47. Gözlemlenen değerlerin yeni eksende toplanması



Şekil 1.48. Gözlemlenen değerlerin yeni eksende gösterimi

LDA analizinin işleyişine ilişkin şekil 1.46 incelendiğinde, iki eksen (X_1 / X_2) üzerinde yer alan gözlemlenen değerlerin, sınıf içi varyansları minimuma ve mesafeleri de maksimuma alınarak, yeni eksen üzerindeki iz düşümleri bulunmaktadır. Şekil 1.47’de yeni eksen üzerine gözlemlenen değerlerin toplandığı ve şekil 1.48’de ise yeni eksenin tek vektörel düzeyde yeniden ortaya konularak, iki boyuttan tek boyuta indirgendeki görülmektedir.



Şekil 1.49. Vektörel düzlemde örnek LDA analizi (k:10)

Şekil 1.49 metin verileri içeren korpustan örneklem oluşturulması maksadıyla elde edilmiş olup, birbirine yakın veri kümeleri içinde varyansların az olduğu, farklı veri kümelerinde de varyansın arttığını söylemek mümkündür.

LDA ve PCA örüntü tanıma ve yüz tanıma gibi birçok uygulama için çok etkili olmakla birlikte, LDA örüntü tanıma için daha iyi sonuçlar verirken, PCA küçük eğitim veri setleri için daha iyi performans gösterebilmektedir (Yang, 2019).

1.11.12. Bellek Temelli Nedenleme

İnsanlar adına tecrübe dediğimiz deneyimlerle karar vermekte olup, bu yöntemde girdiler eski bilinen kayıtların tutulduğu alana girdiğinde, mevcut karma kayıtlara göre sınıflandırma / tahmin işlemi gerçekleştiği yöntem, bellek temelli nedenleme (MBR) olarak bilinmektedir. Bellek organizasyonu ve erişimi, vaka temelli yöntemlerin odak noktası olup, paralel işleme teknikleri kullanılması diğer algoritmalarından ayrılan yönüdür (Aamodt ve Plaza, 1994). MBR modelleme, girdi

değişkenlerinin sayısal, birbirine dik ve standartlaştırıldığı varsayımlarına dayanmaktadır.

1.11.13. Torbalama

Torbalama (Bagging - Bootstrap Aggregation), topluluk öğrenimi ile çoğunlukla sınıflandırıcı oluşturmak maksadıyla kullanılan basit ve etkili bir yöntemdir (Joshi, 2020). Kendi içinde örneklem usulü alt kümelerine ayrılan verinin (Bootstrap örneklem) genelde karar ağaçları ağırlıkta yüksek varyansa sahip veriye uygulandığı modellemedir. Modelleme sonrası elde edilen tahmini değerlerin birleştirilmesi sapmayı minimize edecek ve bu da sistemin daha stabil sonuçlar üretmesini sağlayacaktır. Bu yönüyle rastgele orman yöntemine oldukça benzemektedir. Fakat rastgele ormanda altkümelerin, düğümlerin dallanması için kullanılması ve bunun da veri kümelerinin birbirine benzeme ihtimalini düşürmesi sebebiyle, bazı çalışma alanlarında torbalamaya göre daha işlevli ve etkili olduğu söylenebilir. Bunun dışında topluluk öğrenimi olması sebebiyle, her şekilde tek bir eğitim seti ile eğitilen modellere göre çok daha iyi sonuçlar elde edileceği bilinmektedir (James ve ark., 2013).

1.11.14. Artırma

Artırma (Boosting) modellemesi de topluluk algoritmaları arasında yer almakta olup, varyans ve bias (dizgisel hata) değerinin düşürülmesi ve bunla orantılı olarak tahminin güçlenmesinde kullanılan bir yöntemdir (Dey, 2016). Veri kümesi rastgele alt kümelere bölünür. Torbalamada olan bootstrapler kullanılmaz, oraya aşamalı olarak gelinmektedir. Her bir küme farklı algoritmalarla test edilir, yanlış tahmin edilen veri kümesi, farklı algoritma temelinde öğrenim yapılarak, tekrar test edilir. Bu modellemede ana sistem yapısı, her seferinde ortaya konulan doğru tahminlerden oluşmaktadır. Bu yönüyle de torbalamadan ayrılmaktadır. Çünkü torbalamada kopyalanan veri setlerinin her biri için farklı bir karar ağacı modeli oluşturulmakta ve

tahmin sistemi de bundan etkilenmemektedir. Artırmada ise ağaç yapıları kendinden önceki ağaç yapılarındaki bilgileri kullanarak, etkileşimli (sequential) şekilde, hatalardan dersler çıkararak oluşmaktadır (James ve ark., 2013). En popüler kullanılan uyumlu artırma (Adaptive boost - AdaBoost) modelidir (Dey, 2016).

1.11.15. Gradyan Artırma Makineleri

Gradyan artırma makineleri, ingilizce baş harflerinden oluşan “GBM” olarak da bilinmekte olup, regresyon ve sınıflandırma problemleri için kullanılan öğrenme çeşididir (Sarkar, 2019). Uyumlu artırma modellemesinde olduğu üzere, zayıf öğrenimlerin hatalardan ders çıkararak güçlü öğrenimlere dönüşmesi beklenmekle birlikte, modelin eksikliklerini yüksek nitelik değerindeki veri noktalarını kullanarak elde edilmesi beklenirken, GBM’de ise aynı amaç kayıp fonksiyonlar kullanılarak yapılmaktadır.

1.12. Derin Öğrenme

Derin öğrenme günümüzde en popüler konular arasında yer alan makine öğrenimi olup, hiyerarşik mimarileri kullanarak verilerdeki üst düzey soyutlamaları öğrenmeye çalışan bir makine öğreniminin alt alanıdır (Guo ve ark., 2015). Derin öğrenme, karmaşık yapıları çok katmanlı sinir ağlarıyla öğrenebilen ve klasik makine öğrenim algoritmalarının kolaylıkla öğrenemedikleri modellemeleri gizli katmanlar aracılığıyla çok daha kolay ve başarılı şekilde öğrenen yapılardır (Dey, 2016). Gizli katmanlar, istenen çıktıyı elde etmek için matematiksel fonksiyonlarla hesaplamaların yer aldığı bir yapı olup, katman sayısı değişkenlik gösterebilmektedir. Derin öğrenme yeni bir yaklaşım olmasına karşın, semantik, transfer öğrenim, doğal dil işleme, görselleştirme ve suç inceleme gibi pek çok konuda aktif şekilde kullanılmaya başlanmıştır (Hassan ve ark., 2019, Gheisari ve ark., 2017; Guo ve ark., 2015; Sharma ve Singh, 2017; Young ve ark., 2018,).



Şekil 1.50. Yapay zekâ, makine ve derin öğrenme ilişkisi

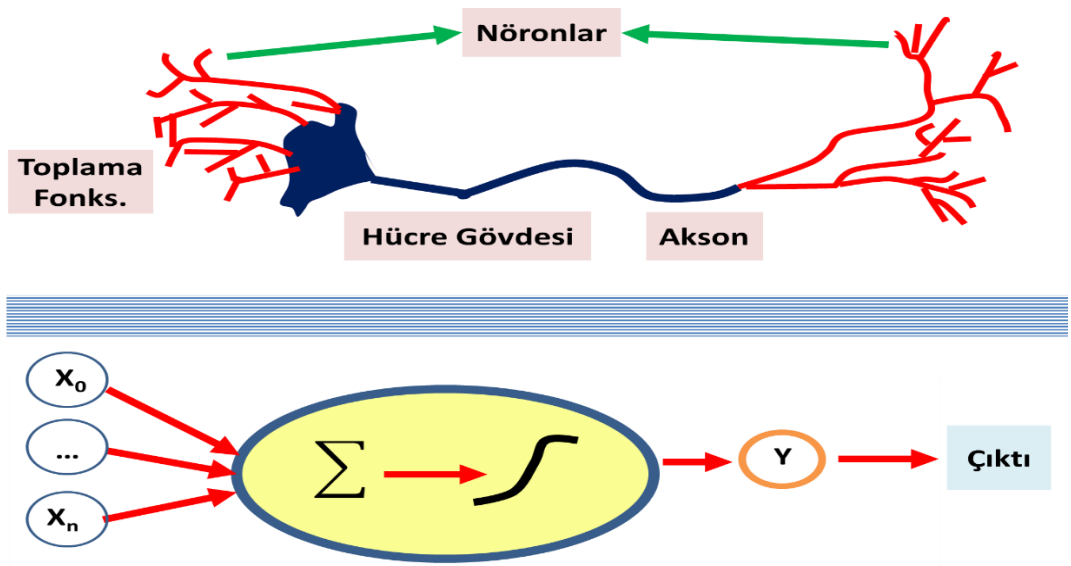
Derin öğrenme, şekil 1.50’de görüldüğü üzere yapay zekâ altında yer alan makine öğreniminin alt kümesidir. Makine öğrenimi olmasına rağmen, makine öğrenimi dışında ayrı bir başlık açmamın temel sebebi, bilhassa büyük veriler üzerinde derin öğrenmenin çok daha başarılı sonuçlar vermesidir. Bu öğrenme yöntemi ile birlikte veri kümesi içinde yer alan girdilerle, denetimli ya da denetimsiz çıktılar tahmin edilmeye veya sınıflandırılmaya çalışılmaktadır (Goodfellow ve ark., 2016)

Lojistik regresyon yaklaşımından en temel farkı, girdi-çıkı arasında gizli katmanların yer almasıdır. Gizli katman sayısının fazla olması, subjektif olarak öğrenmenin derinliğini de ortaya koymaktadır. Subjektif olarak anılmasındaki temel gerekçe günümüzde iki katmanlıdan, binlerce katmanlıya varana kadar öğrenme gerçekleşmekte olup, hepsi de derin öğrenme olarak anılmasına karşın, kendi içinde göreceli farklılıkları barındırmasıdır. Katman sayısı arttıkça öğrenme işlemi için gerekli donanım (GPU (Grafik işleme ünitesi – Graphics Processing Unit) gereksinimi başta olmak üzere) ve süre (parametre sayısının fazla olmasından dolayı) de artacaktır. Ayrıca derin öğrenmenin eğitim sırasında, klasik algoritmalara göre çok daha fazla veri gereksinimi duyduğu bilinmekle birlikte, veri bolluğu yaşadığımız günümüzde bu durum çok da problem olmamaktadır. Fakat incelenecek veri miktarı az ise klasik algoritmalar birçok yönden daha başarılı sonuçlar ortaya koyabilecektir.

Klasik makine öğrenme algoritmaları örüntüyü tanımak maksadıyla istatistiksel analiz temelinde yöntemler kullanırken, derin öğrenmede insan beyninin nöronları benzeri bir modelleme yapılmaktadır. Bu sebeple, kayıpları (loss function) en aza indiren hiperparametre optimizasyonu (hyperparameter tuning) çok daha farklı şekillerde yapılabilmektedir. Derin öğrenme ile insan müdahalesi olmaksızın, özellik çıkarımı yapılabilmesinden dolayı özellikle metin, ses ve görüntü verileri için iyi bir sınıflandırma performansı ortaya koymaktadır.

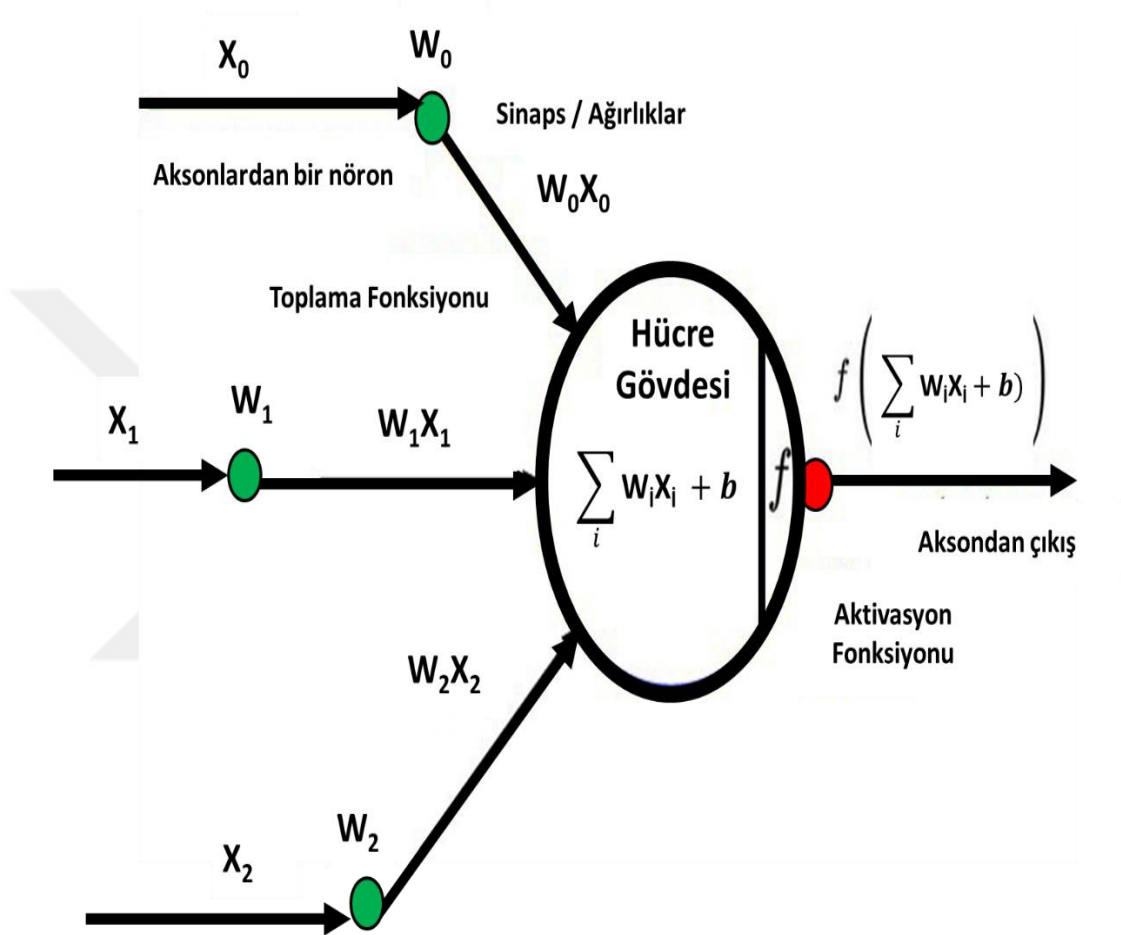
1.12.1. Yapay Sinir Ağları (ANN)

Yapay sinir ağları, insan sinir sisteminde yer alan nöronlar marifetiyle öğrenmenin gerçekleşmesine benzer şekilde, makineler tarafından öğrenmenin gerçekleşmesine dayanan bir yaklaşımdır (Joshi, 2020). Yapay sinir ağları, gerçek dünyada olduğu gibi fonksiyonları tanıma, sınıflandırma ve gerçekleştirme görevini üstlenmiştir (Atalay ve Çelik, 2017). Veri madenciliği genel olarak incelendiğinde, karar ağaçlarının anlaşılmasının, yorumlanmasının ve uygulanmasının ne kadar kolay olduğu konusunda, sinir ağlarını bunun tersi kadar zor olarak tanımlayabiliriz (Koyuncuğil ve Özgülbaş, 2009).



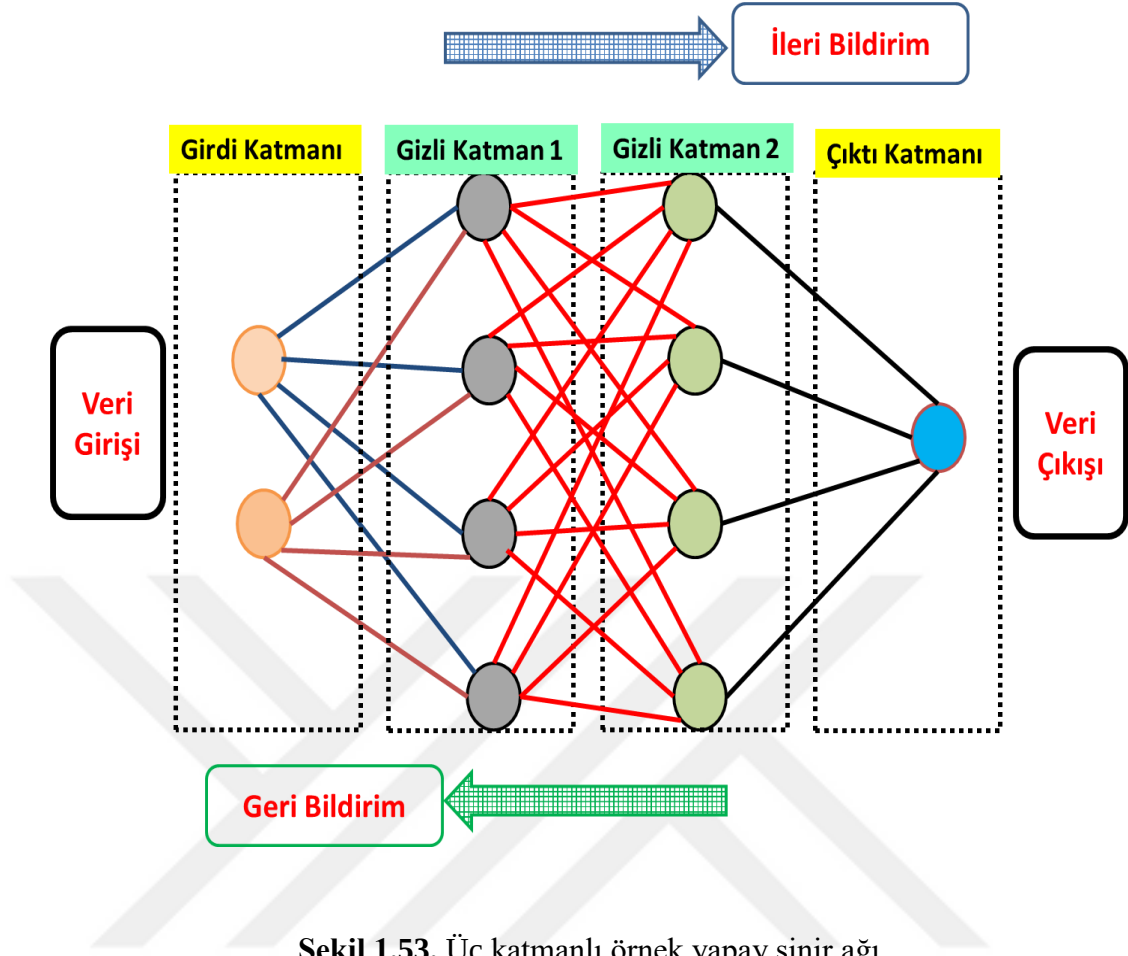
Şekil 1.51. Nöronlar yardımıyla öğrenme modellemesi

Sinir sisteminde nöronlar marifetiyle, ağırlıklarına (sinaps) göre toplama fonksiyonları (dentrit) tarafından gönderilen veriler, hücre gövdesinde aktivasyona maruz kalmakta ve aksonlarla veri çıkışı olmaktadır, nöronların bu etkileşimi şekil 1.51’de görüldüğü üzere insan sinir sistemine benzetilmeye çalışılmıştır.



Şekil 1.52. Nöronlar yardımıyla öğrenmenin matematiksel gösterimi

Yapılan işlemin matematiksel gösterimi, şekil 1.52’de belirtilmiştir. Bu eylemleri bildirim olarak da anılır (feed-forward), fakat zaman zaman yanılısamlarla geri bildirim (yayılm) (back-propagation) de yapılmakta olup, bu durum modelin öğrenme performansını arttırmaktadır. Aynı zamanda en sık kullanılanı geri bildirim (BPNN) (Back Propagation Neural Network) olan modellemedir (Singh ve ark., 2016).



Şekil 1.53. Üç katmanlı örnek yapay sinir ağı

Çalışma süreci şekil 1.53’de örnekleminde gösterildiği üzere, insan öğrenme biçiminin taklidi konumunda olup, yapay nöronlar yardımıyla giriş katmanında yer alan veri, gizli katman / katmanlar aracılığıyla çıkış katmanına iletilmektedir. Yapay sinir ağları, katman sayısına göre tek katmanlı veya çok katmanlı sinir ağlarından oluşabilmektedir. Katman sayısı hesaplanırken, giriş katmanı hesaba katılmaz.

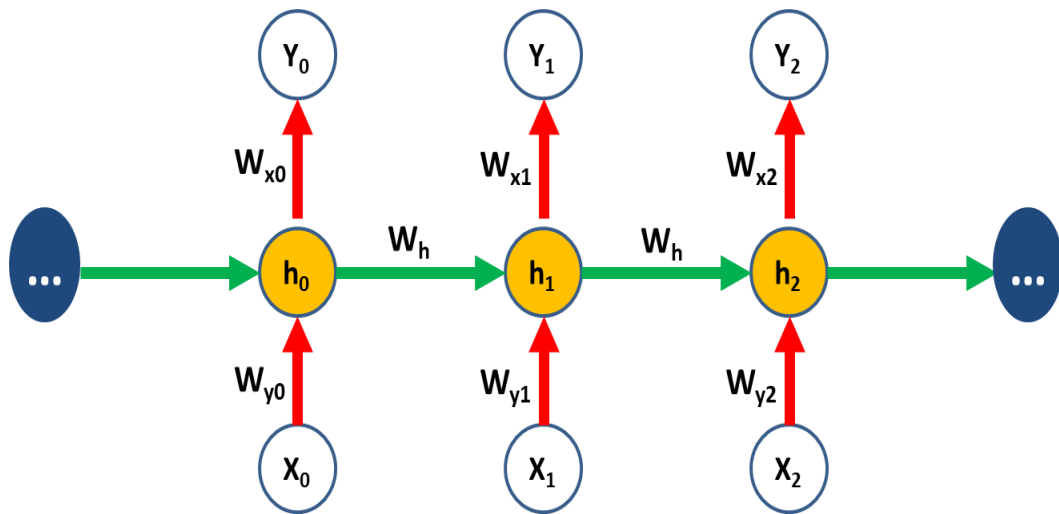
1.12.2. Evrimsel Sinir Ağları (CNN/ConvNet)

Evrimsel sinir ağları, resim incelemeleri özelinde sınıflandırma temelli olarak sinir ağlarını kullanan derin öğrenme yapısıdır. Bu kapsamda derin öğrenme, çoklu gizli katmanlara sahip sinir ağlarını içerir ve genellikle bir dizi işlemde çoklu evrimsel sinir ağlarını (CNN) tarama yapısına sahiptir (Albawi ve ark., 2017). Bir kıvrımlı sinir ağının (CNN) ana yapısı, görüntü gibi girdileri sabit boyutlu bir filtre

veya çekirdek içinden geçirerek bir dizi kıvrımlı özelliklere dönüştürmek için bir kıvrım katmanına sahip olmasıdır (Yang, 2019). Facebook özelinde örnek verecek olursak, yüklediğimiz her fotoğrafta birilerini etiketlemek istediğimiz zaman, sistem kendi sinir ağları içerisinde benzerliğe göre kimlerin olabileceğinin tahminlemesini yapabilmektedir. Yapay sinir ağlarında olduğu gibi evrimsel katmanlar marifetiyle, resim filtreleme ile algılanmaya çalışılmaktadır. Çoklu gizli katman seviyesinde işlem görmekte olup, algılanan nesne matrislere (feature map) dönüştürülerek, özellik çıkarımı yapılmaya çalışılmaktadır (Yang, 2019).

1.12.3. Tekrarlayan Sinir Ağları (RNN)

Klasik sinir ağı modellemesinde girdi ve çıktıların birbirinden bağımsız olduğu varsayımına tabi olarak işlem yapmaktayız. Fakat gerçek hayattaki durum tam olarak bunu karşılamamaktadır. Tekrarlayan sinir ağları ile geçmiş (kısa dönem) hatırladığımız ve geçmişte yaşadıklarımızdan öğrendiklerimizle karar vermekte olduğumuz varsayımı ön plana çıkıştır. Bu sebeple otomatik ilişkilendirme (auto associative) veya geri bildirim ağı (feedback network) olarak da bilinmekte olup, özellikle metin sınıflandırma temelinde aktif olarak kullanılmaktadır (Gomez-Alanis ve ark., 2019).

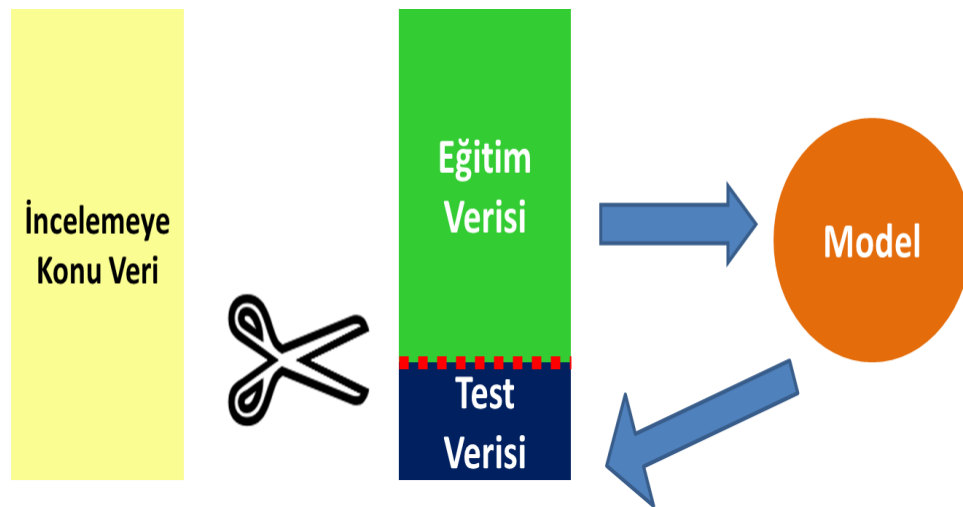


Şekil 1.54. Örnek RNN modeli

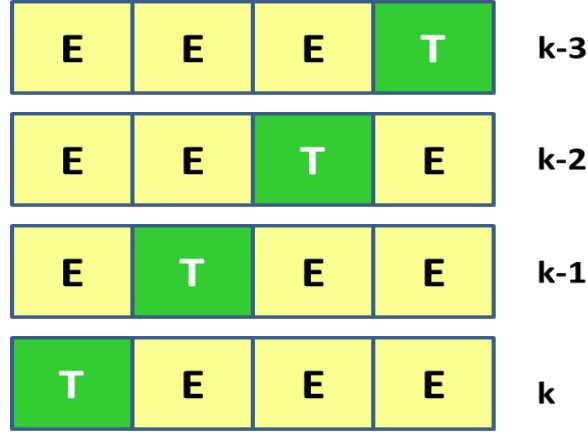
RNN ağı mimarisi, her bir nöron çifti arasında yönlendirilmiş bir bağlantı kullanan ve bağlantıların aynı zamanda mevcut katman içinde de geçerli olduğu ve bağlantı kuvvetleri olarak da zamanla değişen gerçek değerli ağırlıkları temel alarak şekil 1.54’de örneği olduğu üzere modellenmiştir (Yang, 2019). Tekrarlayan sinir ağları, zamanla değişen dinamik veya sıralı verilerle başa çıkmak için bir kapsam önermekte olup, hataların eğitimi ve en aza indirilmesi zaman içinde geri yayılım kullanmaktadır (Joshi, 2020).

1.13. Model Geçerleme ve Performans Metrikleri

Makine öğrenimi modelleri temelinde elde edilen sonuçların performans ölçümlerinin yapılması, değerlendirmenin ortaya koyulabilmesi adına önemlidir. Performans kavramı ise öznel ya da nesnel olabilmekte olup, öznel ölçme yöntemlerinin performans sonuçlarını objektif olarak değerlendirmek oldukça zor ve kesin güvenilirlik taşımamaktadır. Bu sebeple nesnel ölçme yöntemlerinin, yani varlığı kendi gerçekliğine dayanan yöntemlerin kullanılması gereklilik olup, söz konusu yöntemler de çeşitli matematiksel ifadeler içeren metriklerden oluşmaktadır. Genel olarak eğitimde eğitim ve test verileri ile çapraz geçerleme kullanılmaktadır (Sarkar, 2019).



Şekil 1.55. Eğitim ve test verisi kullanımı



Şekil 1.56. K-kat çapraz doğrulama (Örnek k:4)

Eğitim ve test verisi kullanımında makine öğrenimi modellerini eğitmek için kullanılan verisetinin bir kısmı dışarıda bırakılarak, eğitim sürecine katılmaz, test verisi olarak saklanır (Genelde %70 eğitim - %30 test veya %75 eğitim - %25 test) (Joshi, 2020). Eğitim bitimine müteakip, öğrenilenler test verileri üzerinde test edilir (Şekil 1.55). K-kat (K-fold) çapraz geçerlemede ise veriseti n kadar eşit alt kümelere bölünerek, n kadar eğitilip test edilir (leave one out) (Şekil 1.56). Çapraz doğrulama için sadece bir gözlem dışında kalan n örneğimiz ve modelleme üzerinden n kez tekrarlandı. Bu nedenle, seçim yanlılığı minimal fakat doğrulama performansı ise daha iyidir. Genelde verisetinin sayısının az olduğu durumlarda kullanılmaktadır (Rokach ve Maimon, 2015). Bu kapsamda alt maddelerde çeşitli performans metrikleri açıklanmaktadır.

1.13.1. Ortalama Mutlak Hata

Veriseti içerisinde gözlemlenen ve tahmin edilen değerlerdeki hataların negatif ve pozitiflikten arındırılarak, ortalamasının alınmasıdır (Şekil 1.57).

$$\text{Ortalama Mutlak Hata (MAE)} = \frac{1}{n} \sum_{i=1}^n |Y_{\text{tahmin}} - Y_{\text{gözlemlenen}}|$$

Şekil 1.57. Ortalama mutlak hata formülasyonu

1.13.2. Ortalama Kare Hatası

Veriseti içerisinde gözlemlenen ve tahmin edilen değerlerdeki hataların negatif ve pozitiflikten arındırılarak, karelerinin ortalamasının alınmasıdır (Şekil 1.58). MSE’de MAE’ye nazaran aykırı değerler daha büyük hataya sebep olmaktadır.

$$\text{Ortalama Kare Hatası (MSE)} = \frac{1}{n} \sum_{i=1}^n (Y_{\text{tahmin}} - Y_{\text{gözlemlenen}})^2$$

Şekil 1.58. Ortalama kare hatası formülasyonu

1.13.3. Kök Ortalama Kare Hatası

Veriseti içerisinde gözlemlenen ve tahmin edilen değerlerdeki hataların karelerinin (negatif ve pozitiflikten arındırma) ortalaması alınarak, elde edilen değer de karekökünün alınmasıdır (Şekil 1.59). RMSE’de hatanın hassasiyetinin genelde birkaç aykırı değere düşmesi sebebiyle duyarlılığın yüksek olduğu söylenebilmekte olup, bu kapsamda da regresyon problemlerinde en popüler değerlendirme metriği olarak anılmaktadır (Joshi, 2020).

$$\text{Kök Ortalama Kare Hatası (RMSE)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_{\text{tahmin}} - Y_{\text{gözlemlenen}})^2}$$

Şekil 1.59. Kök ortalama kare hatası formülasyonu

1.13.4. Normalleştirilmiş Hata

Özellikle farklı nitelikleri analiz ederken veya polarite analizlerinde normalleştirme oldukça önemlidir. Çünkü hata metrikleri genelde (-) ∞ ile (+) ∞ arasında bir sayı üretmekte olup, bu değerler göreceli olabilmektedir. Söz konusu göreceliliği ortadan kaldırmak adına hatanın alt ve üst sınırları belirlenir, genelde (-) 1 ile (+) 1, 0 ile 1 veya 0 ile 100 seçilir (Joshi, 2020).

1.13.5. Doğruluk ve Hata

Test edilen değer ile tahmin edilen değerlerin yakınlığı karşılaştırılarak doğruluk oranı elde edilmekte olup, model başarı ölçülmesinde kullanılan en popüler yöntemdir (Şekil 1.60) (Danjuma, 2015). Doğruluk oranı hesaplamasında aşağıda formülasyonu görüldüğü üzere, doğru sınıflandırılmış örneklerin (TP + TN), toplamdaki örnek sayısına (TP+ TN+ FP+ FN) oranı kullanılmakta olup, hata oranı ise (1 - Doğruluk oranı) şeklindedir, farklı bir ifadeyle yanlış sınıflandırılmış örneklerin (FP + FN), toplamdaki örnek sayısına (TP+ TN+ FP+ FN) oranıdır (Şekil 1.61).

Tahmin Verileri	Gerçek Veriler	
	1-0 Matrisi	
1	TP	FP
0	FN	TN

Şekil 1.60. Gerçek ve tahmin değerleri matrisi (2*2)

TP (True Positive - Doğru Pozitif): Doğruyu doğru tahmin etme

TN (True Negative - Doğru Negatif): Doğruyu yanlış tahmin etme

FP (False Positive - Yanlış Pozitif): Yanlış doğru tahmin etme

FN (False Negative - Yanlış Negatif): Yanlış yanlış tahmin etme

$$\text{Doğruluk} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Hata} = \frac{FP + FN}{TP + FP + TN + FN}$$

Şekil 1.61. Doğruluk ve hata formülasyonu

1.13.6. Kesinlik

Gerçekte pozitif olan örnek sayısının (TP) pozitif olarak tahmin edilen örnek sayısına (TP + FP) oranı kesinlik değerini vermekte olup, hedeflenenin ne kadar elde edildiğinin göstergesidir (Şekil 1.62).

$$Kesinlik = \frac{TP}{TP + FP}$$

Şekil 1.62. Kesinlik formülasyonu

1.13.7. Anma

Gerçekte pozitif olan örnek sayısının (TP), sonucu doğru (TP + FN) tahmin edilen örnek sayısına oranı anma değerini vermekte olup, yapılan tüm doğru tahminler arasında ne kadarının içerdiği ortaya konulmaktadır (Şekil 1.63).

$$Anma = \frac{TP}{TP + FN}$$

Şekil 1.63. Anma formülasyonu

1.13.8. Özgünlük

Özgünlük, (TN) ile (TN + FP)'nin oranı ile elde edilmekte olup, anmanın tam aksi bir işlem gerektirmektedir (Şekil 1.64).

$$Özgünlük = \frac{TN}{TN + FP}$$

Şekil 1.64. Özgünlük formülasyonu

1.13.9. F-Ölçütü

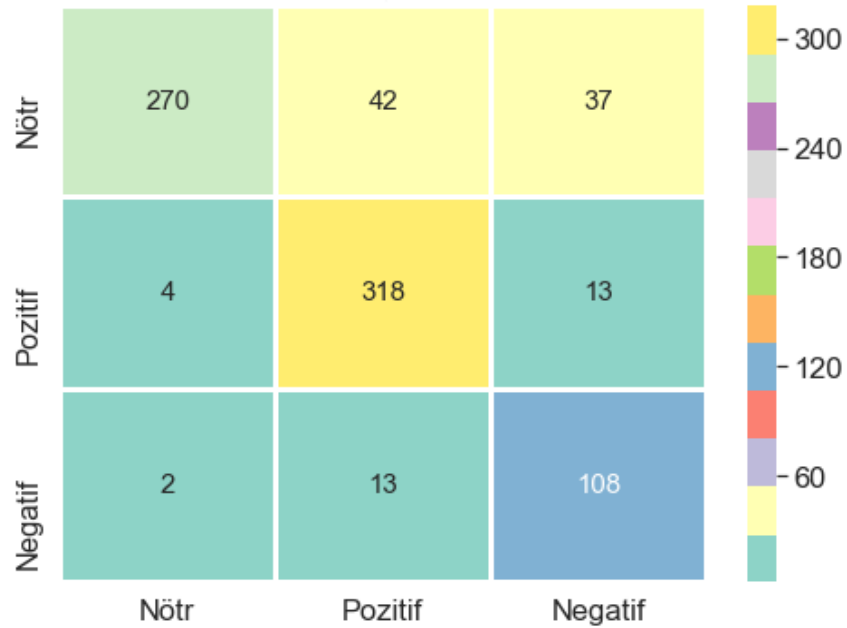
Matematiksel olarak F-ölçütü, kesinlik ve anmanın ağırlıklı harmonik ortalaması olup, en iyi değeri 1, en kötüsü ise 0 olmaktadır (Şekil 1.65). Zaman zaman F1 skoru olarak da anılmaktadır.

$$F \text{ Ölçütü} = \frac{2 * \text{Anma} * \text{Kesinlik}}{\text{Anma} + \text{Kesinlik}}$$

Şekil 1.65. F ölçütü formülasyonu

1.13.10. Karışıklık Matrisi

Özellikle sınıflandırma temelinde yapılan işlemlerde, doğruluğu hesaplamak için karışıklık matrisi kullanılmakta olup, elde edilen değerler n*n boyutunda matris üzerinde gösterilmektedir. Modelleme sonrası ortaya çıkan tespitlerin kategorik biçimde rahatlıkla değerlendirilebilmektedir.

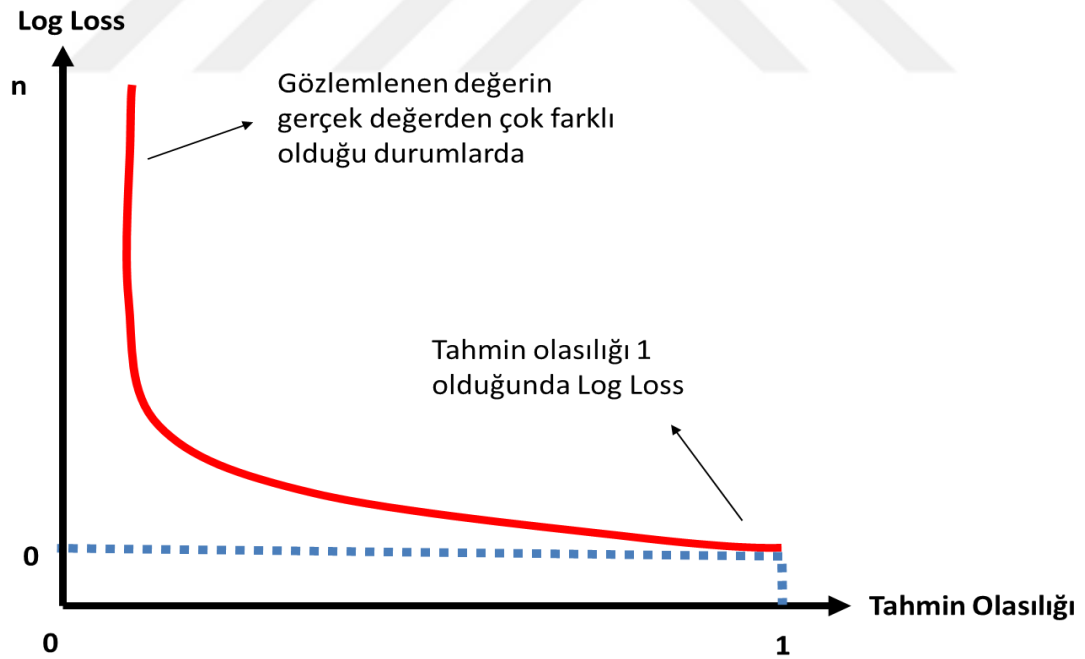


Şekil 1.66. Örnek gerçek ve tahmin değerleri matrisi (3*3)

Şekil 1.66’da polarite analizini içeren 3*3 örnek matris üzerinde görüldüğü üzere, negatif sınıflandırmada kısmen doğruluğun düşük olduğu (%68,35), ardından gelen pozitif sınıflandırmada doğruluğun (%85,25) ve nötr sınıflandırmada ise doğruluğun (%97,82) olduğu görülmektedir.

1.13.11. Logaritmik Kayıp

Logaritmik kayıp, Log loss veya çapraz entropi kaybı (cross-entropy loss) olarak da bilinmekte olup, temel olarak olasılık tahminleri üzerinde tanımlanır ve girdinin 0 ile 1 arasında bir olasılık değeri olduğu bir sınıflandırma modelinin performansını ölçer (Ugur, 2019). Temel yaklaşımı yanlış sınıflandırmanın cezalandırılmasıdır, gözlemlenen değer ile tahmini değer arasındaki fark yüksek çıktığında tahmin olasılığı da düşük çıkacaktır, bunun yanı sıra çok sınıflı verisetlerinde de çalışabilmektedir (Şekil 1.67).



Şekil 1.67. Log Loss ve tahmin ilişkisi

1.13.12. Kappa

Verilerin güvenilirlik ve geçerliliğin sağlanması adına oldukça sık kullanılan istatistik türüdür, gözlemlenen değerin şans eseri oluşup oluşmadığını inceler (Şekil 1.68).

$$Kappa = \frac{\text{Gözlemlenme Olasılığı} - \text{Rastgele Gözlemlenme Olasılığı}}{1 - \text{Rastgele Gözlemlenme Olasılığı}}$$

Şekil 1.68. Kappa formülasyonu

Çizelge 1.2. Kappa değerlerinin yorumlanması

Kappa Değeri	Değerlendirilmesi Gereken Yorum
< 0	Hiç uyum olmaması
0.0 — 0.20	Önemsiz derecede uyum
0.21 — 0.40	Zayıf derecede uyum
0.41 — 0.60	Orta derecede uyum
0.61 — 0.80	İyi derecede uyum
0.81 — 1.00	Çok iyi derecede uyum

Kappa değerleri yukarıda formülasyonu belirtildiği şekilde hesaplanarak, çizelge 1.2’de yer değerler üzerinden yorumlanmaktadır (McHugh, 2012). Kappa değeri 1’e yaklaştıkça uyumun çok yüksek olduğu ve tesadüfi öğrenmenin azaldığı görülmektedir.

1.14. Metin Madenciliği

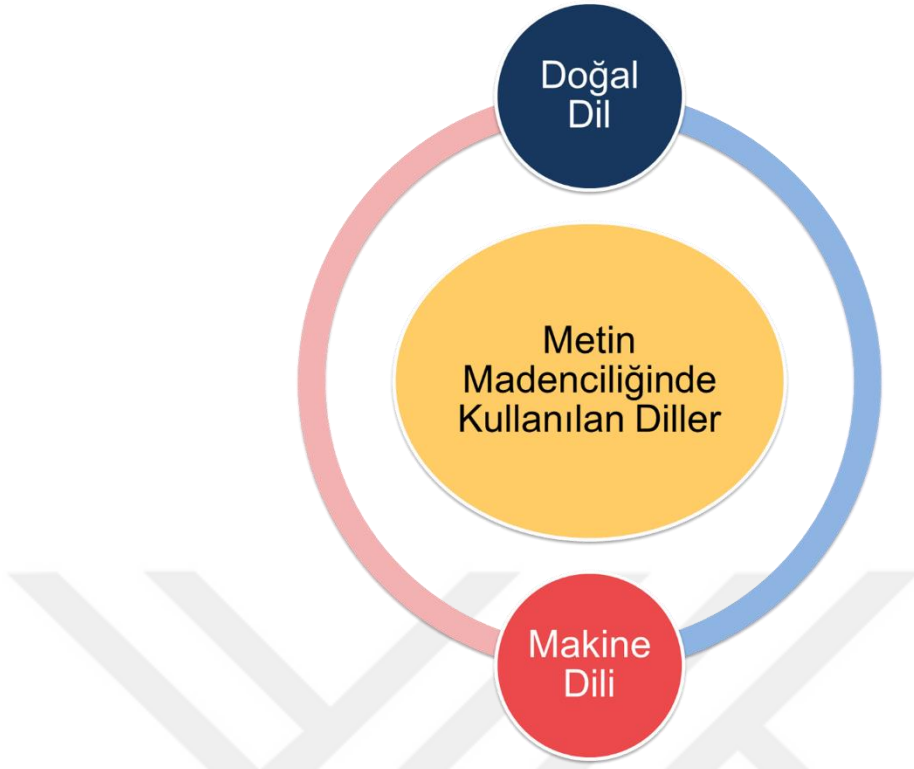
Metin madenciliği (TM - Text Mining) kavramı, literatürde akıllı metin çözümlemesi (intelligent text analysis), metin veri madenciliği (text data mining) ya

da metinde bilgi keşfi (Knowledge discovery in text - KDT) olarak da isimlendirilmekte olup, metin üzerinden anlamlı bilginin çıkarımına odaklanmaktadır (Salloum ve ark., 2018).

Sayısal veri işlemlerinden en büyük farkı, metin verilerinin genelini yapılandırılmamış olması ve eldeki yapılandırılmamış verilerin de sayısal boyuta taşınarak ilişkilerin keşfedilmesidir. Bu sebeple metin madenciliğinde, veri madenciliğine göre doğrudan özellik çıkarımı çok daha zor bir işlem olarak karşımıza çıkmaktadır. Söz konusu zorluğun üstesinden gelebilmek adına farklı bilim alanları ile de ilişkili multi disiplinli bir alan konumundadır.

Dünya üzerindeki verilerin çoğunluğunun metin şeklinde tutulmuş olması da, analiz anlamında metin madenciliğinin önemini arttırmaktadır. Fakat metin verilerinin yapılandırılmamış ve üretimi oldukça basit bir veri türü olduğu da dikkate alınması gereken bir durumdur (Allahyari ve ark., 2017). Söz konusu veriler insanlar tarafından çok rahatlıkla algılanmakta iken, makineler için oldukça zordur. Çünkü metin verileri içerisinde farklı dil ve lehçeleri de konu alan farklı nitelik ve versiyonlarda bilgi bulunması, makineler için okunabilir seviyeye indirgenmesini zorlaştırmaktadır. Metin içindeki bilgilerin önemi de dikkate alındığında, metin verilerinin makinelerce okunarak çeşitli işlemler yapılması maksadıyla, özellikle veri madenciliği, yapay zekâ ve makine öğrenmesi temelinde her geçen gün yeni yöntem ve algoritmalar tasarlanmaya çalışılmaktadır.

Metin üzerinde analizler yapılarak, hipotezler ve eğilimler ortaya konularak, veri örüntüsü hakkında anlamlı bilgi çıkarımı yapılmaktadır. Bu kapsamda anlamlı bilginin elde edilmesinde; bilgi getiri, hece/kelime analizi, kelime frekans dağılımı, örüntü tanıma, etiketleme, bilgi çıkarımı, eğilim analizi, ilişkilendirme madenciliği, özetleme, belgelerden anlamlar çıkarma, makine çeviri, görselleştirme, sınıflama ve kümeleme gibi yöntemler kullanılmaktadır (Allahyari ve ark., 2017; Talib ve ark., 2016; Žižka ve ark., 2019). Söz konusu yöntemler incelendiğinde metin madenciliğinin disiplinlerarası bir alan olduğu ve diğer bilim alanlarıyla sıkı bir ilişkisi olduğu ortadadır.

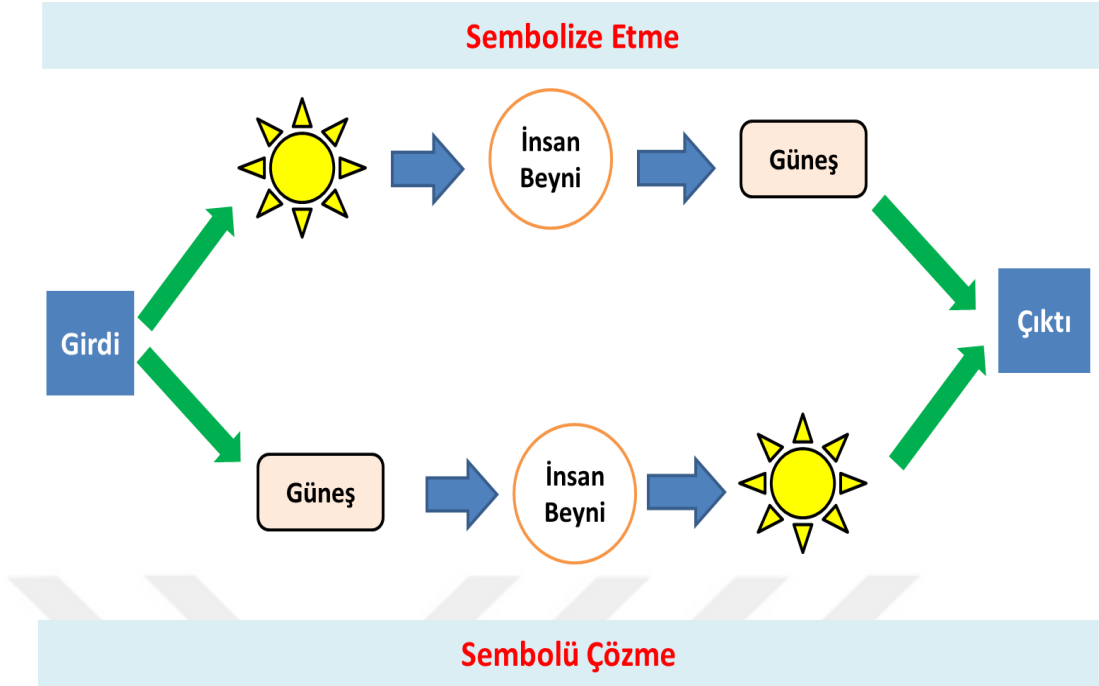


Şekil 1.69. Metin madenciliğinde kullanılan diller

Metin madenciliği kapsamında, insanlar olarak kendi aramızda iletişimde kullanıldığımız doğal dil ile bilgisayar sistemleriyle iletişimde kullandığımız makine dili olmak üzere iki dil bulunmaktadır (Şekil 1.69) (Lane ve ark., 2019). Makine dili yapılandırılmış veri yapısını ihtiva etmekte olmasına karşın, doğal diller çok daha karmaşık bir yapı ortaya çıkartmaktadır. Bu kapsamda müteakip konu başlıklarında, doğal dil ve makine dili konularına değinilecektir.

1.14.1. Doğal Dil

İnsanoğlunun dünyaya geldiği ilk andan günümüze kadar, kendi arasında iletişim kurabilme amacıyla kullandığı görsel, sesli veya yazılı çıktılar doğal dil olarak kabul edilmektedir. Farklı bir ifadeyle doğal dil kavramı, herhangi bir bilgisayar girdisi olmaksızın, doğal yollarla oluşarak, süreç içerisinde iletişim yoluyla gelişen ve geliştirilen dildir (Sarkar, 2019).



Şekil 1.70. İletişimde sembolize etme ve sembolleri çözme

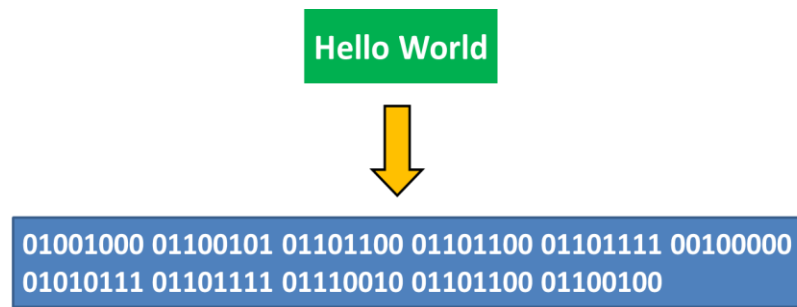
Doğal dillerde kullanılan yapılar, çeşitli referans noktalarına göre isimlendirmeden ortaya çıkmakta olup, bu süreç başlı başına bilişsel bir süreç olarak karşımıza çıkmaktadır (Şekil 1.70). Günümüzde %90'ı 100.000 nüfusun altında konuşanı olan 7117 adet doğal dilin olduğu tahmin edilmekte olup, bunların çoğunun yazma, konuşma ve sembolleri kendi aralarında farklılık göstermektedir (Ethnologue, 2020). Örneğin Türkçede insanların bilgi birikimini arttırmak için okuma yaptıkları nesne “kitap”, bunları koydukları dolap ise “kitaplık” olarak isimlendirilirken; İngilizcede; “book” ve “bookshelf”, Rusçada; “книга” ve “книжная полка”, Arapçada “كتاب” ve “الكتب رف” ve Çince de ise; “书” ve “书架” olarak sembolize edilmektedir. Sembolize etmede farklılıklar olsa da duyguları ifade ederek iletişime geçmede bütün doğal diller benzer bir vazife görmektedir.

Doğal dilin en büyük zorluğu ise yeterince yapılandırılmış olmamasından kaynaklanmaktadır. Yanlış ya da eksik söylenmiş ya da yazılmış olsa bile insanlar doğal dillerden anlam çıkarabilmekte olup, söz konusu anlamı makineden çıkarabilmek ise zorlaşmaktadır. “Arkadaşlarla top oynadık.” cümleciğinden ancak

kısmen anlam çıkarmak mümkündür. Söz konusu örneklemede geçen “top” kelimesi, basketbol, voleybol, futbol vb. hangi kategoride yer alan top olduğu ya da topla hangi oyunun oynandığını çıkarım yapmak zorken, kişiyi tanıyan ve futbola ilgisi olduğunu bilen birisi direk olarak, kişini futbol oynadığı çıkarımını yapabilecektir.

1.14.2. Makine Dili

Makine dilinde yapılan işlemler, her biri 0 ile 1 bitlik değere sahip olan 8 adet komut yapısına sahip olan değer elde edilmesidir (Özgül, 2016). Verilerin her birinin ikili (binary) seviyede olan karşılığı ile işlemcinin, istenilen komutu yapmasını istenmektedir. Programlama dillerinde olduğu üzere derleyiciye ihtiyaç duymadığından doğrudan donanım yapısını kullanmaktadır. Bu sebeple de birinci kuşak olarak anılmaktadır. Birinci kuşak makine dillerini öğrenmek oldukça zordur. Bu sebeple de makine dilinin öğrenimini kolaylaştırma amacıyla ortaya atılan ikinci kuşak programlama dilleri ise derleme (assembly) olarak anılmakta olup, insanların daha kolay anlayabileceği ara yazılımlarla, dil yapısını bilgisayarın işlemcisinin diline çevirmektedir. Çünkü bilgisayar, bilgisayar programları oluşturmak için kullanılan programlama dillerini doğrudan anlayamaz, bu nedenle program kodunun derlenmesi gerekmektedir.



Şekil 1.71. Doğal dilin makine diline dönüşümü

Makine dili veri yapısı itibariyle yapılandırılmıştır. Bütün kurallar nettir, sisteme yapılacak girdilerin tek karşılıkları vardır, makine dillerinin tanımlı olmadığı bir kod parçası yazıldığında sistem hata verir (Lane ve ark., 2019). Bu sebeple de şekil 1.71’de

görüldüğü üzere makine dilindeki verinin analizini yapmak, yapılandırılmamış dil yapılarına göre daha kolaydır.

1.15. Metin Verilerini İşleme ve Anlama

Dünya üzerinde yer alan verilerin % 90'ı yapılandırılmamış şekilde olup, söz konusu veriler; resim, video ve metin gibi birçok farklı biçimde olabilmektedir (Kulkarni ve Shivananda, 2019). Söz konusu verilerin yapılandırılmamış konumda olması, veri gürültüsü (noisy data) sorununu da beraberinde getirmektedir. Bu konuda veri dünyasında sıkça söylenilen “Garbage in, garbage out (GIGO)” özdeyişinde olduğu üzere, düşük kalitede yer alan verilerle yapılan algoritmik öğrenmeler de düşük düzeyde olacaktır. Örneğin “sellam dstum, umarim iyisindir.” cümle yapısını aynı dili konuşan bir insanın dahi anlaması kimi zaman mümkün olmamaktadır. Veri madenciliğinin başarılı şekilde yapılabilmesi için en önemli adımlardan biri veri üzerinde yapılan temizlik ile verinin içeriği ve yapısı gibi temel etmenlerinin ortaya konulmasıdır, farklı bir ifadeyle mümkün olduğunca gürültüden arındırılmasıdır. Makine öğrenimi veya derin öğrenme algoritmaları ile metin üzerinde işlem gerçekleştirmek için öncelikle metin verilerinin işlenmesi ve yorumlamasının yapılması gerekir. İster denetimli veya yarı denetimli, ister denetimsiz olsun, tüm makine öğrenimi algoritmaları sayısal olan giriş özellikleri ile çalışır. Metin verilerini sayısallaştırmadan önce ilk olarak metin verilerinin temizlenmesi, normalleştirilmesi ve ön işleme tabi tutulması gerekmektedir.

1.15.1. URL Etiketleri Kaldırma

URL (Uniform Resource Locator), tekdüzen kaynak bulucu olarak Türkçede tanımlanmakta olup, sayfaların internet üzerindeki yerini ifade etmede kullanılmaktadır (Sahoo ve ark., 2017). URL kullanımı arama motoru optimizasyonu için oldukça gerekli olmakla birlikte duruma dayalı olarak metin madenciliğinde,

metnin dışında bir bağlantı noktası yaratması sebebiyle metinden temizlenmesi önem taşır (Sarkar, 2019).

```
import re
#remove hyperlinks
df1 = "Sadece metin olan görünecek, https://kriminalkonular.blogspot.com/"

df1 = re.sub('https:\\\\.*', '', df1)

print(df1)

Sadece metin olan görünecek,
```

Şekil 1.72. URL etiketi kaldırılması örneği

Şekil 1.72 incelendiğinde, df1’de URL linki ve normal metin bulunmaktadır. Re (Regular expression) kütüphanesi kullanarak, sub metoduyla http adresi boşluk olarak görüntülenmiş, metin ise aynen görünmeye devam etmiştir.

1.15.2. Metni Bölme

Ham veriyi incelemek için öncelikle onu parçalara ayırmak çok daha anlamlı sonuçlar ortaya çıkmasını sağlayacaktır. Bu kapsamda Python özelinde “split” yöntemi, inceleyeceğimiz metin içinde yer alan karakterleri parçalara ayırmada aktif olarak kullanılmaktadır.

```
metin='Hayatımızın giderek sanal hale geldiğini söyleyebiliriz.'

metin.split()

['Hayatımızın', 'giderek', 'sanal', 'hale', 'geldiğini', 'söyleyebiliriz.']

for i in metin.split():
    print(i)

Hayatımızın
giderek
sanal
hale
geldiğini
söyleyebiliriz.
```

Şekil 1.73. Metin bölme örneği

Şekil 1.73 incelendiğinde, split metoduyla boşluk karaktere tanımlı şekilde metnin bölündüğü görülmekte olup, son kelimedeki nokta (“.”) simgesi mevcuttur. Bu örneğe has “.” simgesi de bölme parametresi olarak eklenebilir fakat büyük korpustaki rasyonel bir yöntem olmayacaktır.

1.15.3. Dizgeciklere Ayırma (Tokenleştirme)

Tokenleştirme, bir karakter yapısının token adı verilen cümleler/cümle veya kelimeler/kelime öbeğine ayrılması işlemidir. Metnin kelime ya da cümlecik boyutuna indirgenerek, analize hazırlanması için önemli bir aşama olup, aynı zamanda noktalama işaretlerini de ayrı karakter ögesi olarak tanımladığından, temizleme işleminde de kullanılır (Allahyari ve ark., 2017).

```
from nltk.tokenize import sent_tokenize, word_tokenize

metin='Hayatımızın giderek sanal hale geldiğini söyleyebiliriz.'

word_tokenize(metin)

['Hayatımızın', 'giderek', 'sanal', 'hale', 'geldiğini', 'söyleyebiliriz', '.']

sent_tokenize(metin)

['Hayatımızın giderek sanal hale geldiğini söyleyebiliriz.']
```

Şekil 1.74. Tokenleştirme örneği

Şekil 1.74 incelendiğinde, NLTK (Natural Language Tool Kit) kütüphanesi kullanılarak “word_tokenize” metoduyla cümlelerin tokenleştirildiği ve tekrar eski hale getirilmek için de “sent_tokenize” metodunun kullanıldığı görülmektedir. Noktalama işaretleri gibi özellik arz eden karakterlerin nltk kütüphanesinde olması sebebiyle, örnek üzerinde “.” işaretin ayrıldığı belirlenmiştir. Tokenleştirmede “word_tokenize” yaygın seçim olmakla birlikte, farklı karakter öğelerini uzaklaştırmak için “Treebank tokenizer”, “Penn Treebank”, “WordPunctTokenizer” ve “RegexTokenizer” parametreleri de bulunmaktadır (Bengfort ve ark., 2018).

1.15.4. Karakter Kodlama

Veri madenciliği adına üzerinde değerlendirilme yapılması gereken temel alanlardan biri de karakter kodlamasıdır. Bilgisayar olarak adlandırdığımız yapı, üzerinden elektrik devreleri geçen bir sistem olup, devrede elektrik varsa yaklaşık 5 volt, yoksa da 0 volt olarak işlem görmektedir (Özgül, 2016). Söz konusu değerler ikili (binary) sayma sistemine göre 0 ve 1 olarak karakterize edilerek, kodlanmışlardır. Zamanla 0 ve 1 değerlerinin kodlamada yetersiz kalması ve metin işlemleri için herkesin farklı kod sayıları kullanmasıyla birlikte, 1960'lı yıllarda ASCII (American Standard Code for Information Interchange - Bilgi Alışverişi için Standart Amerikan Kodu) karakter dizisi, IBM çalışanı Bob Bemer tarafından ortaya atılmıştır (Özgül, 2016).

Çizelge 1.3. ASCII karakter tablosu

0 '\x00'	1 '\x01'	2 '\x02'	3 '\x03'	64 '@'	65 'A'	66 'B'	67 'C'
4 '\x04'	5 '\x05'	6 '\x06'	7 '\x07'	68 'D'	69 'E'	70 'F'	71 'G'
8 '\x08'	9 '\t'	10 '\n'	11 '\x0b'	72 'H'	73 'I'	74 'J'	75 'K'
12 '\x0c'	13 '\r'	14 '\x0e'	15 '\x0f'	76 'L'	77 'M'	78 'N'	79 'O'
16 '\x10'	17 '\x11'	18 '\x12'	19 '\x13'	80 'P'	81 'Q'	82 'R'	83 'S'
20 '\x14'	21 '\x15'	22 '\x16'	23 '\x17'	84 'T'	85 'U'	86 'V'	87 'W'
24 '\x18'	25 '\x19'	26 '\x1a'	27 '\x1b'	88 'X'	89 'Y'	90 'Z'	91 '['
28 '\x1c'	29 '\x1d'	30 '\x1e'	31 '\x1f'	92 '\\'	93 ']'	94 '^'	95 '_'
32 '`'	33 '!'	34 '""'	35 '#'	96 '``'	97 'a'	98 'b'	99 'c'
36 '\$'	37 '%'	38 '&'	39 '""'	100 'd'	101 'e'	102 'f'	103 'g'
40 '('	41 ')''	42 '*'	43 '+'	104 'h'	105 'i'	106 'j'	107 'k'
44 ','	45 '-'	46 '.'	47 '/'	108 'l'	109 'm'	110 'n'	111 'o'
48 '0'	49 '1'	50 '2'	51 '3'	112 'p'	113 'q'	114 'r'	115 's'
52 '4'	53 '5'	54 '6'	55 '7'	116 't'	117 'u'	118 'v'	119 'w'
56 '8'	57 '9'	58 ':'	59 ';'	120 'x'	121 'y'	122 'z'	123 '{'
60 '<'	61 '='	62 '>'	63 '?'	124 ' '	125 '}'	126 '~'	127 '\x7f'

ASCII karakter yapısında çizelge 1.3’de görüldüğü üzere 128 adet karakter mevcut olup (Karakter dizisi 0 ile 127 arasında olup, 0’da bir ögedir.), bu da söz konusu sistemin ($2^7=128$), 7 bitlik bir karakter yapısını ihtiva ettiğini göstermektedir. Karakter yapısındaki harf yapıları incelendiğinde, mevcut kodlama yapısının İngilizceye karşılık geldiği rahatlıkla anlaşılmaktadır. Bu sebeple de Türkçeye özgü olarak kullanılan ü, ö, ş, ğ gibi harf karakterlerinin ASCII sisteminde karşılık değeri bulunmamaktadır (Türkçe dilinde Windows-1254 (cp1254) gibi). Bu sebeple “genişletilmiş ASCII” olarak, 8 bitlik sistem içerisinde ($2^8=256$), 256 karakterlik yapılar kullanılmaya başlanmış olup, Dünya dillerinin hepsinin çok farklı karakterleri ihtiva etmesi sebebiyle dillere özgü genişletilmiş ASCII’ler kullanılmakta olup, standart bir genişletilmiş ASCII’den bahsedilememektedir.

ASCII sisteminde her bir harf 1 baytlık karakter yapısına, o da 7 veya 8 bit içinde yer alan değere karşılık gelmekteyken, bu sistemde yer almayan karakterler 1 bayt dışındaki değerlerle anılmaktadır. ASCII sisteminde tek bir baytla bütün dillerdeki karakterler kodlanamayacağı için ASCII dışı karakterler, birden fazla bayt ile kullanılmaktadır. Örneğin ASCII sisteminde tanımlanan “s” harfi, 1 baytlık “s” olarak çıktı üretebiliyorken, ASCII sisteminde tanımlanmayan “ş” harfi, 2 baytlık (b'\xc5\x9f') çıktı üretmektedir.

İngilizce dilini kullanmayan çalışma sahalarında, ASCII sisteminde yer alan kodlama yetersizliğini ortadan kaldırmak için ortaya atılan Unicode kavramı ile birlikte, ucu bucağı sınırsız seviyede olan bir yapı devreye girmiştir (Gleave ve Steinruecken, 2017). Bu sistemle Unicode, doğrudan doğruya karakterleri kodlamayacak, sadece karakter tanımlaması yapacaktı. Unicode sistemi içinde UTF-1, UTF-7, UTF-8, UTF-16 ve UTF-32 isimli kod çözücüler (encoding) bulunmakta olup, bunlar arasında en yaygın kullanılanı UTF-8 kod sistemidir. UTF-8 kod sistemi, UNICODE sistemi içerisinde 1- 4 bayt arasındaki değerleri kullanır.

1.15.4. Metinde Karakter Kodlamanın Tespiti

Metin verilerinin işlenmesinde en sık karşılaşılan durumlardan biri eldeki verinin hangi kod sistemi üzerinden kodlandığıdır. Çünkü Türkçe başta olmak üzere, ASCII kod sisteminin kullanılmadığı durumlarda hangi kod sistemi üzerinden işlem yapıldığı, metnin çözümü (decode) yapılabilmesi adına önemli olmaktadır. ASCII karakter yığını dışında, ideal olan UTF-8 kullanılmasıdır fakat metnin öncelikle algılanıp, standart kodlamaya dönüştürülmesi gerekmektedir (Hardeniya ve ark., 2016). Fransızca paylaşılan metnin “?????uvrent au quotidien” farklı kodlamadaki görüntüsü parantez içinde görüldüğü şekilde olabilmektedir. Bu kapsamda Chardet gibi kütüphaneler aracılığıyla, kullanılan dil tespit edilebilmekte ve gerekli olan dönüşüm yapılabilmektedir.

```
import chardet
Türkçe = 'Şeyma Buğra ve Güzel'.encode('utf-8')
print(Türkçe)
print(chardet.detect(Türkçe))

İngilizce = 'life is good'.encode('utf-8')
print(chardet.detect(İngilizce))
```

```
b'\xc5\x9eeyma Bu\xc4\x9fra ve G\xc3\xbczel'
{'encoding': 'utf-8', 'confidence': 0.87625, 'language': ''}
{'encoding': 'ascii', 'confidence': 1.0, 'language': ''}
```

```
kodçözüm = ['ASCII', 'cp1254', 'latin-1', 'UTF-8', "Big5", "Windows-1251"]
harf = 'Ö'
for text in kodçözüm:
    try:
        print("{} karakteri {} ile {} olarak "
              "ve {} sayısıyla temsil edilmektedir.".
              format(harf, text, harf.encode(text), ord(harf)))
    except UnicodeEncodeError:
        print("{} karakteri {} ile temsil edilemez.".format(harf, text))
```

```
'Ö' karakteri ASCII ile temsil edilemez.
'Ö' karakteri cp1254 ile b'\xd6' olarak ve 214 sayısıyla temsil edilmektedir.
'Ö' karakteri latin-1 ile b'\xd6' olarak ve 214 sayısıyla temsil edilmektedir.
'Ö' karakteri UTF-8 ile b'\xc3\x96' olarak ve 214 sayısıyla temsil edilmektedir.
'Ö' karakteri Big5 ile temsil edilemez.
'Ö' karakteri Windows-1251 ile temsil edilemez.
```

Şekil 1.75. Karakter tespit ve çözümü

Şekil 1.75’de Chardet çözümlemesi ile dil belli bir güven aralığında (confidence) tahmin edilebilmekte ve kod çözümü için aynı karaktere farklı dillerin farklı sonuçlar verdiği de görülmektedir.

1.15.5. Aksanlı Harfleri Kaldırma / Değiştirme

Genellikle herhangi bir metin topluluğunda, özellikle yalnızca İngilizceyi analiz etmek istiyorsanız, özel isim ya da o dile ait özel karakterlerden dolayı aksanlı karakterler / harfler ile karşılaşılabilir (Sarkar, 2019). Bu karakterlerin standartlaştırılması gereklilik olmaktadır. Bu kapsamda “ğ” harfinin “g” harfine dönüştürülmesi, “ı” harfinin kaldırılması bu konuya örnek olabilir.

```
import unicodedata

def remove_accented_chars(text):
    text = unicodedata.normalize('NFKD', text).encode('ascii', 'ignore').decode('utf-8', 'ignore')
    return text

remove_accented_chars('Ateş Buğra Güzel Kâğıt')
```

'Ates Bugra Guzel Kagt'

Şekil 1.76. Aksanlı karakterlere işlem

Şekil 1.76’da Unicodedata Kütüphanesi kullanılarak, aksanlı kelimeler üzerinde UTF-8’e göre yapılan dönüşüm işlemi görülmektedir.

1.15.6. Kullanılan Dilin Tespiti

Metin içinde kullanılan dil veya dillerin tanımlanmaması metin analizi için oldukça önemli bir sorun olup, tanımlamadaki başarısızlık karşımıza sınıflandırma

sorunu olarak çıkmaktadır (Žižka ve ark., 2019). Bu sebeple en basit yaklaşımla dilin bilinen sözcükleriyle sonuç ortaya konulabilmekte olup, karşılaştırmaya konu metnin fazlalığı tespitteki güven aralığını arttıracaktır.

```
from textblob import TextBlob
a = TextBlob("bonjour")
b = TextBlob("merhaba")
c = TextBlob("سلام")
d = TextBlob("ahoj")
e = TextBlob("hola")
print(a.detect_language())
print(b.detect_language())
print(c.detect_language())
print(d.detect_language())
print(e.detect_language())
```

```
fr
tr
fa
cs
es
```

Şekil 1.77. Kullanılan dilin tespiti

Şekil 1.77’de Textblob Kütüphanesi kullanarak, Fransızca(FR), Türkçe(TR), Farsça(FA), Çekçe(CS) ve İspanyolca(ES) “merhaba” anlamını kelimelerin tespiti örnek olarak yapılmış olup, Polyglot, Chardet, Langdetect, Langid, Fasttext, Cld3 ve Spacy gibi kütüphanelerle de aynı işlemi gerçekleştirmek mümkündür.

1.15.7. Özel Karakterlerin Kaldırılması

Metin içinde alfabetik harf gösterimleri dışında yer alan noktalama işaretleri, sayısal ifadeler, emoji gibi özel karakterlerin metinden uzaklaştırılması ön işlem aşamasında yapılmaktadır (Sarkar, 2019). Söz konusu ifadelerin kaldırılmaması metin içinde gürültüye sebep olmaktadır (Hardeniya ve ark., 2016).

```
def ozel_karakterlerin_kaldirilmesi(text, remove_digits=False):

    pattern = r'^a-zçğışöüA-ZÇĞİŞÖÜÖ-9\s]' if not remove_digits else r'^a-zçğışöüA-ZÇĞİŞÖÜÖ\s]'
    text = re.sub(pattern, '', text)
    return text

ozel_karakterlerin_kaldirilmesi("<<<Ortam@ çok ;) değişik!! 100+??", remove_digits=True)

'Ortam çok değişik '
```

Şekil 1.78. Türkçe örnek metinde özel karakterlerin kaldırılması

Şekil 1.78’de görüldüğü üzere Türkçede kullanılan (“ç”, “ğ”, “ş” gibi) harfler korunarak, özel karakterler metinden uzaklaştırılmıştır.

1.15.8. Kısaltmaların Genişletilmesi

Metin içinde kelime veya harf grupları zaman zaman çeşitli kısaltmalarla ifade edilmekte olup, söz konu kısaltmaların bir kısmı sadece baş harflerden oluşabiliyorken, kimisi de belli harf ya da hecenin metinden kaldırılması şeklinde oluşabilmektedir (Sarkar, 2019). Genelde resmi yazılarda az kullanılmakla birlikte, resmi olmayan günlük metinlerde oldukça sık karşılaşılmaktadır. Kısaltmalar tanımlanmadığı takdirde makine tarafından algılanamayacak ve değerlendirme dışı kalacaktır. Söz konusu tanımlanmama, küçük bir korpus içerisinde neredeyse hayati öneme haizken, büyük korpuslarda gözardı edilebilmektedir.

```

df1="TBMM 23 Nisan 1920'de açıldı."
df2="TÜİK istatistik temelli çalışır."
df3="Ben okula gitcem."
df4="You aren't a good man."
df5="You're a good man."
df1 = df1.replace('TBMM','Türkiye Büyük Millet Meclisi')
df2 = df2.replace('TÜİK','Türkiye İstatistik Kurumu')
df3 = df3.replace("gitcem",'gideceğim')
df4 = df4.replace("aren't",'are not')
df5 = df5.replace("You're",'You are')
print(df1)
print(df2)
print(df3)
print(df4)
print(df5)

```

```

Türkiye Büyük Millet Meclisi 23 Nisan 1920'de açıldı.
Türkiye İstatistik Kurumu istatistik temelli çalışır.
Ben okula gideceğim.
You are not a good man.
You are a good man.

```

Şekil 1.79. Kısaltma karakterlerin genişletilmesi

Şekil 1.79’da görüldüğü üzere metin içinde yer alan kısaltmalar, replace metoduyla olması gereken şekle dönüştürülmüştür. Ayrıca RegexpReplacer.replace () fonksiyonu ile herhangi bir değiştirme modelinin her örneğini, karşılık gelen değiştirme modeliyle değiştirilir. Yapılan bu işlemde, değişime konu kısaltma ((r'(\w+)\ve', 'g<1> have') ve (r'(\w+)\n\t', 'g<1> not')) yerine, açık yazılıştaki ifadesi yer alır (Hardeniya ve ark., 2016).

1.15.9. Büyük/Küçük Harf Dönüşümleri

Metin içinde geçen kelimelerin tamamının küçük veya büyük harf yapılarak veya başlık durumunda (title case – cümledeki her kelimenin ilk harfi büyük yazılır) normalizasyonu harf dönüşümlerinin temel amacıdır (Sarkar, 2019). Python özelinde söz konusu dönüşüm, lower, upper ve title metotlarıyla yapılabilmektedir.


```
# lowercase
text = 'DoKuZ Yıl öNcEsİNE MEKtup'
text.lower()
```

'dokuz yıl öncesine mektup'

```
# uppercase
text = 'DoKuZ Yıl öNcEsİNE MEKtup'
text.upper()
```

'DOKUZ YIL ÖNCESINE MEKTUP'

```
# titlecase
text = 'DoKuZ Yıl öNcEsİNE MEKtup'
text.title()
```

'Dokuz Yıl Öncesine Mektup'

Şekil 1.80. Büyük/küçük harf dönüşümleri

Şekil 1.80’de görüldüğü üzere metin içinde (text) yer alan ifade, tamamı küçük harf, tamamı büyük harf veya kelimelerin ilk harfi büyük olacak şekilde tekrar düzenlenmiştir.

1.15.10. Yinelenen Karakterleri Düzeltme

Metin içinde dilin gramer kurallarına uymayan, söz konusu dilin sözlüğünde yer almayan fakat ifade edildiğinde genelde insanlar tarafından algılanabilmekle birlikte, herhangi bir öğretim yapılmaksızın makinelerce anlaşılmayacak ifadeler mevcuttur. Bu kapsamda ifadelerin, makinelerin anlayabileceği şekle dönüştürülmesi önem taşımaktadır. Örneğin; “kaaleeeemmmmm” ifadesi insanlar tarafından anlaşılabilirken, makinelerin anlayabileceği dili de dönüştürülmesi gerekir. Bu aşamada iki yol mevcuttur, bunlardan tekrar eden harf karakterlerinin indirgenmesi, diğeri ise Wordnet ve benzeri kütüphaneler aracılığıyla sözlük yardımıyla benzerlik gösteren kelimenin tespit edilmesidir (Hardeniya ve ark., 2016; Sarkar, 2019).

```

eski_kelime = 'kaaleeeemmmm'
repeat_pattern = re.compile(r'(\w*)(\w)\2(\w*)')
match_substitution = r'\1\2\3'
step = 1
while True:
    # karakter tekrarlarının uzaklastirilmesi
    yeni_kelime = repeat_pattern.sub(match_substitution, eski_kelime)
    if yeni_kelime != eski_kelime:
        print('Adım: {} Kelime: {}'.format(step, yeni_kelime))
        step += 1
        # eski kelimenin guncellenmesi
        eski_kelime = yeni_kelime
        continue
    else:
        print("Düzeltilen Kelime:", yeni_kelime)
        break

```

```

Adım: 1 Kelime: kaaleeeemmmm
Adım: 2 Kelime: kaaleeeemmm
Adım: 3 Kelime: kaaleeeemm
Adım: 4 Kelime: kaaleeeem
Adım: 5 Kelime: kaaleeeem
Adım: 6 Kelime: kaaleem
Adım: 7 Kelime: kaalem
Adım: 8 Kelime: kalem
Düzeltilen Kelime: kalem

```

Şekil 1.81. Tekrar eden karakterlerin doğru düzeltilmesi

```

eski_kelime = 'happpppppyy'
repeat_pattern = re.compile(r'(\w*)(\w)\2(\w*)')
match_substitution = r'\1\2\3'
step = 1
while True:
    yeni_kelime = repeat_pattern.sub(match_substitution, eski_kelime)
    if yeni_kelime != eski_kelime:
        print('Adım: {} Kelime: {}'.format(step, yeni_kelime))
        step += 1
        eski_kelime = yeni_kelime
        continue
    else:
        print("Düzeltilen Kelime:", yeni_kelime)
        break

```

```

Adım: 1 Kelime: happpppppyy
Adım: 2 Kelime: happpppppy
Adım: 3 Kelime: happppppy
Adım: 4 Kelime: happpppy
Adım: 5 Kelime: happppy
Adım: 6 Kelime: happy
Adım: 7 Kelime: hapy
Adım: 8 Kelime: hapy
Düzeltilen Kelime: hapy

```

Şekil 1.82. Tekrar eden karakterlerin yanlış düzeltilmesi

```

from nltk.corpus import wordnet
eski_kelime = 'happpppppyy'
repeat_pattern = re.compile(r'(\w*)(\w)\2(\w*)')
match_substitution = r'\1\2\3'
step = 1
while True:
    if wordnet.synsets(eski_kelime):
        print("Final correct word:", eski_kelime)
        break
    yeni_kelime = repeat_pattern.sub(match_substitution, eski_kelime)
    if yeni_kelime != eski_kelime:
        print('Adım: {} Kelime: {}'.format(step, yeni_kelime))
        step += 1
        eski_kelime = yeni_kelime
        continue
    else:
        print("Wordnet Üzerinden Düzeltilen Kelime:", yeni_kelime)
        break

```

```

Adım: 1 Kelime: happpppppyy
Adım: 2 Kelime: happpppppy
Adım: 3 Kelime: happppppy
Adım: 4 Kelime: happpppy
Adım: 5 Kelime: happppy
Adım: 6 Kelime: happy
Adım: 7 Kelime: happy
Final correct word: happy

```

Şekil 1.83. Tekrar eden karakterlerin Wordnet üzerinden düzeltilmesi

Şekil 1.81’de görüldüğü üzere Türkçe “kaaleeeemmmmm” ifadesi üzerinde tekrar eden karakterler uzaklaştırılarak, “kalem” kelimesi elde edilmiştir. Türkçede dil yapısı itibariyle kelime içinde aynı karakter ögesi genelde peşi sıra gelmemektedir. Fakat İngilizce başta olmak üzere pek çok dilde bunun örnekleri mevcuttur. Şekil 1.82’de İngilizce yazımı yanlış yazılan “happpppppyy” ifadesinin tekrar eden harf karakterleri çıkarıldığında “hapy” kelimesi elde edilmiş olup, söz konusu ifadenin anlamsal karşılığı bulunmaktadır. Şekil 1.83’de ise İngilizce yazımı yanlış yazılan “happpppppyy” ifadesinin Wordnet kütüphanesi üzerinden kontrolü, eğer sözlükte varsa o kelimenin kullanılması, yoksa da tekrar eden harf karakterlerinin çıkarılması komutları verilmiş olup, “happy” sonucu elde edilmiştir.

1.15.11. Yazım Düzeltme

İncelemeye konu metin verisi içerisinde insan veya programlama hatasından kaynaklı yazım yanlışlıkları olabilmektedir. Söz konusu hata veri kalitesini olumsuz yönde etkileyerek, ana korpusun büyüklüğüne göre zaman zaman inceleme için hayati önem taşıyabilmektedir. Yazım düzeltme işlemi uygulanarak, yapılan hatanın tespiti, hatanın üstesinden gelmek için aday düzeltmelerin üretilmesi ve sıralaması metin verileri üzerine yapılmaktadır.

Düzeltilme işlemini incelemeye konu dilde mukayeseye konu n-gram olup olmadığına odaklanır, varsa doğru olarak kabul edilirken, yoksa da alternatif tespitler çıkarılmaya çalışılmaktadır (Nonghuloo ve Krishnamurthi, 2017; Žižka, 2019). İki başlıca yöntemle yapmak mümkündür, bunlardan ilki incelenecek dile ait olması gereken yazım niteliklerini taşıyan korpusu sisteme öğretmek, benzerlik oranı üzerinden tahminleme yapmak, diğeri ise mevcut kütüphanelerin kullanılmasıdır. İngilizce ve birçok dil için TextBlob, SymSpell, Pyspellchecker ve Autocorrect gibi kütüphaneler kullanılabilirken, Türkçede ise Zemberek, ITU Turkish NLP Web Service API ve Turkishnlp gibi kütüphaneler kullanılabilmektedir. Kullanılan dillerde mevcut dilin kullanım çeşidinin ya da yanlış yazımının neredeyse sonsuz büyüklükte bir veri kümesi doğuracağından dolayı, söz konusu kütüphaneler arasında kesin güvenilirlikle “daha iyi” şeklinde tahminleme yapmak imkânsızdır. Dil işleminin karışık yapısından kaynaklı hiçbir kütüphanenin mükemmel sonuçlar ortaya koyamadığı birçok çalışmada mevcuttur (Gudmundsson ve Menkes, 2018; Jo ve ark., 2019). Bu sebeple de en önemli konu, incelemeye konu veri setinin büyüklüğüdür. Veriseti büyüdükçe, yazım hatalarının incelemedeki etkisi giderek azalmaktadır. Aynı zamanda veri setindeki büyüklük derin öğrenmedeki başarı oranını da etkileyeceğinden, derin öğrenme ile düzeltme işlemleri kısmi başarı ile yapılabilmektedir (Otter ve ark., 2018; Sooraj ve ark., 2018).

```
from textblob import Word
w = Word('hapyp')
w.correct()
```

```
'happy'
```

```
w = Word('mountth')
w.spellcheck()
```

```
[('mouth', 0.5694444444444444),
 ('month', 0.2222222222222222),
 ('mounted', 0.1493055555555555),
 ('mount', 0.05902777777777776)]
```

Şekil 1.84. Yazım düzeltme ve alternatif değerler

Şekil 1.84 incelendiğinde Textblob Kütüphanesi üzerinden kelime bazında “hapyp” kelimesinin doğrusu “happy” olarak tahmin edilmişken, “mountth” kelimesinde ise alternatifler de istenmiş ve 4 farklı sonuç tahmini ortaya çıkarak, bunların arasında da en güçlüsü %56,944 ile “mouth” kelimesi olmuştur.

```
import turkishnlp
from turkishnlp import detector
obj = detector.TurkishNLP()
obj.download()
obj.create_word_set()

lwords = obj.list_words("Telefnu rehberree kayııt etmeiy düşünüyodum amaaa vazgçtm")
print(obj.auto_correct(lwords))
```

```
Download is successful
```

```
['telefonu', 'rehberlere', 'kayıt', 'etmeyi', 'düşünüyodum', 'amaca', 'vazgeçti']
```

Şekil 1.85. Türkçe cümle düzeltme

Şekil 1.85 incelendiğinde, Turkishnlp Kütüphanesi üzerinden yapılan cümle analizinde yazılmaya çalışılan cümle “Telefonu rehberlere kayıt etmeyi düşünüyordum ama vazgeçtim” iken, ekrana “Telefnu rehberree kayııt etmeiy düşünüyodum amaaa vazgçtm” yazıldığı ve sistemin de söz konusu cümleyi “telefonu rehberlere kayıt

etmeyi düşünüyordum amaca vazgeçti” şeklinde çıktı ürettiği belirlenmiş olup, ana anlam olarak iyi tahminler olsa da kesin güvenilirlikle sonuç elde edilememektedir.

```
from textblob import Sentence
w = Sentence("I ws gonig to coooome to u tody, but I ddn't, mybe I'd comme tomrrw.")
w.correct()

Sentence("I was going to coooome to u body, but I don't, maybe I'd comme tomorrow.")
```

Şekil 1.86. İngilizce cümle düzeltme

Şekil 1.86 incelendiğinde, TextBlob Kütüphanesi üzerinden yapılan cümle analizinde yazılmaya çalışılan cümle "I was going to come to you today, but I didn't, maybe I could come tomorrow." iken, ekrana "I ws gonig to coooome to u tody, but I ddn't, mybe I'd comme tomrrw." yazıldığı ve sistemin de söz konusu cümleyi "I was going to coooome to u body, but I don't, maybe I'd comme tomorrow." şeklinde çıktı ürettiği belirlenmiş olup, tıpkı Türkçe cümle analizinde olduğu gibi ana anlam olarak iyi tahminler olsa da kesin güvenilirlikle sonuç elde edilememektedir.

1.15.12. Kök Bulma

Kök bulma (Stemming), sözcüğün çekimsel eklerden arındırılmış halinin ortaya ö-çıkırılması olup, morfolojik ve analiz anlamında önem taşımaktadır (Žižka, 2019). Dünya üzerinde konuşulan dilden dile değişmekle birlikte genelde kelimenin aldığı ekler, çoğu zaman anlamda değişiklik ortaya çıkarmamakta fakat “gözlük” örneğinde olduğu gibi bambaşka kavramları da ifade edebilmektedir.

```

from snowballstemmer import TurkishStemmer
turkStem=TurkishStemmer()
turkStem.stemWord("gözlüğe")

'gözlük'

turkStem.stemWord("gözlüğü")

'gözlük'

turkStem.stemWord("gözlüğünden")

'gözlük'

turkStem.stemWord("gözlükden")

'gözlük'

turkStem.stemWord("gözünden")

'göz'

```

Şekil 1.87. Türkçe kök bulma

Şekil 1.87’de Snowball Stemmer Kütüphanesi kullanılarak, “gözlüğe”, “gözlüğü”, “gözlüğünden”, “gözlükden” ve “gözünden” kelimelerinin kökleri bulunmuştur.

```

from nltk.stem import LancasterStemmer
lancaster = LancasterStemmer()
lancaster.stem('walking')

'walk'

lancaster.stem('walks')

'walk'

from nltk.stem import PorterStemmer
stemmer = PorterStemmer()
stemmer.stem('walks')

'walk'

stemmer.stem('walked')

'walk'

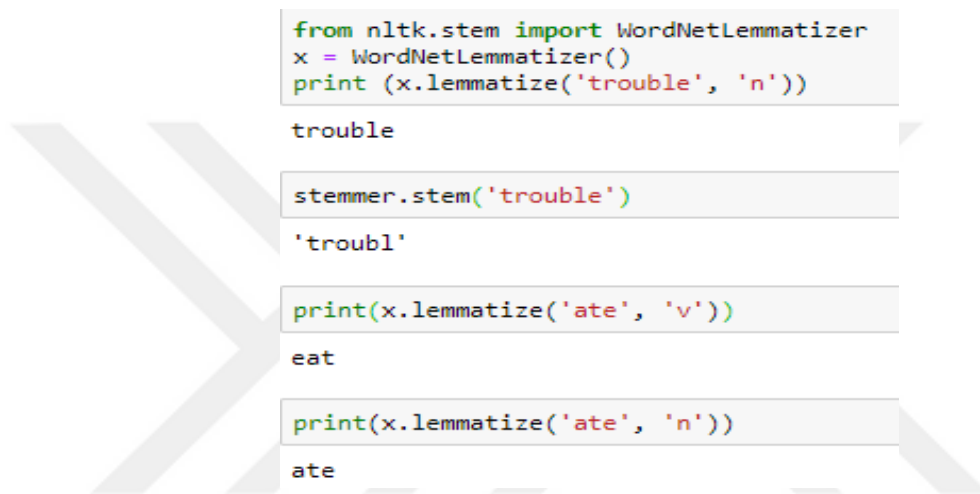
```

Şekil 1.88. İngilizce kök bulma

Şekil 1.88’de NLTK Kütüphanesi üzerinden Lancaster ve Porter algoritmaları kullanılarak; “walking”, “walks” ve “walked” kelimelerinin kökleri bulunmuştur.

1.15.13. Kök Çözümleme

Kök çözümleme (Lemmatization) işlemi de stemming işleminde olduğu gibi kelime köklerinin tespitine odaklanmaktadır. Fakat farklı algoritmik temellere dayanırlar, stemming işleminde kelimenin eksik ya da fazla kırpıldığı hata durumları daha fazla olmaktadır (Hardeniya ve ark., 2016). Kök çözümleme ise, kelimeleri metin içindeki görevlerine yani isim, fiil, zarf gibi niteliklere göre işlem yapmaktadır.



Şekil 1.89. Kök çözümleme ve kök bulma

Şekil 1.89’da görüldüğü üzere NLTK Kütüphanesi üzerinden “trouble” kelimesi stemming işlemine tabi tutulduğunda sonuç “troubl” çıkmaktayken, isim olarak kök çözümleme işlemi yapıldığında “trouble” bulunmuştur. Diğer örnekte ise “ate” kelimesinin fiil ve isim olarak kök çözümlemesi yapılmış olup, fiilde kelime kökü “eat” olmaktadır, isimde aynı kalmaktadır.

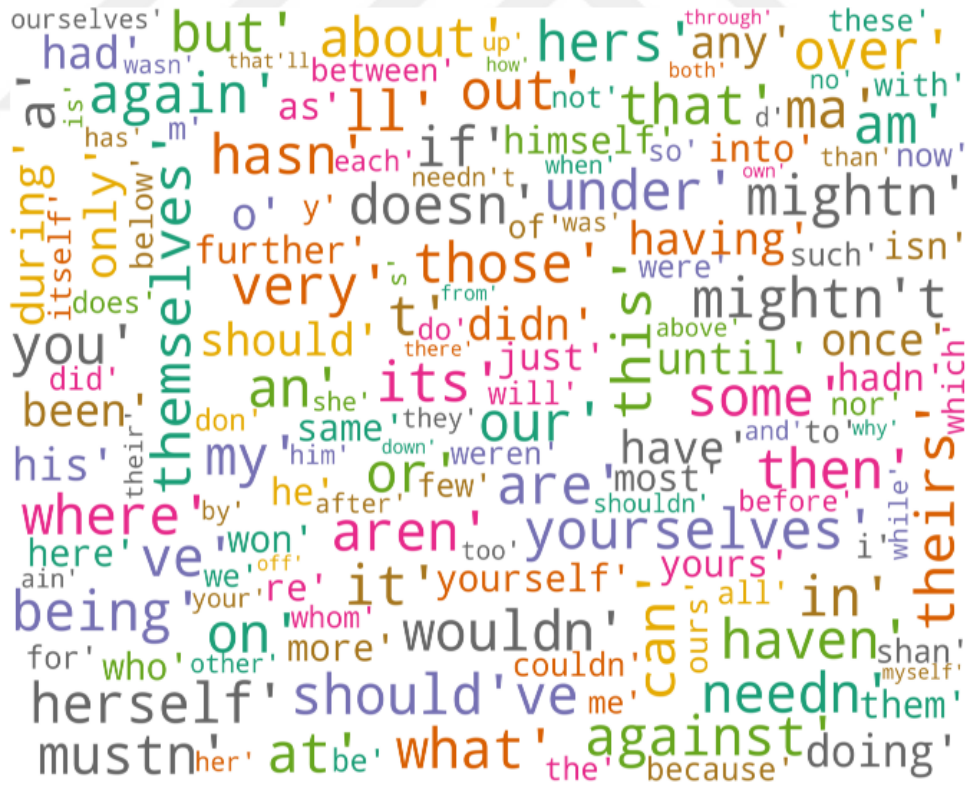
1.15.14. Durağan Kelimelerin Kaldırılması

Metin içinde önemi hiç veya az olan kelimelerin ana korpus içinden kaldırılarak, metnin diğer kelimelerini koruyup analize hazır hale getirmek için yapılan

önişlemlerden biri de durağan kelimelerin (Stopwords) uzaklaştırılmasıdır (Hardeniya ve ark., 2016).



Şekil 1.90. Türkçe durağan kelimeler



Şekil 1.91. İngilizce durağan kelimeler

Evrensel bir durağan kelime listesi bulunmamakta birlikte, NLTK üzerinde 22 farklı dile ait durağan kelime listesi bulunmaktadır. Söz konusu liste üzerinden Türkçe diline ait olan 53 tanesi şekil 1.90’da, İngilizce diline ait olan 179 tanesi de şekil 1.91’de kelime bulutu şeklinde gösterilmiştir.

```
from nltk.corpus import stopwords
stoplist = stopwords.words('turkish')
text = "bu bir yazılı sınav ama ben sonucu biliyorum"
cleanwordlist = [word for word in text.split() if word not in stoplist]
print(cleanwordlist)
```

['bir', 'yazılı', 'sınav', 'ben', 'sonucu', 'biliyorum']

```
from nltk.corpus import stopwords
stoplist = stopwords.words('english')
text = "this is a written exam but I know the result"
cleanwordlist = [word for word in text.split() if word not in stoplist]
print(cleanwordlist)
```

['written', 'exam', 'I', 'know', 'result']

Şekil 1.92. Durağan kelimelerin kaldırılması

Şekil 1.92’de görüldüğü üzere Türkçe "bu bir yazılı sınav ama ben sonucu biliyorum" ifadesinden ['bir', 'yazılı', 'sınav', 'ben', 'sonucu', 'biliyorum'] kaldığı, İngilizce "this is a written exam but I know the result" ifadesinden de ['written', 'exam', 'I', 'know', 'result'] kaldığı görülmektedir. Örnekte olduğu üzere durağan kelimeler arasında bir takım olumsuzluk kelime ve eklerinin de olması, bunların metinden uzaklaştırıldığında anlamda değişikliğin oluşabileceği unutulmamalıdır (Sarkar, 2019).

1.15.15. Sözcük Türü Etiketleme

Kelimeler bir dilde en küçük birimdir, bağımsızdır, kendi içinde bir anlamı ihtiva etmelerine rağmen cümle içinde bağımsız değildir. Zaman zaman anlam ve biçimde

farklılıklar olabilmekte, bu da analiz yapmayı güçlendirmesi sebebiyle sözcük türlerine göre etiketleme (POS – Part of Speech) yapılmaktadır. Metin içinde yer alan sözcüklerin söz dizimsel konumlarına ve cümle içindeki durumlarına göre genel olarak isim, fiil, sıfat ve zarf şeklinde metnin bölümlerinin etiketlenmesidir. Etiketleme işlemi iki aşamada gerçekleşmekte olup, birinci aşamada kelimelerin ait olduğu köklerin makine öğrenimiyle bulunması mevcut iken, ikinci aşamada söz konusu kelime kökünün türünün yine makine öğrenimi temelinde belirlenmesi süreci olup, yapılan bu işlem sayesinde anlamsal içerikler ortaya çıkarılmaktadır.

Kelimeler cümle içinde yer alan en küçük anlamsal öğelerdir. Bu sebeple de kelimelerin iyi anlaşılması; kelime → ifade → cümlecik → cümle zincirinde olduğu cümlenin yani metnin iyi anlaşılmasıdır. Kelimelerin anlaşılması sürecinde ise daha önce eğitilmiş kural tabanlı veriler ile mukayese edilerek, sözcük türünün etiketlendiği bir öğrenme bulunmaktadır.

Kelimelerin köklerinin, türlerinin ve kullanım sıklığının belirlenerek, ortak bir havuzda analizi ise BoW (Bag of Words) modeli olarak bilinmektedir. Sözcük türü etiketlemede asıl amaç, kelimelerden yola çıkarak, verinin genelini anlamlandırabilmektir.

```
sentence = "Ali okula gidecekti ama ben izin vermedim."  
import nltk  
nltk_pos_tagged = nltk.pos_tag(nltk.word_tokenize(sentence))  
pd.DataFrame(nltk_pos_tagged, columns=['Word', 'POS tag']).T
```

	0	1	2	3	4	5	6	7
Word	Ali	okula	gidecekti	ama	ben	izin	vermedim	.
POS tag	NNP	MD	VB	JJ	NN	NN	NN	.

Şekil 1.93. Türkçe POS etiketleme

```

sentence = """"Ali was going to school but I did not allow it""""
import pandas as pd
import spacy
nlp = spacy.load("en_core_web_sm", parse=True, tag=True, entity=True)
sentence_nlp = nlp(sentence)
# POS tagging with Spacy
spacy_pos_tagged = [(word, word.tag_, word.pos_) for word in sentence_nlp]
pd.DataFrame(spacy_pos_tagged, columns=['Word', 'POS tag', 'Tag type']).T

```

	0	1	2	3	4	5	6	7	8	9	10
Word	Ali	was	going	to	school	but	I	did	not	allow	it
POS tag	NNP	VBD	VBG	IN	NN	CC	PRP	VBD	RB	VB	PRP
Tag type	PROPN	AUX	VERB	ADP	NOUN	CCONJ	PRON	AUX	PART	VERB	PRON

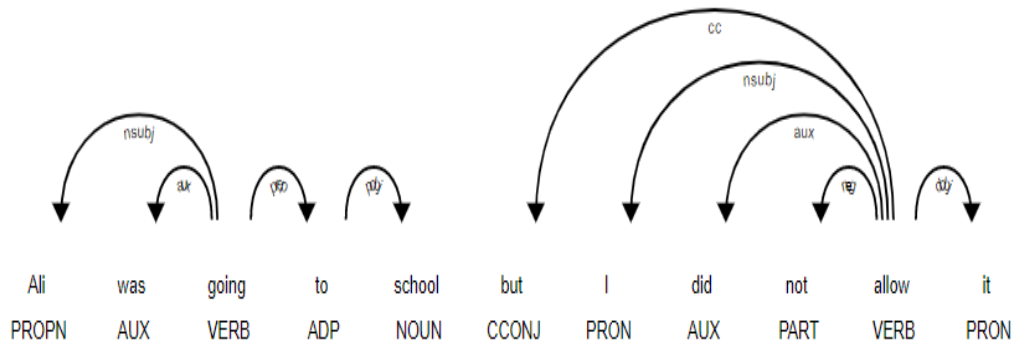
```

import nltk
nltk_pos_tagged = nltk.pos_tag(nltk.word_tokenize(sentence))
pd.DataFrame(nltk_pos_tagged, columns=['Word', 'POS tag']).T

```

	0	1	2	3	4	5	6	7	8	9	10
Word	Ali	was	going	to	school	but	I	did	not	allow	it
POS tag	NNP	VBD	VBG	TO	NN	CC	PRP	VBD	RB	VB	PRP

Şekil 1.94. İngilizce POS etiketleme

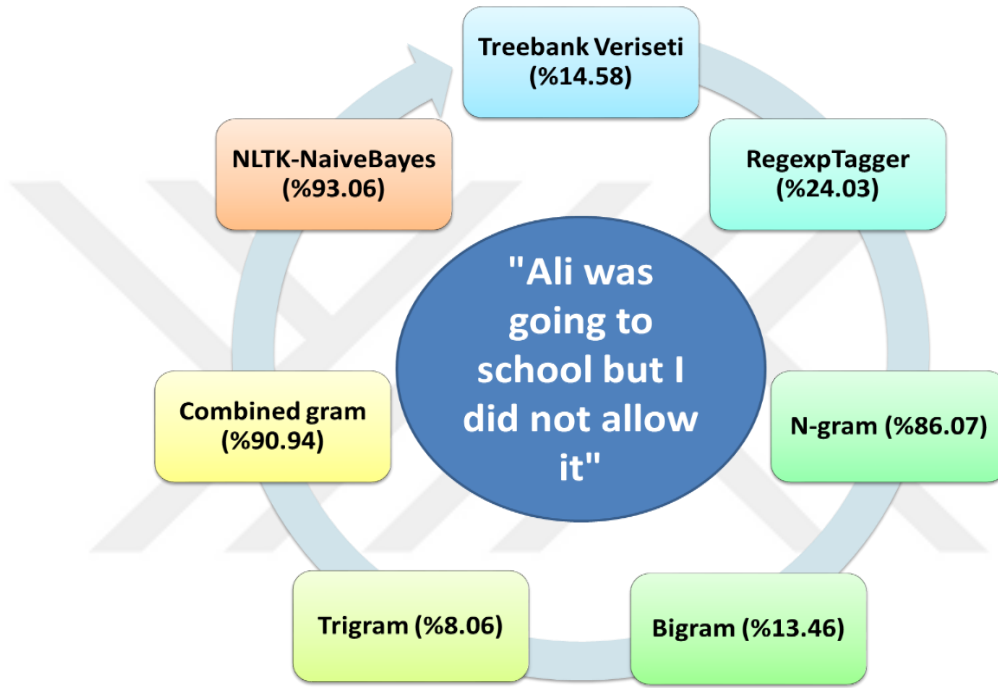


Şekil 1.95. Örnek cümlelerin ilişkisel analizi

Çizelge 1.4. Treebank POS etiketleri

Nu.	Etiket (Tag)	Açıklama (EN)	Açıklama (TR)
1	CC	Coordinating conjunction	Eşyapısal bağlaç
2	CD	Cardinal number	Numara
3	DT	Determiner	Belirteç
4	EX	Existential	Varolan
5	FW	Foreign word	Yabancı kelime
6	IN	Preposition or subordinating conjunction	Edat ya da yan bağlaç
7	JJ	Adjective	Sıfat
8	JJR	Adjective, comparative	Sıfat, karşılaştırmalar
9	JJS	Adjective, superlative	Sıfat, üstünlük
10	LS	List item marker	Liste ögesi işaretleyicisi
11	MD	Modal	Kip
12	NN	Noun, singular or mass	İsim, tekil veya kitle
13	NNS	Noun, plural	İsim, çoğul
14	NNP	Proper noun, singular	Özel isim, tekil
15	NNPS	Proper noun, plural	Özel isim, çoğul
16	PDT	Predeterminer	Ön belirteç
17	POS	Possessive ending	İyelik sonu
18	PRP	Personal pronoun	Kişisel zamir
19	PRP\$	Possessive pronoun	İyelik zamiri
20	RB	Adverb	Zarf
21	RBR	Adverb, comparative	Zarf, karşılaştırmalı
22	RBS	Adverb, superlative	Zarf, en üstün
23	RP	Particle	Ön ek veya son ek
24	SYM	Symbol	Sembol
25	TO	<i>to</i>	
26	UH	Interjection	Ünlem
27	VB	Verb, base form	Fiil, temel form
28	VBD	Verb, past tense	Fiil, geçmiş zaman
29	VBG	Verb, gerund or present participle	Fiil, zarf-fiil veya İngilizcede ing ile biten faaliyet ismi
30	VCN	Verb, past participle	Fiil, geçmiş zaman ortacı
31	VBP	Verb, non-3rd person singular present	Fiil, 3. şahıs olmayan tekil durum
32	VBZ	Verb, 3rd person singular present	Fiil, 3. şahıs tekil durum
33	WDT	Wh-determiner	Soru belirteçleri
34	WP	Wh-pronoun	Soru zamirleri
35	WP\$	Possessive wh-pronoun	İyelik soru zamirleri
36	WRB	Wh-adverb	Soru zarfı

Şekil 1.93’de "Ali okula gidecekti ama ben izin vermedim." cümlesi NLTK kullanılarak, şekil 1.94’de ise “Ali was going to school but I did not allow it” cümlesi Spacy ve NLTK kütüphaneleri kullanılarak öğelerine ayrılmış, öğelerine ayrılmıştır. Ayrılmış öge parçalarının isimlendirmeleri ise Penn Treebank formatında çizelge 1.4’de gösterilmekte (Santorini, 1990) olup, şekil 1.95’de ise karar ağacı deseninde ilişkisel durum görselleştirilmiştir.



Şekil 1.96. Farklı yöntemlerle POS etiketleme ve doğruluk oranları

Söz konusu POS etiketlemelerinin tamamı kütüphanelerin ana korpusunda yer alan etiketli verilerden oluşmaktadır. Metin işlemenin doğası gereği etiketleme işlemi çok da kolay olmamaktadır, örneğin “yaz” kelimesine baktığımızda; “bu soruyu yaz” cümleciğinde fiil iken, “bu yaz çok güzel” cümleciğinde ise isim rolünü üstlenmiştir. Şekil 1.96’da görüldüğü üzere “Ali was going to school but I did not allow it” örnek cümlesi üzerinden POS etiketlemede, NLTK Treebank üzerinden %14.58 doğruluk, RegexPTagger kütüphanesi üzerinden %24.03 doğruluk, n-gram ile %86.07, bigram ile %13.46, trigram ile %8.06, n-gram/bigram/trigramların birleşimi (combined gram) ile %90.94, NLTK üzerindeki verilerin Naive Bayes’deki eşit olasılık durumuna göre %93.06 doğruluk elde edilmiştir. Bu oranlar örnek cümle üzerinden yapılmıştır, farklı

cümlelerde farklı performanslar ortaya çıkabilecek olup, bu da metin madenciliğinin oldukça değişken parametrelerini içinde barındırdığını ortaya koymaktadır. Makine öğrenmesi temelinde farklı parametrelere göre modellenebilecek POS etiketleyicileri genelde Gizli Markov Modeli (Hidden Markov Model - HMM) ve Koşullu Rastgele Alan (Conditional Random Field - CRF) temelinde çalışmaktadır (Hardeniya ve ark., 2016)

1.15.16. İsimlendirilmiş Varlık Tanıma

İsimlendirilmiş varlık tanıma (Named Entity Recognition - NER), bir cümledeki kelimelerin varlık olarak tanımlanmasını ifade eder; daha farklı ifadeyle ilgili kişinin adı, yeri, organizasyonu gibi bilgilerin tanımlanmasıdır.

```
import spacy
from spacy import displacy
sp = spacy.load('xx_ent_wiki_sm')
sen = sp(u"""Ankara ve Istanbul sorusunu bilmiyorum ama Google,
          Opera ve Mozilla bildiğim en popüler arama motorlarıdır.""")
for entity in sen.ents:
    print(entity.text + ' - ' + entity.label_ + ' - ' + str(spacy.explain(entity.label_)))
```

Ankara - LOC - Non-GPE locations, mountain ranges, bodies of water
Istanbul - LOC - Non-GPE locations, mountain ranges, bodies of water
Google - ORG - Companies, agencies, institutions, etc.
Opera - MISC - Miscellaneous entities, e.g. events, nationalities, products or works of art
Mozilla - MISC - Miscellaneous entities, e.g. events, nationalities, products or works of art

```
displacy.render(sen, style='ent', jupyter=True)
```

Ankara LOC ve Istanbul LOC sorusunu bilmiyorum ama Google ORG ,
Opera MISC ve Mozilla MISC bildiğim en popüler arama motorlarıdır.

Şekil 1.97. İsimlendirilmiş varlık tanıma

Şekil 1.97’de görüldüğü üzere "Ankara ve Istanbul sorusunu bilmiyorum ama Google, Opera ve Mozilla bildiğim en popüler arama motorlarıdır." cümlesi üzerinde

Spacy kütüphanesi kullanılarak yapılan incelemede Ankara ve İstanbul'un yer olarak tanımlandığı, Google'ın şirket olarak, Opera ve Mozilla'nın ise bilinen çeşitli varlıklar olarak tanımlandığı görülmektedir.

1.15.17. Söz Dizimsel Analiz (Ayrıştırma)

Söz dizimsel (Parsing) analizde cümlelerin ayrıştığı, özne nesne, sıfat gibi temel öğelerden de faydalanarak söz diziminin nasıl olması gerektiğini belirleyerek metni anlamaya çalışan modellemedir. Örneğin “Ben severim seni” cümlesi insan tarafından anlamlandırması oldukça basit iken, makine tarafından anlaşılması bir dizi farklı işlemi beraberinde getirerek, bu cümlenin belirli bir olasılık tabanında “Ben seni severim” şeklinde algılanmasını içerir. Bu yüzden de dilin temel gramer kurallarına gereksinim duyularak, fiil olan kelimenin nerde duracağı, özneyi hangi yapının takip edeceğinin belirlenmesi önemlidir (Žižka ve ark., 2019). Söz konusu algılama yani öğrenme sürecinde temel olarak kural tabanlı ve olasılıksal olarak iki yaklaşım çeşidi mevcuttur (Hardeniya ve ark., 2016).

Kural tabanlı yaklaşımda, dilin gramer ve söz dizimsel sınırlamaları önemlidir, genel olarak içeriksiz gramer (Context-Free Grammar - CFG) ve regex temelli yaklaşımlar buna örnektir (Hardeniya ve ark., 2016). Olasılıksal yaklaşımda ise, dilin gramer ve söz dizimsel sınırlamaları belirli olasılık modellemelerinde öğrenilir, olasılıklı içeriksiz gramer (Probabilistic Context-Free Grammar - PCFG) ve Stanford kütüphanesinin söz dizimsel ayrıştırıcısı buna örnek olmaktadır (Hardeniya ve ark., 2016). Genel olarak ise, derin öğrenme mantığında çalışan olasılıksal yaklaşımların giderek daha fazla kullanıldığı söylenebilir.

1.16. Metin Temsil Yöntemleri

Metin ifadelerinin kendi içinde anlamsal farklılıklarının ve öznelliklerinin fazla olması sebebiyle zaman zaman yanlış ya da eksik iletim kökenli sorunlar

oluşabilmektedir. Bu sebeple söz konusu metni oluşturan kelimelerin öneminin ve anlamının hiçbir değişime uğramadan sayısal temsillerinin oluşturulması gerekliliktir. Sayısal yapıya dönüştürülmesi aynı zamanda makinelerin metni anlamlandırılmasını da sağlamaktadır. Sayısallaştırma işlemi vektörel düzlemde yapılmakta olup, bir vektör uzayı modeli yardımıyla, metin belgeleri vektör boyutlarını oluşturan belirli terimlerin sayısal vektörleri olarak matematiksel bir modele dönüştürülür. Bu kapsamda alt maddelerde metin temsili için kullanılabilecek bazı yöntemler belirtilmiştir.

1.16.1. Kelime Torbası (Bag of Words - BoW)

Kelime torbası modelinde, metin içinde yer alan tüm kelimeler bir küme içerisine toplanır ve küme içerisinde elde edilen veriler makine öğreniminde yaygın kullanılan vektörlere dönüştürülür. Yapılandırılmamış veriler için en basit vektör uzayı temsil formudur (Žižka ve ark., 2019).

	açık	bugün	güneşli	günler	hamburger	hava	kahverengi	köpekler	kızartmasını	olur	patates	sıcak	tatlıdır	yağmurlu	yemeyi	çok	özledim
0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0
1	0	0	1	1	0	0	0	0	0	1	0	1	0	0	0	0	0
2	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0
3	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1
4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1

Şekil 1.98. Kelime torbası gösterimi

Kelime Torbası modeli, her bir metin belgesini sayısal bir vektör olarak temsil eder; burada her boyut, bileşim kümesindeki belirli bir sözcüktür ve değeri, belgedeki sıklığı, oluşumu veya hatta ağırlıklı değerler olabilmekte olup, gösterimi 0-1 mantığına dayanmaktadır (Sarkar, 2019). Metin verilerinin tek bir kümede toplanması sırasında kelime sırası, sözcük dizimleri ve gramer konuları göz ardı edilir, sadece kelimelerin matematiksel gösteriminden oluşan bir çanta oluşturmak amaçtır. Şekil 1.98’de görüldüğü her cümlede mevcut ifadeler 1 olarak anılırken, o cümlede olmayıp diğer cümlede bulunan ifadeler 0 ile anılmaktadır. Bu yöntemde n kadar benzersiz kelime

kullanılıyorsa, n kadar vektörel düzlem kullanmak gerekir. Her kelimenin ayrı bir boyutta gösterimi verinin genel boyutunu arttırmakta ve aynı zamanda sözcükler arasındaki ilişkisel analizi sınırlandırmaktadır. Çok sayıda yer alan boyuttaki özelliklerin işlenmesi daha fazla zaman gerektirdiğinden, yüksek sayıda özellik genellikle hesaplama karmaşıklığı açısından da sorun olarak karşımıza çıkmaktadır (Žižka ve ark., 2019).

1.16.2. N-Gram

N-gram modeli, genel olarak kelime torbası yönteminin uzantısı konumunda olup, kelime torbasında dikkate alınmayan kelime sırasının hesaplamada dikkate alınmış halidir (Pagliardini ve ark., 2017; Sarkar, 2019). Örneğin; “IOS, Android’den daha hızlıdır.” cümlesi ile “Android, IOS’dan daha hızlıdır.” cümlesi arasında kelime çantası yöntemine göre herhangi bir fark yokken, aradaki anlam farkı ancak n-gram ile ortaya konulabilmektedir. Herhangi bir kelime unigram veya 1-gram olarak tanımlanır. Bigramlar iki kelimeden, trigramlar üç kelimeden ve n-gram ise n kelimeden oluşmaktadır.

	açık kahverengi	bugün hava	güneşli günler	günler sıcak	hamburger yemeyi	hava yağmurlu	kahverengi köpekler	köpekler çok	kızartmasını özledim	patates kızartmasını	sıcak olur	yemeyi özledim	çok tatlıdır
0	0	1	0	0	0	1	0	0	0	0	0	0	0
1	0	0	1	1	0	0	0	0	0	0	1	0	0
2	1	0	0	0	0	0	1	1	0	0	0	0	1
3	0	0	0	0	0	0	0	0	1	1	0	0	0
4	0	0	0	0	1	0	0	0	0	0	0	1	0

Şekil 1.99. Bigram gösterimine ilişkin örnek

N-gram yöntemiyle şekil 1.99’da görüldüğü üzere kelimelerin cümle içindeki sırası, önceliği veya sonralığı anlamlıdır. Kelime çantasında olduğu üzere, metin içinde mevcutsa 1, metin içinde benzersizse 0 olarak gösterim yapılır, bu gösterim şekli “one-hot” olarak da isimlendirilmektedir.

1.16.3. Terim Frekansı-Ters Doküman Frekansı (TF-IDF)

TF-IDF kavramı, İngilizcede “term frequency (TF)” olarak anılan terim frekansı terimi ile yine İngilizcede “inverse document frequency (IDF)” olarak anılan ters metin frekansı kavramlarının kısaltmalarından oluşmaktadır (Alelyani ve ark., 2018). Yöntem, bahse konu iki metriğin birleşiminden oluşmaktadır. BoW modelinin, büyük verilerde ortaya çıkarabileceği potansiyel hataların olması, bunun yanı sıra özellikle pek sık rastlamadığımız fakat çeşitli kategorileri ihtiva eden kelimeler özelinde de etkili olabilecek yeni bir model arayışı sonucunda TF-IDF modeli ortaya çıkmıştır (Sarkar, 2019). Aynı zamanda bu yöntem, terim içinde yer alan kelimelerin çıkarılarak, terim ağırlık değeri atamada en sık kullanılan yöntemlerden biridir (Wang ve ark., 2010).

$$tfidf = \text{Terim Frekansı (tf)} * \text{Ters Metin Frekansı (idf)}$$

Şekil 1.100. TF-IDF formülasyonu

Yöntemin genel matematiksel gösterimi (tfidf) yukarıdaki gibi olup (Şekil 1.100), terim frekansı (tf) herhangi bir terimin metin içinde ne kadar geçtiğinin hesaplanmasıdır. Matematiksel gösterimi aşağıda belirtilmekte olup, A belgesinde n kelimesinin sıklığının gösterimi f_{nA} ile yapılmıştır (Şekil 1.101).

$$tf(n, A) = f_{nA}$$

Şekil 1.101. TF formülasyonu

Ters terim frekansı (idf), herhangi bir terimin toplam belge sayısındaki her terim için belge sıklığına bölünmesi ve daha sonra sonuca logaritmik ölçeklendirme uygulanması ile hesaplanır. Paydada 0’a bölünmeden kaynaklı hatalardan dolayı +1 eklenir. A toplam belge sayısını, n de terimin bulunduğu belge sayısını ifade etmekte olup, formülasyonu aşağıda belirtilmiştir (Şekil 1.102).

$$idf(n, A) = 1 + \log \frac{N}{1 + df(n)}$$

Şekil 1.102. IDF formülasyonu

	Bugün	hava	yağmurlu.	Patates	Açık	günler	sıcak	olur.	kızartmasını	Hamburger	çok	yemeyi	köpekler	Güneşli	tatlıdır.	özledim.	kahverengi
0	0.58	0.58	0.58	0.00	0.00	0.0	0.0	0.0	0.00	0.00	0.00	0.00	0.00	0.0	0.00	0.0	0.00
1	0.00	0.00	0.00	0.00	0.00	0.5	0.5	0.5	0.00	0.00	0.00	0.00	0.00	0.5	0.00	0.0	0.00
2	0.00	0.00	0.00	0.00	0.45	0.0	0.0	0.0	0.00	0.00	0.45	0.00	0.45	0.0	0.45	0.0	0.45
3	0.00	0.00	0.00	0.61	0.00	0.0	0.0	0.0	0.61	0.00	0.00	0.00	0.00	0.0	0.00	0.5	0.00
4	0.00	0.00	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.61	0.00	0.61	0.00	0.0	0.00	0.5	0.00

Şekil 1.103. TF-IDF gösterimine ilişkin örnek

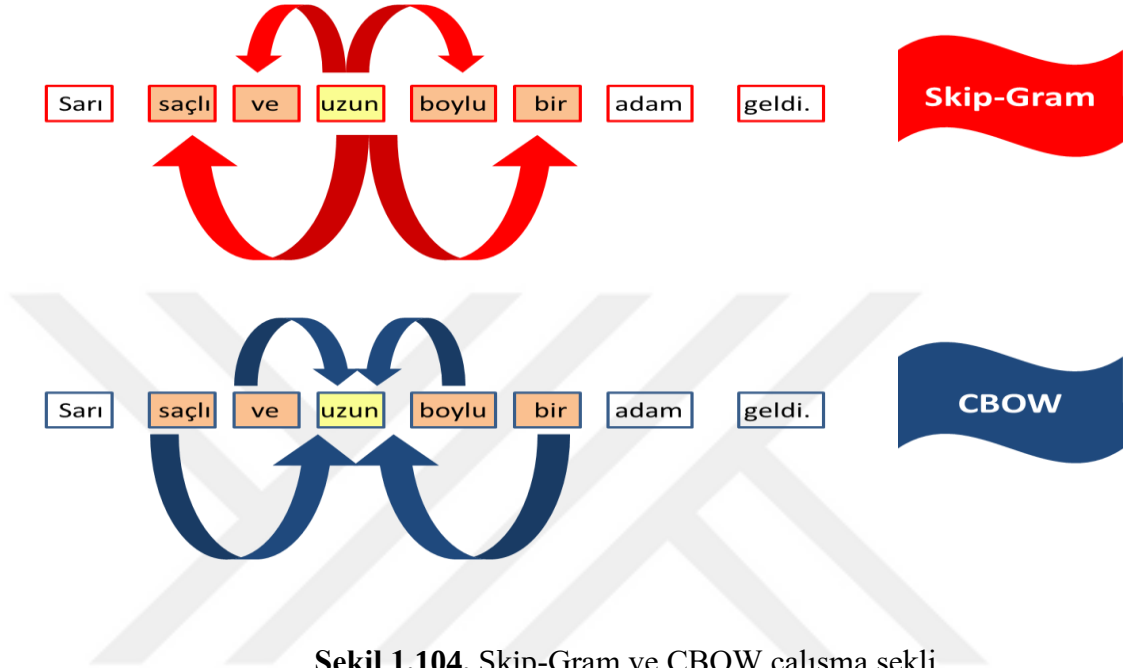
Şekil 1.103’de görüldüğü üzere incelemeye konu korpus içinde yer alan kelimeler metin içinde geçtikleri sıklığa göre yukarıda belirtilen formülasyon değerleri kullanılarak ağırlık değeri almaktadır. Ayrıca yapılan işlemi formülasyon gerek duymadan, Python programlama dilinde Scikit-Learn kütüphanesindeki fit_transform metoduyla TF-IDF hesaplanabilmekte ve bu da en yaygın kullanımlarından biri olarak karşımıza çıkmaktadır.

1.16.4. Kelime Vektörleri (Word2Vec)

Word2Vec denetimsiz ve tahmin temelli olarak kelime yapılarını vektörel düzlemde göstermek için kullanılan metin temsil yöntemi olup, ilk kez Google çalışanı Mikolov ve arkadaşları tarafından 2013 yılında ortaya atılmıştır (Bengfort ve ark., 2018). Metin temsili için bağlamsal ve anlamsal analizdeki güçlü tekniğiyle en yaygın kullanılan yöntemlerden biri olan Word2Vec, Skip-Gram ve CBOW (Continuous Bag of Words) olarak ikiye ayrılmaktadır (Ma ve Zhang, 2015). Söz konusu alt yöntemlerin ikisi de sinir ağlarını kullanmaktadır.

CBOW modelinde, kenardaki kelimelerden faydalanılarak ortada yer alan kelimeler tahmin edilmeye çalışılırken, Skip-Gram yönteminde ortadaki kelimelerden

kenardaki kelimeler tahmin edilmeye çalışılmaktadır (Şekil 1.104), yöntemler farklı olsa da her ikisinde de etrafındaki kelimelerden yola çıkarak yüksek oranda ilişkisel benzerliklerin tespitine odaklanılmıştır. Skip-Gram yöntemi küçük korpuslar üzerinde daha başarılı iken, çalışması CBOW'a göre daha yavaş olmaktadır (Sarkar, 2019).



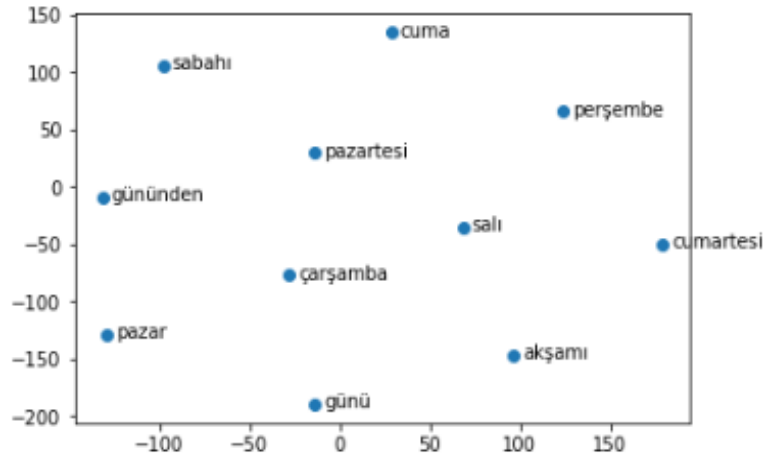
Şekil 1.104. Skip-Gram ve CBOW çalışma şekli

```
model.wv['pazartesi']
array([-4.32596743e-01, -5.73503137e-01, -2.51854151e-01, 1.74076632e-01,
       -1.15351766e-01, 6.79918826e-01, 3.02452445e-01, 2.06144169e-01,
       -5.16328871e-01, 2.89702024e-02, -3.98362279e-02, 4.23722059e-01,
       -8.12329934e-04, 3.24783146e-01, -3.24015081e-01, -7.20486394e-04,
       -2.35945791e-01, 1.41409770e-01, 7.04829693e-01, 1.68207943e-01,
       2.70622551e-01, -7.35547990e-02, -4.70357761e-02, 7.73448348e-01,
       1.64687619e-01, 1.01710176e-02, 7.15121627e-01, -1.60151228e-01,
       -1.40068099e-01, 1.57064907e-02, -8.02338243e-01, -3.92922908e-01,
       -3.39531898e-01, -3.48197341e-01, 1.93387493e-01, 1.20804757e-01,
       -3.60574037e-01, 1.67459592e-01, 5.03587306e-01, -4.01524603e-01,
       2.09635094e-01, 2.46924192e-01, 3.34720343e-01, -1.44494489e-01,
       -5.48628084e-02, -3.95119339e-01, -5.36114071e-03, 2.18418375e-01,
       3.17077428e-01, 1.37076080e-01, -3.79678845e-01, 3.16482633e-01,
       2.23112017e-01, -2.62077928e-01, 3.46372247e-01, 1.58605665e-01,
       1.44381687e-01, -2.83535689e-01, -4.37250137e-02, -1.29521668e-01,
       7.44397759e-01, -3.68199259e-01, 4.19067740e-01, -6.66296780e-02,
       -1.00970425e-01, 3.14206690e-01, -1.76413029e-01, 4.64249887e-02,
       4.63643581e-01, 7.61264041e-02, -8.76022100e-01, -2.10279912e-01,
       -4.39036071e-01, 1.09441578e+00, 8.65977257e-02, -1.11116163e-01,
       -2.16201693e-01, 6.99643135e-01, -8.77166390e-01, -8.80329013e-02,
       2.49705955e-01, 3.77545893e-01, -3.17547739e-01, 7.71237314e-02,
       -1.42481044e-01, 3.19165409e-01, 2.78861582e-01, -2.54730374e-01,
       1.13141919e-02, 4.13943492e-02, 1.79535851e-01, 8.48362923e-01,
       1.95746690e-01, -4.22854066e-01, 1.14064328e-02, 3.19808692e-01,
       -8.22148770e-02, -1.16329513e-01, -3.10655385e-01, 9.21263695e-02],
      dtype=float32)
```

Şekil 1.105. Kelimenin vektörel düzeyde gösterimi

```
model.wv.most_similar('pazartesi')
[('perşembe', 0.9375725984573364),
 ('cumartesi', 0.9241599440574646),
 ('çarşamba', 0.9111599326133728),
 ('salı', 0.8965775966644287),
 ('cuma', 0.8947243690490723),
 ('günü', 0.8835302591323853),
 ('akşamı', 0.8571591973304749),
 ('pazar', 0.827623724937439),
 ('sabahı', 0.823012769222595),
 ('gününden', 0.7515261173248291)]
```

Şekil 1.106. En benzer kelimeler



Şekil 1.107. En benzer kelimelerin grafik gösterimi

Word2Vec modelinin nasıl işlem gördüğünü örnek olarak gösterebilmek maksadıyla, Gensim kütüphanesi kullanılarak, Hürriyet gazetesinin verilerini içeren korpus üzerinden “pazartesi” kelimesinin vektörel karşılığı elde edilmiş (Şekil 1.99), söz konusu kelimeye en yakın ilişkide olan kelimeler, ilişki oranlarıyla bulunmuş ve grafikte gösterilmiştir (Şekil 1.105, Şekil 1.106 ve Şekil 1.107). Word2Vec modelinde genel olarak, korpus içinde yer alan tüm kelimelerin üzerinden geçilmiş, CBOW veya Skip-Gram algoritmalarıyla etrafındaki veya ortadaki kelime/kelimeler tahmin edilmeye çalışılarak, kelimelerin birliktelik durumu vektörel düzleme yansıtılmaktadır (Rong, 2014). Kelimeler üzerinde yapılan işlem Word2Vec olarak, belgeler üstünde yine denetimsiz öğrenme temelinde yapılması ise Doc2Vec olarak isimlendirilmektedir.

1.16.5. GloVe

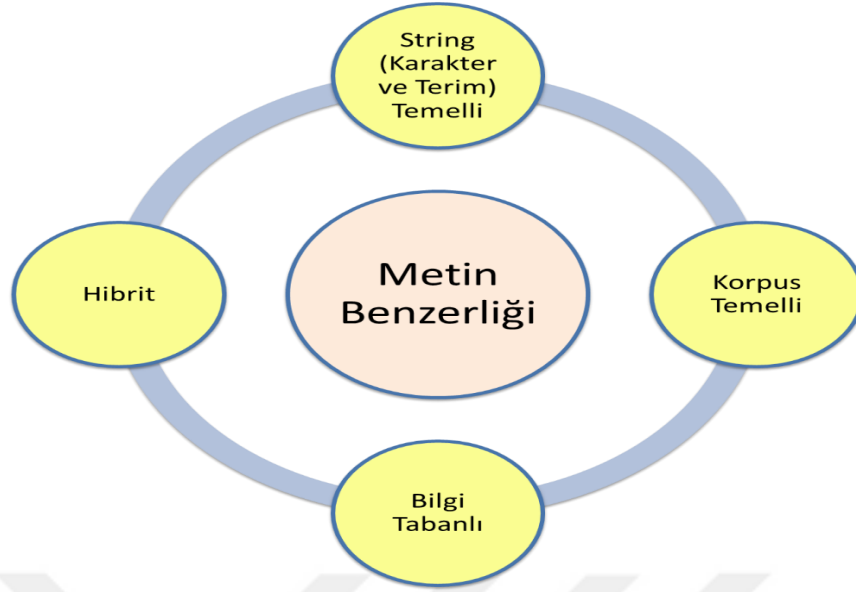
GloVe (Global Vectors for Word Representation) modeli, genel olarak Word2Vec'e benzemekle birlikte, özellikle büyük korpuslar üzerindeki vektörleri elde etmek için kullanılabilecek denetimsiz öğrenme modelidir (Sarkar, 2019). Vektör alanı oluşturması, Word2Vec ile aynı olmakla birlikte, GloVe modeli analiz öncesi kelimelerin istatistiksel analizini yaptıktan sonra vektörleri oluşturmayı başlar, eğitim süreci ise Word2Vec'e göre daha hızlıdır (Lane ve ark., 2019).

1.16.6. FastText

FastText modeli, 2016 yılında Facebook tarafından Word2Vec'in uzantısı olarak tanıtılarak geliştirilen modellemedir (Sarkar, 2019). Diğer modelleme türlerinde (n-gram hariç) morfolojik kelime yapıları yok sayılmakta, yapay sinir ağlarına girdi olarak kelime başlık başına ayrı bir varlık olarak girdi edilmektedir. FastText modellemesi, kelimeleri girdi olarak vermek yerine, n-gram seviyesindeki yapıları girdi olarak sisteme tanıtmaktadır, bu yüzden de her kelimeyi n-gram BoW gibi kabul eder (Lane ve ark., 2019). Sisteme girdilerin kelime yerine, n-gram olarak tanıtılması eğitim süresini uzatacak fakat bunun yanı sıra az kullanılan kelime yapılarının da sistem tarafından görülmesini sağlayacaktır.

1.17. Metin Benzerliği

Metin benzerliği, metin madenciliği içerisinde tasnifleme ve kümeleme temelinde çalışmalarda bir gereklilik olması sebebiyle önemi giderek artan bir çalışma konusudur.



Şekil 1.108. Metin benzerliği yöntemleri

Şekil 1.108’de gösterildiği üzere benzerlik; string, korpus, bilgi ve hibrit temelinde dört ana başlıkta sınıflandırılmaktadır (Gomaa ve Fahmy, 2013). Genel olarak benzerlik ölçümü sözcüksel (Lexical) ve anlamsal (Semantical) olarak yapılabilmektedir (Sarkar, 2019). Bu kapsamda metni oluşturan temel öge olan sözcükler arasında ve devamında cümlede, paragrafta ve belgenin tamamında aynı işlemin yapılarak ilişki durumunun ortaya konulması string temelli olup, bu yöntemde kendi içerisinde karakter ve terim temelli olarak ikiye ayrılmaktadır (Prasetya ve ark., 2018). Korpus temelli yaklaşım, daha önceki mevcut kavram kümesine dayanır ve anlam bilimsel bir yaklaşımdır (Prasetya ve ark., 2018). Bilgi tabanlı yaklaşım, anlamsal analizle elde edilen bilgileri kullanarak, kelimeler arası benzerliği ölçmeye çalışırken (WordNet gibi), hibrit tabanlı ise birden fazla benzerlik ölçme yönteminin birlikte kullanılmasıdır (Gomaa ve Fahmy, 2013).

Metin benzerliğini ortaya koymak için Hamming mesafesi, Kosinüs mesafesi, Öklid mesafesi, Manhattan mesafesi, Jaccard benzerliği, BM25 benzerliği, TF-IDF, Levenshtein mesafesi ve n-gram gibi yöntemler kullanılmakta olup (Gomaa ve Fahmy, 2013; Prasetya ve ark., 2018), en çok kullanılanı ise kosinüs ve TD-IDF yöntemleridir (Sarkar, 2019). TD-IDF doküman benzerliğinde terim frekansı baz alınarak özellikle

kullanılmakta olup, kosinüs yönteminde ise vektörel düzlemde iki veya daha fazla değerin aralarındaki açımın kosinüsünden faydalanılmaktadır. Benzerlik gösteren vektörel yapılar ($\cos 0^\circ$) 1 veya 1'e yakın, birbiriyle ilgisizse 0 veya 0'a yakın ($\cos 90^\circ$), birbiriyle zıt karakterde ise de -1 veya -1'e yakın bir ($\cos 180^\circ$) değer çıkacaktır (Sarkar, 2019).

1.18. Konu Modelleri

Veri miktarının her gün artmasıyla birlikte, ihtiyaç duyulan bilgiye erişimin giderek zorlaştığı bilinmektedir. Söz konusu zorluğun üstesinden gelebilmek için çeşitli metin madenciliği teknikleri kullanarak, büyük metinleri anlama adına konu modellemesi yapılması ihtiyaç haline gelmiştir (Blei, 2012). Konu modellemenin temel amacı, bir metin içindeki gizli ve gizli anlamsal yapıları keşfetmek için matematiksel ve istatistiksel teknikleri kullanmaktır (Blei ve Lafferty, 2009). Konu modellemesi bir makine öğrenmesi olup, konu modellemesi yapmak için kullanabileceğimiz çeşitli algoritmalar vardır. Konu modelleri, konu olarak temsil edilen bir belge grubundan anahtar temaları veya kavramları çıkarma süreci etrafında döner ve her konu, bir belge veya belge grubundan sözcük / terim dizinini olarak temsil edilebilmektedir (Sarkar, 2019). Bu terimler belirli bir konuyu, temayı veya kavramı ifade etmekte ve her bir konu da, bu terimlerin taşıdığı anlamsal yapı sayesinde diğer konulardan kolayca ayırt edilmektedir. Konu modelleri, temel kavramları çıkarmak ve tasvir etmek için büyük metin belgeleri grubunu özetlemeden daha iyi ve yararlı sonuçlar vermekte, metin içindeki gizli kalıpları çıkarmaya da olanak sağlamaktadır (Sarkar, 2019). Bu kapsamda en yaygın olanları, Gizli anlamsal analiz (Latent Semantic Analysis (LSA / LSI)), Olasılıksal gizli anlamsal analiz (Probabilistic Latent Semantic Analysis (pLSA)) ve Gizli dirichlet allocation (Latent Dirichlet Tahsisi (LDA)) olarak bilinmektedir (Sarkar, 2019).

1.18.1. Gizli Anlamsal Analiz (LSA)

Gizli anlamsal çözümlemenin (LSA) temel amacı, metinleri anlamsal analiz yapmak için vektör tabanlı gösterim kullanılarak, metinler arasındaki benzerliğin hesaplanmasıdır. LSA, verileri yeniden düzenlemek için tekil değer ayrışmasını (SVD) kullanır (Rehurek ve Sojka, 2010) (SVD, vektör uzayının tüm boyutlarını yeniden yapılandırmak ve hesaplamak için matris kullanan bir yöntemdir.). LSA yöntemi, Word2Vec ve GloVe'a göre daha hızlı eğitim sağlamakta ve büyük belgeler arasında daha iyi performans göstermekte olup, bunun yanı sıra korpusu daha az verimli kullanmakta ve özellikle soru-cevap gibi olaylarda kelimeleri daha doğru nedenleyebilmektedir (Lane ve ark., 2019).

LSA kelimeleri çeşitli konularda toplamak için TF-IDF vektör matrisini kullanarak, kelimeleri BoW vektörleri üzerinde bütünleştirmeye çalışan bir algoritmadır. Vektörel düzlemdeki konuları, boyutlarını ve çeşitliliği optimize etmek için temel ilk olarak benzer kelimelerin aynı kümelemede toplanması olduğundan daha fazla birlikte olma eğilimindedir (Sarkar, 2019). Söz konusu eğilim anlamsal yönden de gizli ilişkilerin ortaya çıkmasını sağlayacaktır. Ayrıca modellemede ortaya çıkacak konu sayısı TF-IDF vektör sayısından daha az çıkacağı için LSA aynı zamanda boyut sayısını azaltma yöntemidir (Lane ve ark., 2019; Rehurek ve Sojka, 2010).

1.18.2. Olasılıksal Gizli Anlamsal Analiz (PLSA)

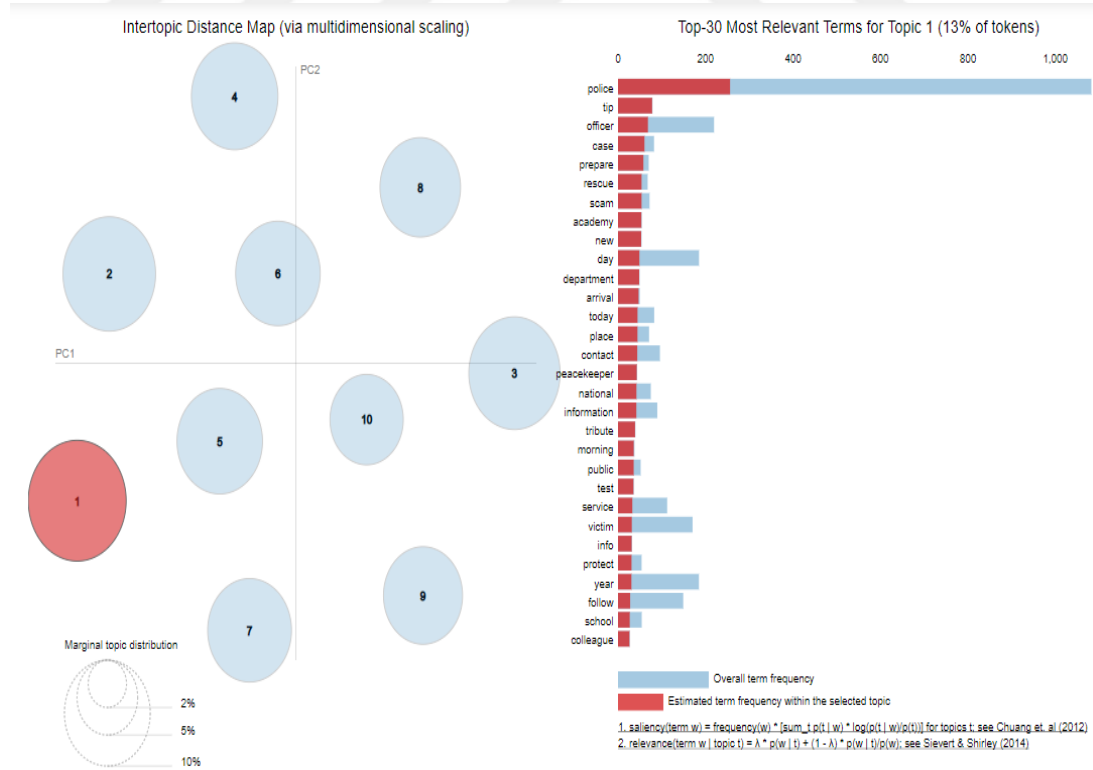
Olasılıksal gizli anlamsal analiz (PLSA), LSA modellemesini optimize etmek için ortaya atılan bir yaklaşımdır. Özellikle olasıksal anlamda modellemeyi geliştirilmeye çalışılmış olup, birden fazla anlamı olan kelimelerin netleştirilmesi ve ortak noktası olan kelimelerin gruplanarak, benzerliklerinin ortaya çıkarılması amaçtır (Sarkar, 2019). Kelimeler arası ilişkilendirmede, bir sözlüğe başvurmadan kelime kullanımının farklı bağlamlarını tanımlamak ve ayırt etmeyi ortaya koymak ana yaklaşımdır. Metin işlemede, otomatik soru önerisi tanımlamada oldukça başarılı

sonuçlar ortaya koymuştur. Eğitimi sırasında ise aşırı uyum (overfitting) problemiyle karşılaşlabilmektedir.

1.18.3. Gizli Dirichlet Tahsisi (LDA)

Gizli Dirichlet tahsisi (LDA), her belgenin olasılıksal anlam dizinini, benzer konuların kombinasyonunu ortaya koymak maksadıyla ortaya atılan olasılık tabanlı yaklaşımdır (Blei ve Lafferty, 2009). Genel olarak amacı metin içerisinde yer alan kelimelerin çoğunluğunu kapsayan bir konu modeli ortaya koymaktır.

Şekil 1.109’da Scikit-Learn Kütüphanesi kullanılarak, örnek yapılan 10 adet konuya sahip modellemeye baktığımızda, 1 numarada vektörel düzlemde gösterilen konu modellemesinin tokenlerin yaklaşık %13’ünü ihtiva ettiği görünmektedir. Bu da belgelerin çeşitli konu karışımlarından üretildiğini ortaya koymaktadır.



Şekil 1.109. LDA konu modellemesi ve token yapısı

LSA ile temelde aynı varsayımlara dayanmakta olup, temel fark LDA'nın konu dağılımının Dirichlet olduğunu varsaymasıdır (Dağılımsal hipotez ve istatistiksel hipotezle belgelerin birkaç konu hakkında olabileceği dağılım hipotezi) (Rehurek ve Sojka, 2010). LSA'da ise konu modellemede herhangi bir dağılım öngörülmemekte olup, konular vektörel temsillere göre yapılmaktadır. LDA'da belirlenen konulara kelimelerin tahsis edilmesi istatistiksel sonucu, LSA'nın doğrusal matematiğinden daha kesindir (Lane ve ark., 2019).

LDA kullandığı yaklaşım farklılığıyla en basit ve hızlı boyut küçültme ve sınıflandırma modellerinden biridir (Lane ve ark., 2019). Özellikle doğal dil işlemedeki makine öğrenimi modelleri arasında en popüler metin modelleme tekniğidir (Sarkar, 2019).

1.19. Tezin Amacı ve Kapsamı

Bilgisayar teknolojilerinin gelişimi ve internet kullanımının son yıllarda yaygınlaşması neticesinde, mevcut sosyal ağlar kitle iletişim araçlarına alternatif bir iletişim ortamı oluşturmuştur. Mevcut sosyal ağların kullanımı arttıkça da “Foursquare” gibi yer ve mekân paylaşımı, “Instagram, Facebook” gibi resim ve ifade paylaşımı ve bunun yanı sıra “Twitter” gibi hayata ilişkin düşünce ve müşterek örgütlenmelerin yapılabileceği özel siteler çoğalarak, insanların iletişim anlayışında bazı değişimlere yol açmıştır.

Dijital dünyanın giderek gerçek hayatın yerini alması ve hemen hemen her bireyin sanal bir kimliğinin olduğu günümüzde, ortaya çıkan tablonun suç disiplini açısından incelenmesi araştırmanın genel kapsamını oluşturmaktadır. Dijitalleşmeye bağlı olarak, suç oranlarının da giderek arttığı günümüzde, suç ve suçla mücadele edebilmek adına, yapılan bu çalışmada Twitter üzerinden yapılan paylaşımların veri madenciliği ile analiz edilerek, makine öğrenimi yardımıyla elde edilen bilgilerin, güvenlik güçleri tarafından suçu önlemeye yönelik yapılacak proaktif faaliyetlerde kullanılabilirlik durumunun belirlenmesi, içinde suç ihtiva eden paylaşımın önceden

belirlenebilmesi ve ayrıca ulusal/uluslararası alanda faaliyet gösteren güvenlik güçlerinin paylaşımlarının toplum destekli güvenlik hizmeti açısından analizinin yapılması amaçlanmaktadır. Yapılan bu çalışma, sosyal ağlarda suç kavramını ele alacak olmasından dolayı, özellikle adli bilişim, kriminoloji ve kriminalistik literatürü başta olmak üzere, Türkçe metin madenciliği alanına kuramsal ve ampirik olarak katkı yapması beklenmektedir.

Araştırma kapsamında aşağıda belirtilen hipotezlerin ve araştırmanın alt amaçları olarak çizelge 1.5’de belirtilen araştırma sorularının cevaplanması amaçlanmaktadır.

Çizelge 1.5. Hipotezler ve araştırma soruları

Çalışmadaki Hipotezler ve Hipotezlere Bağlı Araştırma Soruları	
Hipotez – 1	
Güvenlik güçlerinin suçla mücadelede toplum desteğini sağlamaya yönelik sosyal ağlarda yaptığı paylaşımların, aynı ülke içinde veya uluslararası boyutta farklılaşmaması beklenmektedir.	
Hipotez – 1’e Bağlı Araştırma Soruları	
1	Kolluk kuvvetlerinin sosyal ağ kullanım alışkanlıkları nasıldır? (Hangi sıklıkla, hangi cihazlar üzerinden, ne tür mesajlar verdiği vb.)
2	Kolluk tarafından paylaşılan mesajların makine öğrenimine dayalı duygu durumu (pozitif, negatif veya nötr) nasıldır?
3	Kolluk tarafından yapılan paylaşımların konu modellemesi nasıldır, ağırlıklı işlediği konular neler? (LDA Analizi)
4	Ülkemizde faaliyet gösteren Jandarma ve Polis tarafından yapılan paylaşımlara halkın verdiği tepki üzerinden mukayeseli korelasyon analizi ile anomali tespiti yapmak mümkün müdür?
5	İtalya’da faaliyet gösteren Arma dei Carabinieri ve Polizia di Stato tarafından yapılan paylaşımlara halkın verdiği tepki üzerinden mukayeseli korelasyon analizi ile anomali tespiti yapmak mümkün müdür?
6	Fransa’da faaliyet gösteren Gendarmerie Nationale ve Police Nationale tarafından yapılan paylaşımlara halkın verdiği tepki üzerinden mukayeseli korelasyon analizi ile anomali tespiti yapmak mümkün müdür?
7	İspanya’da faaliyet gösteren Guardia Civil ve Policía Nacional tarafından yapılan paylaşımlara halkın verdiği tepki üzerinden mukayeseli korelasyon analizi ile anomali tespiti yapmak mümkün müdür?

Çizelge 1.5. Devam

8	Ülkemizde, İtalya’da, Fransa’da ve İspanya’da faaliyet gösteren jandarma ve polis tarzı kolluk kuvvetlerinin sosyal ağdaki etki gücü nedir?
9	Ülkemizde, İtalya’da, Fransa’da ve Fransa’da faaliyet gösteren jandarma ve polis tarzı kolluk kuvvetlerinin sosyal ağdaki paylaşımları arasında benzerlik var mıdır?
Hipotez - 2	
Makine öğrenimi modellemeleriyle sosyal ağlardaki Türkçe metinler üzerinde siber zorbalık tespiti yapmak mümkündür.	
Hipotez – 2’ye Bağlı Araştırma Soruları	
1	Siber zorbalık içeren ve içermeyen mesajlar arasında keşifsel analizde farklılık var mı?
2	Siber zorbalığın tespiti denetimli öğrenme kapsamında yapılabilir mi?
3	Siber zorbalığın tespiti denetimsiz öğrenme kapsamında yapılabilir mi?
4	Farklı makine öğrenimi modelleri arasında F1 skor oranları nasıldır?
5	Suçla mücadelede makine öğrenim modellerinin kullanımı işgücü ve kolluksal etkinlik gücünü artırır mı?
6	Veri setinin büyük veri üzerindeki sonuçları nasıldır? Yeterli mi, geliştirilmeli mi?
7	Türkçe metinlerde tespitin optimize edilmesi kapsamında dil gramer yapısı etkili mi?
Hipotez - 3	
Veri madenciliği teknikleriyle sosyal ağlar üzerinde keşifsel veri analizi yapılarak, açık kaynak istihbaratı elde edilebilir.	
Hipotez – 3’e Bağlı Araştırma Soruları	
1	Herhangi bir konuya ilişkin atılan mesajların duygu durumu tespit edilebilir mi?
2	Olumsuz mesajların (terör, hakaret, zorbalık vb. içeren) UTF sisteminde kullandığımız saat ile ilişkisi var mıdır?
3	Herhangi bir konuda en etkin # ve @ etiketlerini görmek mümkün müdür?
4	@ ile # işaretli etiketler arası veri ağı analizi yapmak mümkün mü?
5	Denetimsiz öğrenme kapsamında konu modellemesi yapılabilir mi?
6	Anahtar kelime analizi araması yapılarak, büyük veriden anahtar kelimeyle bağlantılı ilişkili veriler öğrenilebilir mi?
7	@ etiketli etki gücü en yüksek hesaplar üzerine keşfedici analiz yapılabilir mi?
8	Haber ajanslarından elde edilen bilgi, insanların sosyal medyada yazdıklarının ne kadarını karşılıyor?

Çizelge 1.5’de belirtilen hipotez ve alt araştırma sorularının cevaplandırılması ile sosyal ağların karar destek birimlerine bütüncül yaklaşımla, iletişimden

müdahaleye fayda sağlayacağı değerlendirilmektedir. Tartışma bölümünde belirtilen mevcut literatür taramasına istinaden, bu çalışma sosyal ağlardaki suçların incelenerek, uygulamada yapay zekâ kullanılarak tespit edilmesi ve sosyal ağdaki genel örüntülerin tespiti ve elde edilen bilgilerin suç ve suçla müdahalede bütüncül olarak ortaya kullanılması açısından sosyal ağları ele alan ilk çalışma olacaktır. Bu kapsamda suç özelinde kriminolojik ve kriminalistik öğelerin tezde incelenmesi, tezin özgün değerini oluşturmaktadır.



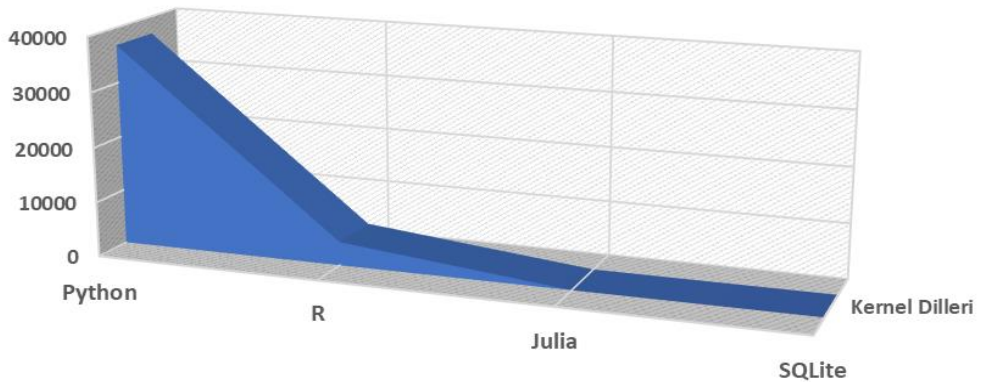
2. GEREÇ VE YÖNTEM

“Önleyici kolluk faaliyetleri kapsamında sosyal medyada veri madenciliği: suçları önlemek için Twitter Analytic’i kullanma” konu başlığı altında yapılan çalışmaya ilişkin kullanılan gereç ve yöntemler aşağıda maddelendirildiği üzere, alt başlıklarda açıklanmaktadır.

- Verilere erişmede kullanılan programlama dili
- Veriler üzerinden yapılan işlemlerde kullanılan kütüphaneler
- Veri elde elde ve ön işleme
- Verinin sayısallaştırılması
- Verinin analizi

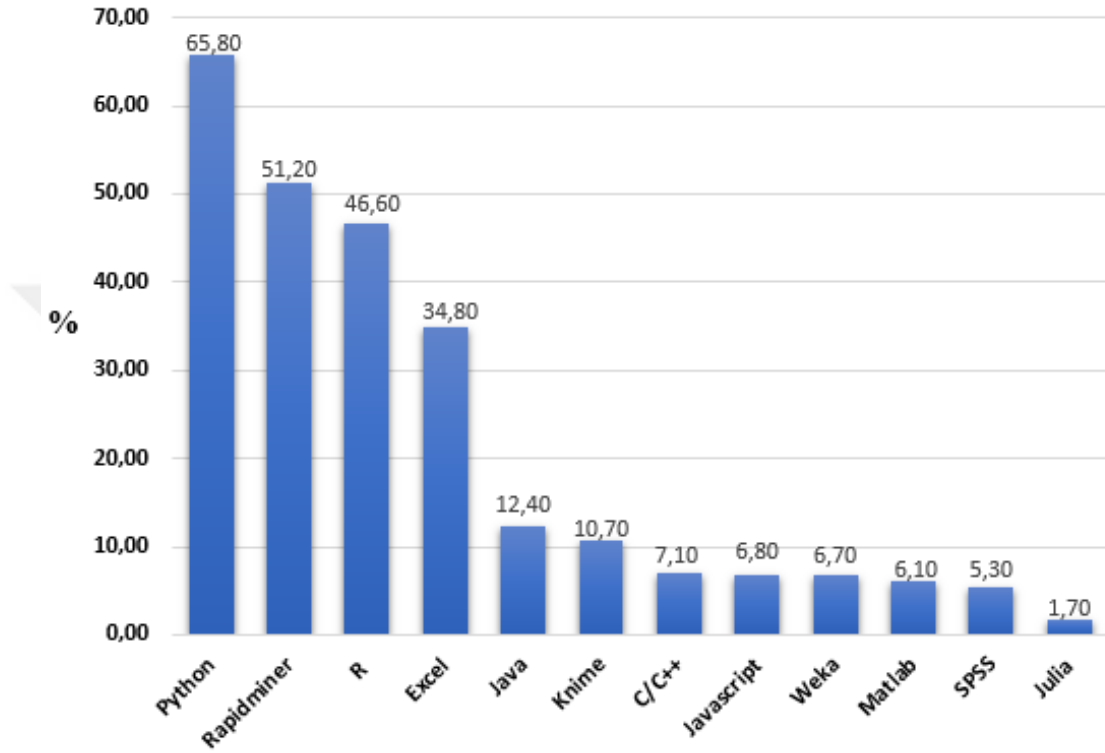
2.1. Verilere Erişmede Kullanılan Programlama Dili

Büyük verilerden inceleme yaparak, metin örgüleri üzerinde makine öğrenimine dayalı sonuçlar çıkarılması ve elde edilen bilgilerin görselleştirilmesi çalışmanın ana gereksinimini oluşturmaktadır. Bu kapsamda kullanılacak programlama dilinin seçimine ilişkin yapılan araştırmada;



Şekil 2.1. Kaggle üzerinden makine öğreniminde kullanılan diller

– Google’ın yürütücülüğünde faaliyet gösteren veri dünyasının büyük platformlarından olan Kaggle (2020) isimli sitede makine öğrenimi temelinde kernel olan işlemlerin %89.6’sının Python üzerinden yapıldığı (Python-36824, R-4267, Julia-5, SQLite-2) şekil 2.1’de gösterildiği üzere belirlenmiştir.



Şekil 2.2. 2019 yılında veri bilimi ile ilgili kullanılan dil ve yazılımlar

– Veri madenciliği konusunda en büyük platformlardan olan KDnuggets üzerinden Piatetsky (2020) tarafından yapılan çalışmada; 2017, 2018 ve 2019 yıllarında kullanıcılarına “What Analytics, Big Data, Data Science, Machine Learning software you used in the past 12 months for a real project?” (Gerçek bir proje için son 12 ayda kullandığınız Analitik, Büyük Veri, Veri Bilimi ve Makine Öğrenimi yazılımı nedir?) sorusunu yöneltmiş olup, elde edilen cevaplarda 2019 yılı için de %65.8 oranında Python’un kullanıldığı belirlenmiş ve söz konusu çalışmanın bazı çıktıları yüzde cinsinden şekil 2.2’de gösterilmektedir.

Genel kullanım oranlarına bakıldığında sık kullanımı ile dikkat çeken Python, çoklu programlama paradigmalarını destekleyen genel amaçlı bir programlama dili olup, içinde sınıflar, nesneler, veriler ve yöntemler dahil olmak üzere birçok nesne yönelimli programlama kavramı da bulunmaktadır. Veri tabanı yönetimi, makine öğrenmesi, derin öğrenme, doğal dil işleme ve web geliştirme gibi amaçlarla aktif kullanılmaktadır (Sarkar, 2019). Bu sebeple bu çalışmada Python dili kullanılmıştır.

2.2. Veriler Üzerinde Yapılan İşlemlerde Kullanılan Kütüphaneler

Python ile verilerin Twitter üzerinden çekilmesi, işlenmesi ve modellenmesinde kullanılan kütüphaneler ile söz konusu kütüphanelerin en bilinen işlevlerinin tanımlanmasına ilişkin açıklamalar aşağıda belirtilmektedir.

- Gensim kütüphanesi, büyük veri içerisinde konu modelleme, doküman indeksleme ve benzerlik analizi için kullanılmaktadır.
- JSON kütüphanesi, JSON (JavaScript Object Notation) formatında her bir uygulamanın sunucularla iletişim kurması için geçici kod gereksinimini ortadan kaldırmak üzere tasarlanmış bir kodlama düzenidir.
- Matplotlib kütüphanesi, sistemin MATLAB üzerinden çalışmasını sağlayan komut işlevi olup, her pyplot işlevi bir şekil üzerinde bazı değişiklikler yaparak, çizim alanı içerisinde şekil oluşturur.
- NLTK (Natural Language Toolkit) kütüphanesi, insan dili verileriyle çalışacak Python programları oluşturmak için kullanılmakta olup, sözcük kaynağı için sınıflandırma, belirteç, kaynak, etiketleme, ayrıştırma ve anlamsal akıl yürütme gibi işlevleri yerine getirebilmektedir.
- NumPy (Numerical Python) kütüphanesi, bilimsel hesaplamaların hızlı bir şekilde yapılmasını sağlayan matematik temelli işlevleri içerisinde barındıran modüldür.
- Pandas kütüphanesi, incelemenin ve analizin veri çerçevesi (dataframe) içinde yapılmasını sağlayan yüksek performanslı ve açık kaynak kodlu modüldür.

- Scikit-Learn Kütüphanesi, Sklearn olarak da bilinmekte olup, makine öğrenimi yapılabilmesi için içinde çeşitli öğrenme algoritmalarını barındıran modüldür.
- Seaborn kütüphanesi, matplotlib tabanlı bir Python görselleştirme modülü olup, istatistiksel grafikler oluşturmak için üst düzey bir kullanım sağlar.
- Spacy kütüphanesi, doğal dil işleme için geliştirilmiş olup, etiketleme, ayrıştırma, tokenleştirme ve adlandırılmış varlık tanıma gibi özellikleri içerisinde barındıran modüldür.
- Tensorflow kütüphanesi, makine öğrenimi temelinde kullanılan ve Google tarafından 2017 yılında piyasaya sürülen modüldür.
- Textblob kütüphanesi, metin verilerini işlemede ve çeviri gibi metin görevlerinin icrasında kullanılan modüldür.
- Tweepy kütüphanesi, twitter geliştirici hesabından alınan giriş kodları ile verilerin çekimi için Twitter’a erişimde kullanılan modüldür.
- WordCloud kütüphanesi, özellikle metin temelli verilerden kullanım sıklığı vb. kriterler de dikkate alınarak kelime bulutu yapılmasında kullanılan modüldür.

2.3. Verileri Elde Etme ve Ön İşleme

Araştırma kapsamında sosyal ağlar temelinde yapılacak veri incelemesi için Twitter isimli sosyal ağ hesabı örneklem olarak seçilmiş olup; söz konusu ağın örneklem olarak seçilmesinin nedenleri olarak;

- Twitter özelinde kullanıcıların duygu ve düşüncelerini ifade etmede Facebook ve Instagram’a kıyasla gizliliğin biraz daha düşük olması sebebiyle, Twitter’ın özgür düşünce ve paylaşımlara diğer ağlara göre daha uygun bir platformda olması,
- Facebook ve Instagram gibi popüler diğer platformların kişisel blog tarzı kullanımı biraz daha yaygın iken, Twitter kullanıcılarının etiketler yardımıyla gündem yaratıp, örgütlenme imkânı bulması yer almaktadır.

2.3.1. Twitter

Twitter, kayıtlı kullanıcıların sosyal ağ içerisinde paylaşım yapabildiği, başkalarının paylaşımlarının beğenebildiği ve söz konusu başkalarına ait paylaşımların tekrar paylaşılabilirdiği (retweet) dünya üzerinde oldukça yaygın kullanımı olan bir sosyal ağ platformudur. Twitter analytic üzerinden verilere ilişkin temel performans metriklerini yayınlamaktadır. WeaereSocial (2020) tarafından yapılan incelemede, ülkemizde 11.800.000’lik aktif kullanıcı kitlesi olan Twitter, Dünya genelinde; Amerika, Japonya, İngiltere, Suudi Arabistan ve Brezilya’dan sonra en fazla kullanımın olduğu 6’ncı ülkedir. Ülkemizdeki kullanım oranları, Facebook ve Instagram’ın gerisinde kalsa da, kullanım amaçları diğerlerinden oldukça farklı olup, örgütlenme ve fikir madenciliği diğer ağlara nazaran daha kolay yapılabilir.

Gönderilen her mesaj “tweet” olarak tanımlanırken, varsayılan olarak paylaşım yapılan tweet’ler herkes tarafından görülebilmekte olup, ancak gönderi dağıtımında takipçilerle olacak şekilde de kısıtlamaya gidilebilmektedir. Twitter’a web sitesi, akıllı telefonlarda yer alan uygulamalar ve bazı ülkelerde uygulandığı üzere kısa mesajla da paylaşım yapılabilir. 2017 yılına 140 karaktere kadar paylaşımlara izin veren Twitter, sonrasında 280 karaktere kadar insanların paylaşım yapabildiği platform halini almıştır.



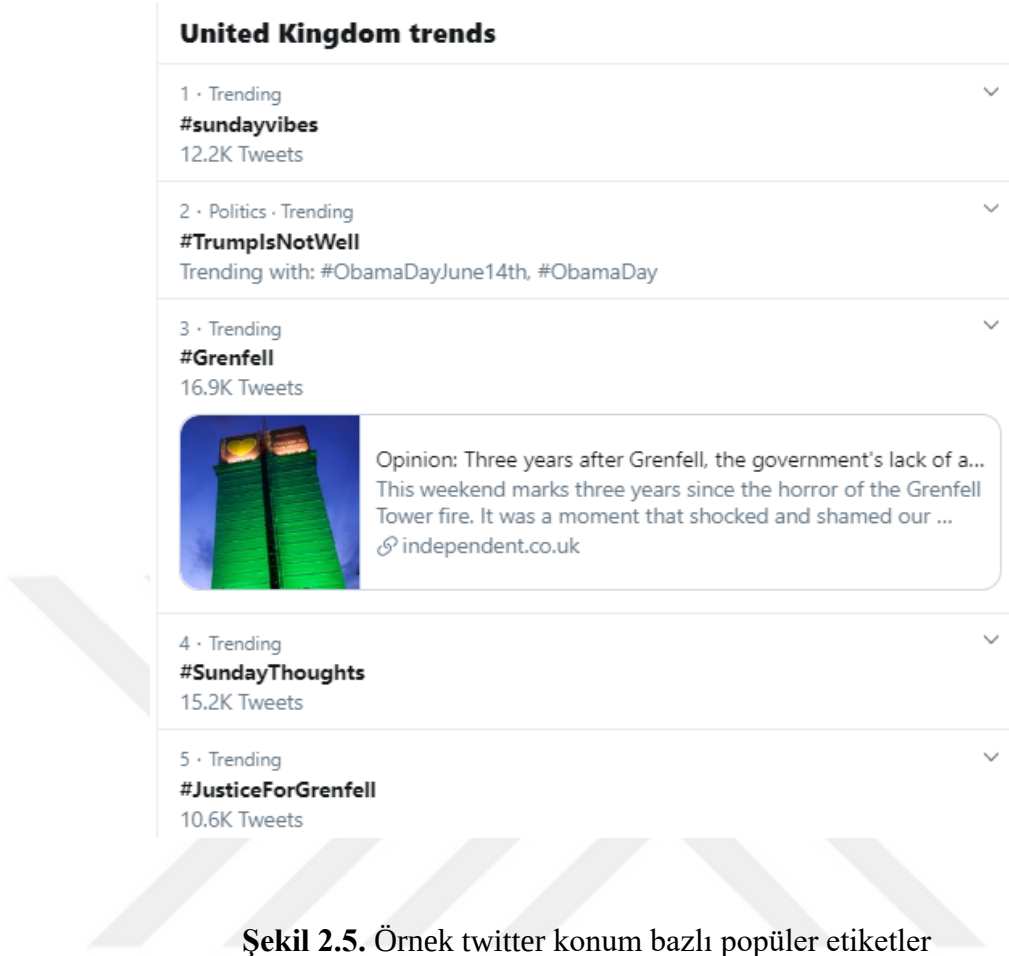
Şekil 2.3. Örnek kişisel twitter sayfası girişi

Şekil 2.3’de örnek olarak gösterildiği üzere, takip edilmek istenen kullanıcı “follow” (takip et) sekmesi tıklanarak, takip edilebilmekte olup, yaptığı paylaşımlar hakkında twitter takipte bulunan kullanıcılara bildirimde bulunarak bilgi akışı sağlamaktadır.



Şekil 2.4. Örnek twitter paylaşımlarının genel görünümü

Şekil 2.4’de örnek olarak gösterildiği üzere, twitter paylaşımları ters zamansal dizilim bulunmakta olup, işletmede sıklıkla bahsedilen prensiplerden biri olan LIFO (Last In First Out) çalışma şekli olarak belirlenmiştir, bu sebeple de en son yapılan paylaşımlar en üst sırada, farklı bir ifadeyle en görünür yerde bulunmaktadır. Paylaşım altına, diğer kullanıcıların yorum yazabilmesi, paylaşımı beğenmesi ve paylaşımı yeniden paylaşması gönderi altındaki kutucuklardan yapılmaktadır.

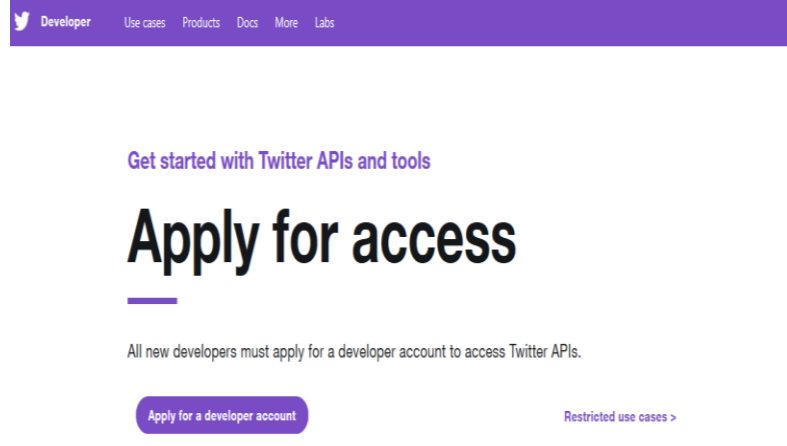


Şekil 2.5. Örnek twitter konum bazlı popüler etiketler

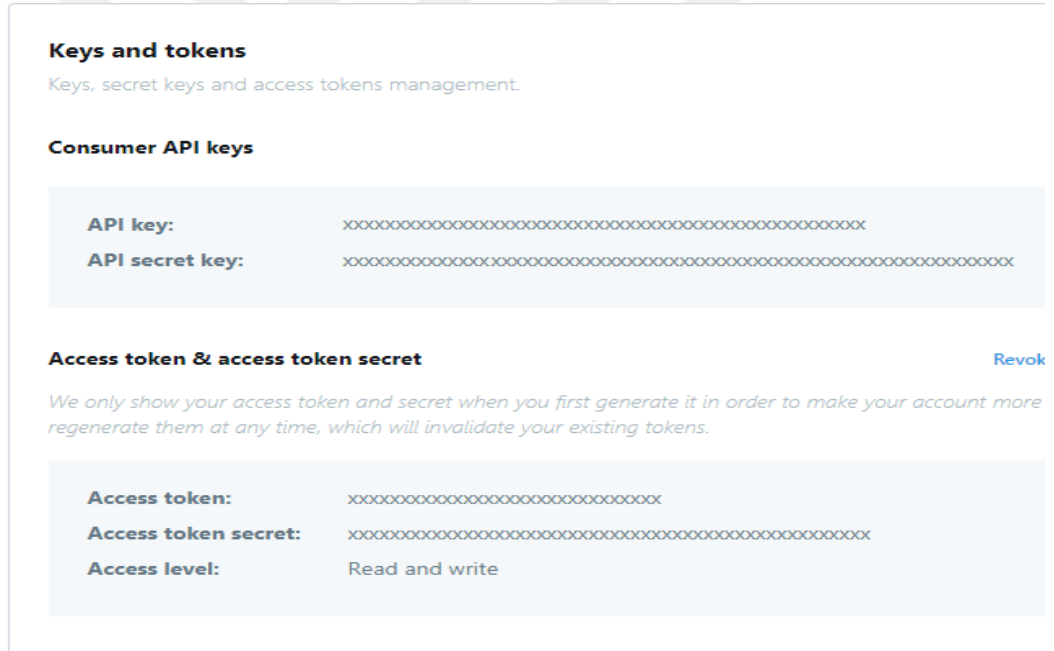
Şekil 2.5’de örnek olarak gösterildiği üzere, twitter üzerinden konum bazlı olarak popüler konuşulan konular belirlenebilmekte olup, söz konusu konular üzerinden sosyal ağ içinde sık konuşulan konuları takip edebilmek mümkün olmaktadır.

2.3.2. Veri Elde Etme

Araştırmaya konu sosyal ağdan elde edilen veriler, Python programlama dili kullanılarak, Twitter Developer API (Twitter Geliştirici Kütüphanesi) üzerinden yapılmıştır.

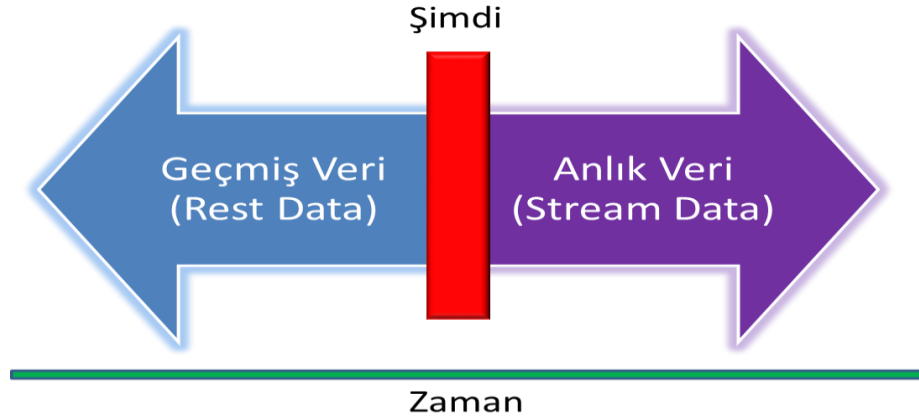


Şekil 2.6. Twitter Developer API sayfası genel görünümü



Şekil 2.7. Twitter API token ve şifre edinimi

Şekil 2.6’da ekran resmi görüldüğü üzere <https://developer.twitter.com> adresinden geliştirici hesabı açılarak, şekil 2.7’de ekran resmi görülen (token şifrelerinin üstü xxx şeklinde kapatılmıştır.) geliştirici hesap şifreleri alınmıştır.



Şekil 2.8. Twitter’den çekilebilecek veriler

Şekil 2.8’de görüldüğü üzere twitter üzerinden geçmiş ile anlık veriler çekilebilmektedir (Rest API – Stream API). Geçmiş veriler, herhangi bir anahtar kelime vasıtasıyla konu üzerinde konuşulan konular olabileceği gibi, belirli bir kullanıcıya ait veriler de olabilmektedir (Bonzanini, 2016). Anlık verilerde ise, mevcut zaman ve sonrası ana temel olup, anahtar kelime vasıtasıyla elde edilen güncel verilerdir.

Twitter Developer API’nin standart kullanımına ilişkin genel bilgiler aşağıda belirtilmektedir (Twitter, 2020; Bonzanini, 2016);

- Geçmiş veriler üzerinden aramada yedi günlük zaman kısıtlaması mevcuttur, verilerin hepsi verilmez, örneklem tweetler verilir.
- Varsayılan erişim seviyesi 400 parça anahtar kelimeye, 5.000 kullanıcı kimliğine ve 25 adet 0.1-360 derece konum kutusuna izin verir.
- Genel kural olarak Twitter verilerini ayrıştırmada JSON ayrıştırıcı kullanımı daha iyi sonuç vermektedir.
- Verilen sonuçlar tüm genel durumların rastgele örneğini döndürür, varsayılan erişim düzeyi tarafından döndürülen tweetler aynıdır, bu nedenle iki farklı istemci bu uç noktaya bağlanırsa aynı tweetleri almaktadır.

– Anahtar kelime üzerinden yapılan çalışmalarda, anahtar kelime tweetin gövdesi içindeki metinle eşleşirse, veri akışa bırakılır. Bu bir tokenleştirilmiş eşlemedir, anahtar kelime dizesinin tweet gövdesinin tokenleştirilmiş metniyle eşleştiği anlamına gelmektedir.

– Emojilerin anahtar kelime gibi kullanıldığı durumlarda işleyiş anahtar kelimelerle aynıdır, unicode üzerinden işlemler yapılır.

– “@” simgesi kullanıcı etiketlemede kullanılmaktadır.

– “#” simgesi hashtag etiketlemede kullanılmaktadır.

– Veri çekiminde dil sınırlaması girilebilmekte fakat elde edilen verilen kullanıcının seçtiği dil üzerinden elde edilmekte olup, anahtar kelimeyle birlikte dil kısıtının girilmesi daha anlamlı veriler elde etmeyi sağlamaktadır.

– Veri çekiminde konum sınırlaması konulabilmekte veya konum bazlı olarak da çalışabilmekte olup, konum bilgilerinin twitter tarafından çok az seviyede verilmesi sebebiyle seçim, doğal afet vb. gibi özel durumlar dışında anlamlı analiz yapılması biraz daha zordur.

– Anahtar kelime araması, tırnak (“”) içerisinde işlem görmektedir.

– Geçmiş veri çekiminde yapılan paylaşım verilerinin diğer kullanıcılar tarafından ne tepki aldığı (yorum, beğenilme veya tekrar paylaşma) anlık veriye nazaran daha anlamlı bir şekilde incelenebilir.

– Tweetlerin hangi kaynaktan (cep telefonu, web vb.) gönderildiğine ilişkin veriler alınabilmekte fakat bu kıstas veri süzgeci olarak kullanılamamaktadır.

– Paylaşılan video ve resimlere ilişkin veriler URL şeklinde alınabilmektedir (Sadece twitter’a yüklenenler için geçerli olup, Periscope gibi diğer uygulamalardan yapılan işlemlerde geçersizdir.).

– Twitter API, uygulamalara erişimi sınırlar, bu sınırlar kullanıcı başına veya daha belirleyici olarak erişim belirteci başına belirlenmekte olup, uygulama yalnızca kimlik doğrulaması yaparak çalışmakta ve sınır aşıldığında veri çekimine kısıt koyulmaktadır.

– Hız sınırları 15 dakikalık aralıklara ayrılmıştır. Tüm uç noktalar kimlik doğrulaması gerektirir, bu nedenle kimliği doğrulanmamış aramalar ve hız sınırları kavramı yoktur, GET istekleri için iki başlangıç grubu vardır: 15 dakikada bir 15 çağrı ve 15 dakikada bir 180 çağrı şeklinde işlem yapılmaktadır.

- Her 15 dakikada uygulamaya ilişkin istek limitlerinin toplamı günlük (24 saatlik zaman diliminde) 100.000 istekten fazla olamamaktadır.
- Zaman sınırlarına uyarak kullanıcılara ait, son 3200 civarı tweet çekmek mümkündür, söz konusu sayıya retweet sayıları dahildir.
- Her veri çekim isteğinde çekilecek tweet sayısının (max_count) en fazla 200 olması gerekmektedir.
- Veri çekilirken istek kısıtlamalarına uyulmuş olup, herhangi bir hata (İnternet hızı, sistem kısıtlaması vb.) alınması durumunda anlık veriler, saat ve tarih bazlı kayda alınarak, yeni döngüde tekrar veri çekimine devam edilmiştir.

Şekil 2.9. Örneklem Twitter mesajında yer alan bilgiler

2.3.3. Veri Ön İşleme

Verilerin işlem safhasında gereksiz yoğunluk yaratmaması ve verilerden daha anlamlı sonuçlar elde edilmesi maksadıyla şekil 2.9’da yer alan başlıklardan kullanılanları ayrı bir csv formatında toplanarak, bunlar üzerinde de bazı ön işleme çalışmaları yapılmıştır.

Veri ön işleme kapsamında yapılan çalışmalar;

- Veri ön işlemede aşamasında öncelikle büyük veri (Korpus) içinde yer alarak analizde kullanılmayan kısımlar temizlenmiş ve veri çekiminde kısıtlamalardan kaynaklı alınan hatalarla tarih/saat bazlı dosyalar birleştirilmiştir.

- Veri ön işleme safhasında en sık karşılaşılan problem ise ana yazılım uygulamalarının ASCII (İngilizce: American Standard Code for Information Interchange, Türkçe: Bilgi Değişimi İçin Amerikan Standart Kodlama Sistemi) olmasıdır. ASCII kodlama sistemi latin alfabesi üzerine kurulu 7 bitlik bir karakter kümesi olduğundan, Türkçede yer alan ç, ş, ü, ö gibi harfleri tanımlamamaktadır. Türkçe verinin çekilmesi ve okunması sırasında hep “UTF-8” (Unicode Transformation Format – 8 Bitlik) dilinin çözümleme olarak kullanılmıştır.

- Farklı harf karakterlerini metinden temizlemek için, “re.sub” metodu kullanılarak (Örneğin Türkçe için “text = re.sub("[^a-zçğışöüA-ZÇĞİŞÖÜ]", ' ')", farklı karakterler İngilizce ve Türkçe’ye özel olarak temizlenmiş ve farklı karakter olması durumunda da ilgili yere boşluk bırakılması sağlanmıştır.

- Örneğin İngilizce’de “I” harfinin küçük yazıldığında “i” olması gibi nedenlerle, İngilizce için ASCII değerler tablosunda harfler küçültülmüş (text.lower), diğer dillerde de UTF-8 değerler tablosunda harfler küçültülerek analizde olabilecek karmaşanın önüne geçilmiştir.

- Yapılan paylaşımlar bölme (split) metoduyla kelimelere ayrılarak, içerisinde İngilizce ve Türkçe duraklama kelimeleri (Stopwords) olması durumunda korpuslardan temizlenmesine ilişkin döngü hazırlanarak, metindeki durağan kelimeler uzaklaştırılmıştır (Türkçe için {'da', 'her', 'nereye', 'asında', 'siz', 'ne', 'hep', 'az', 've', 'belki', 'mü', 'birkaç', 'acaba', 'o', 'de', 'hem', 'yani', 'şey', 'biz', 'birşey', 'biri', 'bu', 'ama',

'gibi', 'nerede', 'daha', 'neden', 'tüm', 'hepsi', 'veya', 'bazı', 'niçin', 'ya', 'şu', 'hiç', 'ile', 'niye', 'defa', 'çünkü', 'için', 'kez', 'eğer', 'en', 'ise', 'çok', 'mu', 'kim', 'nerde', 'sanki', 'diye', 'ki', 'nasıl', 'mı'}, İngilizce için {'doing', 'hasn't', 'during', 'hadn't', 'ain', 'won', 'isn't', 'no', 've', 'before', 'you're', 's', 'she's', 'up', 'that'll', 'being', 'as', 'only', 'shouldn't', 'won't', 'or', 'until', 'herself', 'if', 'between', 'has', 'off', 'then', 'while', 'yourself', 'yours', 'y', 'all', 'couldn't', 'doesn't', 'those', 'haven't', 'own', 'very', 'that', 'you', 'for', 'to', 'but', 'mustn't', 'been', 'most', 'so', 'than', 'can', 'shouldn't', 'from', 'itself', 'just', 'shan', 'into', 'with', 'ma', 'once', 'you'd', 'other', 'and', 'should've', 'm', 'did', 'should', 'needn't', 'against', 'it', 'ourselves', 'aren't', 'i', 'hers', 'didn't', 'o', 'hasn't', 'how', 'mightn't', 'its', 'the', 'yourselves', 'she', 'are', 'same', 'here', 'further', 'under', 'wouldn't', 't', 'by', 'this', 'you'll', 'be', 'do', 'it's', 'him', 'not', 'few', 'himself', 'about', 'down', 'nor', 'we', 'such', 'needn't', 'some', 'didn't', 'shan't', 'weren't', 'is', 'was', 'after', 'an', 'a', 'over', 'now', 'what', 'wouldn't', 'they', 'them', 'because', 'below', 'both', 'above', 'whom', 'aren't', 'isn't', 'wasn't', 'there', 'will', 'their', 'at', 'll', 'you've', 'mightn't', 'had', 'your', 'each', 'mustn't', 'am', 'couldn't', 'our', 'of', 'on', 'he', 'themselves', 'too', 'in', 'more', 'why', 'have', 'don't', 'hadn't', 'which', 'theirs', 'ours', 'having', 'out', 'when', 'don't', 'where', 'her', 'his', 'me', 'again', 'd', 're', 'through', 'haven't', 'weren't', 'myself', 'does', 'who', 'were', 'my', 'any', 'these', 'doesn't', 'wasn't'}).

– Özel karakterleri karşılık gelen Türkçe karakterle, harflerin değişimi sağlanmıştır (Türkçe’de örneğin; özel harfler = {'î': 'i', 'â': 'a'}).

– İngilizce analiz için stemming ve lemmatization işlemleri yapılmış olup, Türkçe için Zemberek, Snowballstemmer, Turkishlemmatization gibi kütüphaneler mevcuttur, fakat çıktıları arasında zaman zaman başarısız örnekler oluşabilmektedir. (Türkçe’de örneğin yapılmaya çalışılan; “okul”, “okula”, “okulda” ve “okuldan” sözcüklerinin hepsini okul olarak tanımlayabilmesidir. Bu şekilde yapılmasa da analiz yapılabilmekte olup, kök tanımlama yapıldığı taktirde daha başarılı olacağı değerlendirilmiştir.)

– Gereksiz analizin önüne geçebilmek adına ve Türkçe’de yeterince iyi kök bulma kütüphanelerinin eksikliğinden kaynaklı kelimenin metin içinde geçiş sayısı dikkate alınmış olup, özellikle büyük korpuslar temelinde iyi sonuçlar alınmış olup, kullanımı da oldukça yaygındır (Örneğin “kalem” kelimesi, metin içinde iki veya daha fazla metin içinde geçmesi durumunda analize alınmış, böylece mümkün olduğunca

tekil yanlış yazımlar dışlanarak, makine öğreniminde sisteminde daha düşük eforla öğrenim sağlanmaya çalışılmıştır.). Metin verilerinin sosyal medyada oldukça bozuk kullanımından dolayı, bu şekilde analiz özellikle Türkçe büyük verilerde gerekliliktir.

– Türkçe için kök bulma işleminin uygulanmamasından kaynaklı, yanlış yazılabilecek bazı özel tanımlamalar yapılmıştır (Örneğin `df[column] = df[column].str.replace("pıslık","pislik"), df[column] = df[column].str.replace("gerizekali","gerizekalı")` gibi)

2.4. Verilerin Sayısallaştırılması

Ön işlemde geçirilen veri, terimlere ayrılarak her bir veri genel bir terimler sözlüğü elde edilmiştir. Elde edilen sözlük üzerinden kelimeler torbası ve TF-IDF türlerinde tasnifleme yapılarak, metin verisinin sayısal ifadeye dönmesi sağlanmıştır.

2.5. Verilerin Analizi

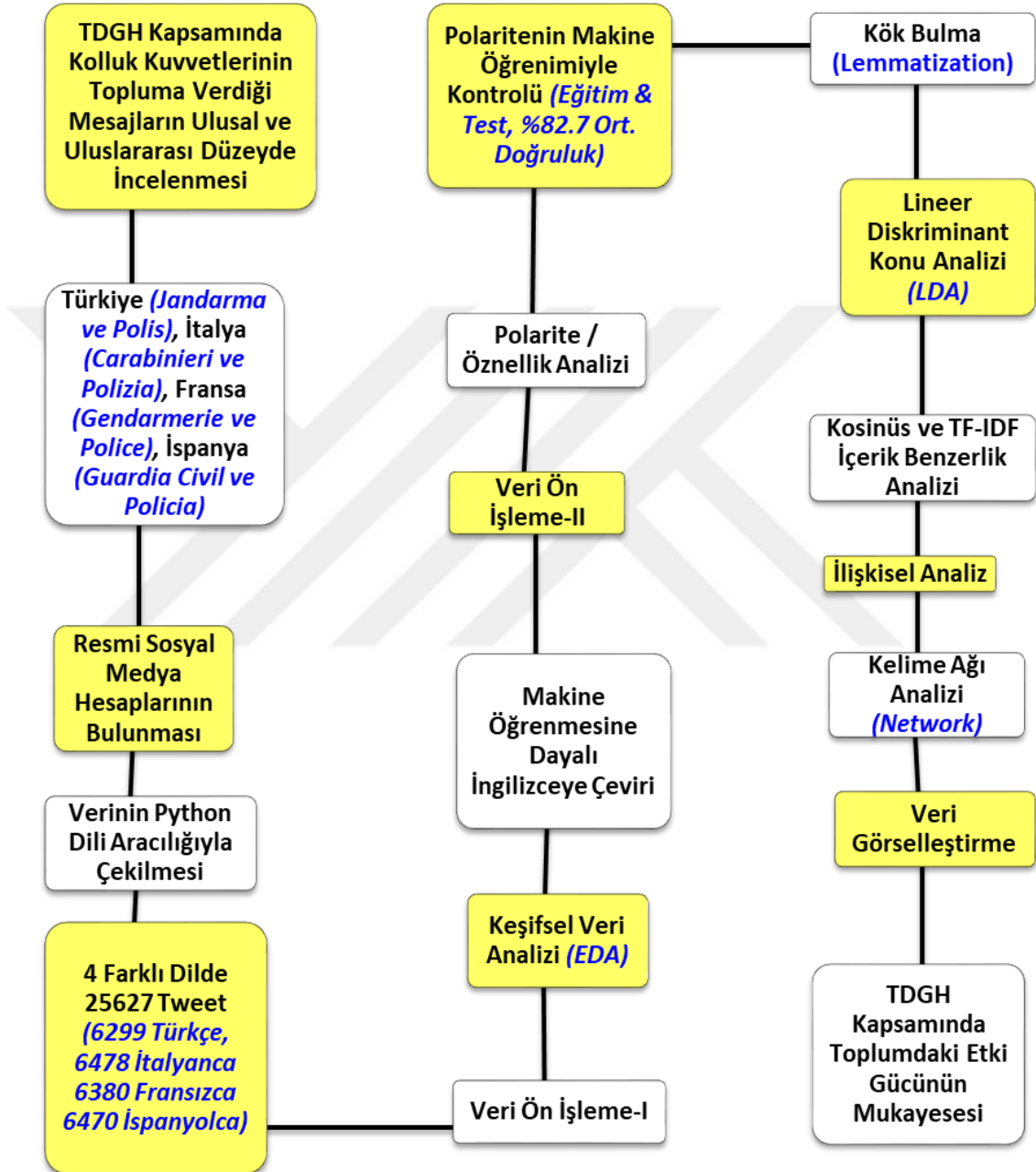
Tez kapsamında üç farklı çalışma konusu ele alınmış olup, her konu farklı analiz metotları altında incelenmiştir. Çalışmada genel olarak; metin madenciliği temelinde keşifsel veri analizi, metin görselleştirme, makine çevirisi, LDA konu modelleme, makine ve derin öğrenme algoritmalarına dayalı sınıflandırma, duygu polarite analizi, cosinüs ve TF-IDF benzerlik analizleri, kullanıcı/kelime ağ analizi, zamansal/anlamsal anomali, denetimsiz öğrenme kapsamında t-SNE modellemesi üzerinden vektörel düzeyde kümeleme gibi analizler yapılmıştır. Analizler kapsamında yapılan tüm işlemler, Windows bilgisayar üzerinden Python 3.8 kullanılarak yapılmıştır. İncelenen konular ile konulara ait işlem prosesleri ve analiz kapsamında kullanılan kütüphane ve metotlar ilgili konu başlığında açıklanmıştır.

2.5.1. TDGH Kapsamında Kolluk Kuvvetlerinin Paylaşımlarının İncelenmesi

Araştırma kapsamında elde edilen veriler; Türkiye, İtalya, Fransa ve İspanya’da faaliyet gösteren jandarma ve polis teşkilatlarının resmi Twitter hesapları üzerinden dört farklı dilde (Türkçe, İtalyanca, Fransızca ve İspanyolca) elde edilen bilgilerden oluşmaktadır. Bu çalışma “Güvenlik güçlerinin suçla mücadelede toplum desteğini sağlamaya yönelik sosyal ağlarda yaptığı paylaşımların, aynı ülke içinde veya uluslararası boyutta farklılaşmaması beklenmektedir.” hipotezi ve bu hipoteze ait aşağıda belirtilen araştırma sorularını cevaplandırmaya amaçlamaktadır.

- Kolluk kuvvetlerinin sosyal ağ kullanım alışkanlıkları nasıldır? (Hangi sıklıkla, hangi cihazlar üzerinden, ne tür mesajlar verdiği vb.)
- Kolluk tarafından paylaşılan mesajların makine öğrenimine dayalı duygu durumu (pozitif, negatif veya nötr) nasıldır?
- Kolluk tarafından yapılan paylaşımların konu modellemesi nasıldır, ağırlıklı işlediği konular neler? (LDA Analizi)
- Ülkemizde faaliyet gösteren Jandarma ve Polis tarafından yapılan paylaşımlara halkın verdiği tepki üzerinden mukayeseli korelasyon analizi ile anomali tespiti yapmak mümkün müdür?
- İtalya’da faaliyet gösteren Arma dei Carabinieri ve Polizia di Stato tarafından yapılan paylaşımlara halkın verdiği tepki üzerinden mukayeseli korelasyon analizi ile anomali tespiti yapmak mümkün müdür?
- Fransa’da faaliyet gösteren Gendarmerie Nationale ve Police Nationale tarafından yapılan paylaşımlara halkın verdiği tepki üzerinden mukayeseli korelasyon analizi ile anomali tespiti yapmak mümkün müdür?
- İspanya’da faaliyet gösteren Guardia Civil ve Policía Nacional tarafından yapılan paylaşımlara halkın verdiği tepki üzerinden mukayeseli korelasyon analizi ile anomali tespiti yapmak mümkün müdür?
- Ülkemizde, İtalya’da, Fransa’da ve İspanya’da faaliyet gösteren jandarma ve polis tarzı kolluk kuvvetlerinin sosyal ağdaki etki gücü nedir?

– Ülkemizde, İtalya’da, Fransa’da ve İspanya’da faaliyet gösteren jandarma ve polis tarzı kolluk kuvvetlerinin sosyal ağdaki paylaşımları arasında benzerlik var mıdır?



Şekil 2.10. TDGH kapsamında kolluk kuvvetlerinin paylaşımlarının işlem prosesi

Analiz kapsamında uygulanan işlem prosesi şekil 2.10'da diyagram halinde gösterilmekte olup;

- Dört farklı dilde 25627 tweet (6299 Türkçe (3071 adet Jandarma ve 3228 adet Polis), 6478 İtalyanca (3232 adet Carabinieri ve 3246 adet Polizia), 6380 Fransızca (3239 adet Gendarmerie ve 3141 adet Police), 6470 İspanyolca (3243 adet Guardia Civil ve 3227 adet Policia)) elde edilmiştir.

- Verilerin elde edilmesinde öncelikle ASCII karakter yapısı kullanılmış olup, fakat Türkçe'de, İtalyanca'da, Fransızca'da ve İspanyolca'da farklı harf karakter (Örneğin ş ve ü harfleri gibi) yapılarının kullanımından dolayı dönüşümde sıkıntılar yaşandığından UTF-8 unicode karakter kullanılarak veriler kodlanmıştır.

- Makine çevirisi yapıldığından dolayı ve metin içinde kullanılan dile has harf karakterlerinin kaybolmaması için metin ön işleme iki safhada icra edilmiş olup, ilk safhada sadece rakam, emoji ve karakter boşlukları (kelime arasındaki boşluklar hariç) kaldırılmış olup, tarih/saat, beğeni ve retweet yönleriyle keşifsel veri analizi yapılmıştır.

- Textblob kütüphanesi kullanılarak, Google translate üzerinden Türkçe, İtalyanca, Fransızca ve İspanyolca paylaşımlar ortak olarak İngilizceye çevrilmiştir. Müteakiben İngilizce durağan kelimeler cümleden kaldırılarak, ön işlem metin temizleme aşamaları tekrarlanmıştır.

- Paylaşımların duygu durumları, Textblob kütüphanesi üzerinden tanımlanan kelimeler yardımıyla yapılmıştır. Olumlu kelimeler kümesi derecesine göre $X > 0$, yine olumsuz kelimeler kümesi de $X < 0$ olarak ele alınmış olup, bunların dışında yer kelimeler ise $X = 0$ nötr olarak kabul edilmiştir. Her cümlenin polarite durumu, kelime değerlerinin toplamının 0 değerine karşı konumuna göre olmuş, aynı zamanda cümlelere ait öznellik durumu da ortaya konulmuştur. Polarite değerleri kendi içerisinde teyit maksatlı makine öğrenimine tabi tutulmuştur.

- Zamansal ve anlamsal anomali paylaşımların tespiti için Matplotlib ve Sklearn kütüphaneleri, bunların dışında kelime bulutları gösterimleri için de Wordcloud kütüphanesi kullanılmıştır.

- Konu modelleme öncesi NLTK üzerinden WordNet kullanılarak, kelime kökleri bulunmuştur ve TF-IDF üzerinden vektörel düzlemde frekans değerleri

hesaplanmıştır. Konu modellemesinde LDA kullanılmış olup, iterasyon değeri 5 olarak ele alınmıştır (n_components=10, learning_method='online', verbose=True). pyLDavis kütüphanesi üzerinden interaktif görsel elde edilmiştir.

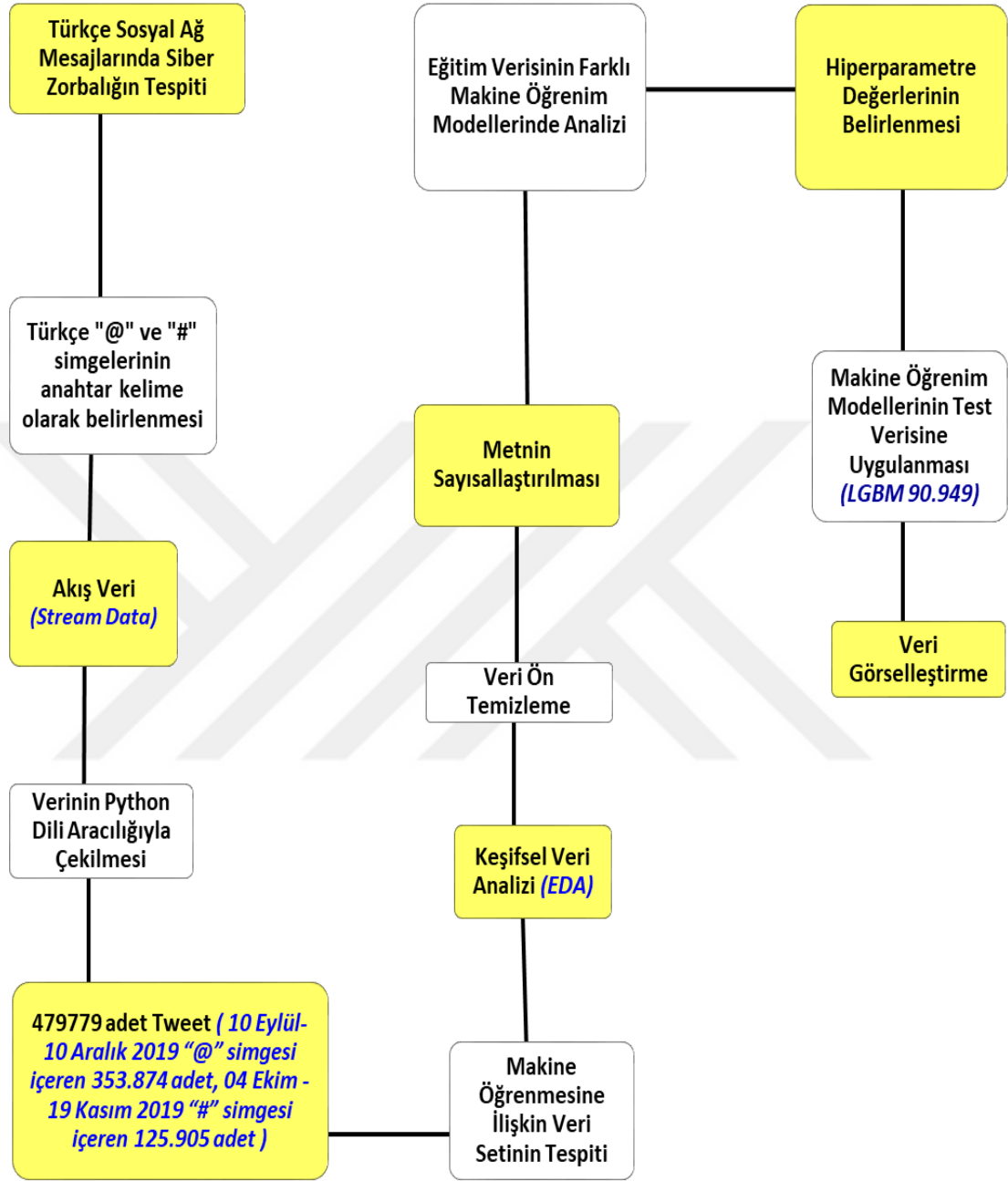
– Benzerlik için ikili analizlerde cosinüs, bütün kolluk birimlerinin karşılaştırılmasında ise TF-IDF kullanılmış olup, ayrıca kelimelerin diğer hangi kelime grubuna bağlandığı ve en sık kullanılan kelimelere ilişkin ilişkisel ve görsel analizler yapılmıştır.

2.5.2. Türkçe Sosyal Ağ Mesajlarında Siber Zorbalığın Tespiti

Twitter örnekleminde sosyal ağlar üzerinde yapılan paylaşımlardan yola çıkarak makine öğrenmesi temelinde otomasyonel olarak paylaşımın siber zorbalık türü bir mesaj olup olmasının tespiti, çalışmanın amacıdır. Makine öğrenimi yapılabilmesi için Türkçe tek siber zorbalık veri seti olan, Bozyiğit (2018) tarafından yayınlanan çalışma öğrenme amaçlı olarak kullanılmıştır (Bozyiğit ve ark., 2019). Söz konusu veri seti içerisinde 1500'ü siber zorbalık içeren, 1500'ü de siber zorbalık içermeyen 3000 sosyal medya mesajı yer almaktadır. Bu çalışma “Makine öğrenimi modelleriyle sosyal ağlardaki Türkçe metinler üzerinde siber zorbalık tespiti yapmak mümkündür.” hipotezi ve bu hipoteze ait aşağıda belirtilen araştırma sorularını cevaplandırmaya amaçlamaktadır.

- Siber zorbalık içeren ve içermeyen mesajlar arasında keşifsel analizde farklılık var mı?
- Siber zorbalığın tespiti denetimli öğrenme kapsamında yapılabilir mi?
- Siber zorbalığın tespiti denetimsiz öğrenme kapsamında yapılabilir mi?
- Farklı makine öğrenimi modelleri arasında F1 skor oranları nasıldır?
- Suçla mücadelede makine öğrenim modellerinin kullanımı işgücü ve kolluksal etkinlik gücünü artırır mı?
- Veri setinin büyük veri üzerindeki sonuçları nasıldır? Yeterli mi, geliştirilmeli mi?

– Türkçe metinlerde tespiti optimize edilmesi kapsamında dil gramer yapısı etkili mi?



Şekil 2.11. Makine öğrenimine dayalı siber zorbalık tespiti işlem prosesi

Analiz kapsamında uygulanan işlem prosesi şekil 2.11’de diyagram halinde gösterilmekte olup;

– 10 Eylül 2019 - 10 Aralık 2019 tarihleri arasında “@” simgesi içeren 353.874 adet tweet ile 04 Ekim 2019 - 19 Kasım 2019 tarihleri arasında “#” simgesi içeren 125.905 adet tweet olmak üzere, 479779 adet tweet (bunların arasında veri temizliği sonrası 7779 adet tweet boş (null) olarak kalmış, bunlar ana korpusdan çıkartılarak tweet sayısı 472000 olarak alınmıştır.) makine öğrenimi performansının gerçek verilerde boyutu ne denli azalttığını görebilmek adına kullanılmıştır.

– Keşifsel veri analizi yapılarak, kelimelerdeki karakter sayıları, cümlelerdeki kelime sayıları ve kelime karakter uzunlukları tespit edilerek, eğitim ve test verileri arasında anomali yaratan bir durum olup olmadığı incelenmiştir.

– Makine öğrenmesine ilişkin 3000 mesajlık denetimli veri seti kullanılmış olup, mevcut veri setinin (%50 - %50) dengeli dağılım gösterdiği belirlenmiştir.

– Metin içinde kullanılan dile has harf karakterleri, rakam, emoji ve karakter boşlukları (kelime arasındaki boşluklar hariç) kaldırılmıştır.

– Sosyal medyada informal dilin yaygın kullanımı sebebiyle içeriğinde kötü anlam ifade ettiği belirlenen sözcüklerin yanlış ya da eksik yazımları da tespit edilerek, yapılan tespitler doğrultusunda ön işlem safhasında gerekli düzeltmeler yapılmıştır (Örneğin; “okuz”, “ökuz” ve “öküz” sözcüğü “öküz” olarak işleme alınmıştır.).

– İngilizce ve birçok dil için TextBlob, SymSpell, Pyspellchecker ve Autocorrect gibi kütüphaneler kullanılabiliyorken, Türkçede ise Zemberek, ITU Turkish NLP Web Service API ve turkishnlp gibi kütüphaneler kullanılabilmektedir. Kullanılan dillerde mevcut dilin kullanım çeşidinin ya da yanlış yazımının neredeyse sonsuz büyüklükte bir veri kümesi doğuracağından dolayı, söz konusu kütüphaneler arasında kesin güvenilirlikle “daha iyi” şeklinde tahminleme yapmak imkânsızdır. Dil işlemenin karışık yapısından kaynaklı hiçbir kütüphanenin mükemmel sonuçlar ortaya koyamadığı birçok çalışmada mevcuttur. Türkçe metinlerde yapılan stemming ve lemmatization işlemlerinin İngilizce dili kadar başarılı olmadığı gözlemlenmiştir. Bunun en temel sebeplerinden ilki Türkçe dilinin yaygınlığının İngilizceye göre daha az olması sebebiyle başarılı kelime tasnif örneklerinin bulunmaması ve diğer bir sebebi de Türkçenin sondan çekim alan dil olması sebebiyle İngilizceye göre çözümlemesinin daha zor olmasıdır. Ayrıca Türkçe sondan eklemeli bir dil olduğu için köke indirgeme anlam yönünü değiştirmektedir. Tüm bu sebeplerden dolayı stemming ve lemmatization uygulanmamıştır.

– Metinleri sınıflandırmada en önemli nokta, makine öğrenimi modellerine uygun olarak metinlerin sayısal hale getirilmesidir. Bu çalışmada kelime çantası (BoW (Bag-of-Words)) metodu uygulanarak özellik vektörü oluşturulmuştur. Kelime Torbası özelliklerini çıkarmak için öncelikle unigram içeren bir kelime dağarcığı oluşturulmuş ve belge frekansı 2'nin altında olan terimler göz ardı edilmiştir. Sonrasında ise elde edilen unigramların TF-IDF özellikleri hesaplanmıştır. Tüm işlemler, Python Scikit-learn kütüphanesi kullanılarak yapılmıştır.

– Textblob kütüphanesi kullanılarak, Türkçe durağan kelimeler cümleden kaldırılmıştır.

– Eğitim ve test veri seti %70 ve %30 oranlarında bölünmüştür.

– Test verisi üzerinden yarı denetimli öğrenmenin en iyi sınıflamayla gerçekleşmesi için eğitim verisi üzerinde öğrenmenin farklı algoritmalar temelinde optimum düzeyde olması sistemin ideal düzenidir. Bu sebeple eğitim veri seti üzerinde çeşitli makine öğrenimi modelleri sırayla uygulanmış olup, kullanılan modeller; Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, Karar Ağacı Modeli (Decision Tree Model), Ekstra Ağaçlar Sınıflandırması (Extra Trees Classifier), Lineer Diskriminant Analizi (Linear Discriminant Analysis), Karesel Diskriminant Analizi (Quadratic Discriminant Analysis), AdaBoost (Adaptive Boosting), Gradyan Artırma Modeli (Gradient Boosting Model), Rastgele Orman (Random Forest), Lojistik Regresyon Modeli (Logistic Regression Model), Perceptron Sinir Ağı (Perceptron Neural Network), Lineer Karar Destek Sınıflandırması (Linear Support Vector Classification), XGBoost, K-En Yakın Komşu Modeli (K - Nearest Neighbors Model (KNN)), Destek Vektör Makineleri (Support Vector Machine (SVM)), Stokastik Gradyan Azaltma (Stochastic Gradient Descent (SGD)), Light Gradyan Artırma Modeli (Light Gradient Boosting Model (LGBM)) ve Voting Sınıflandırma Modeli (Voting Classifier Model) olarak yer almaktadır.

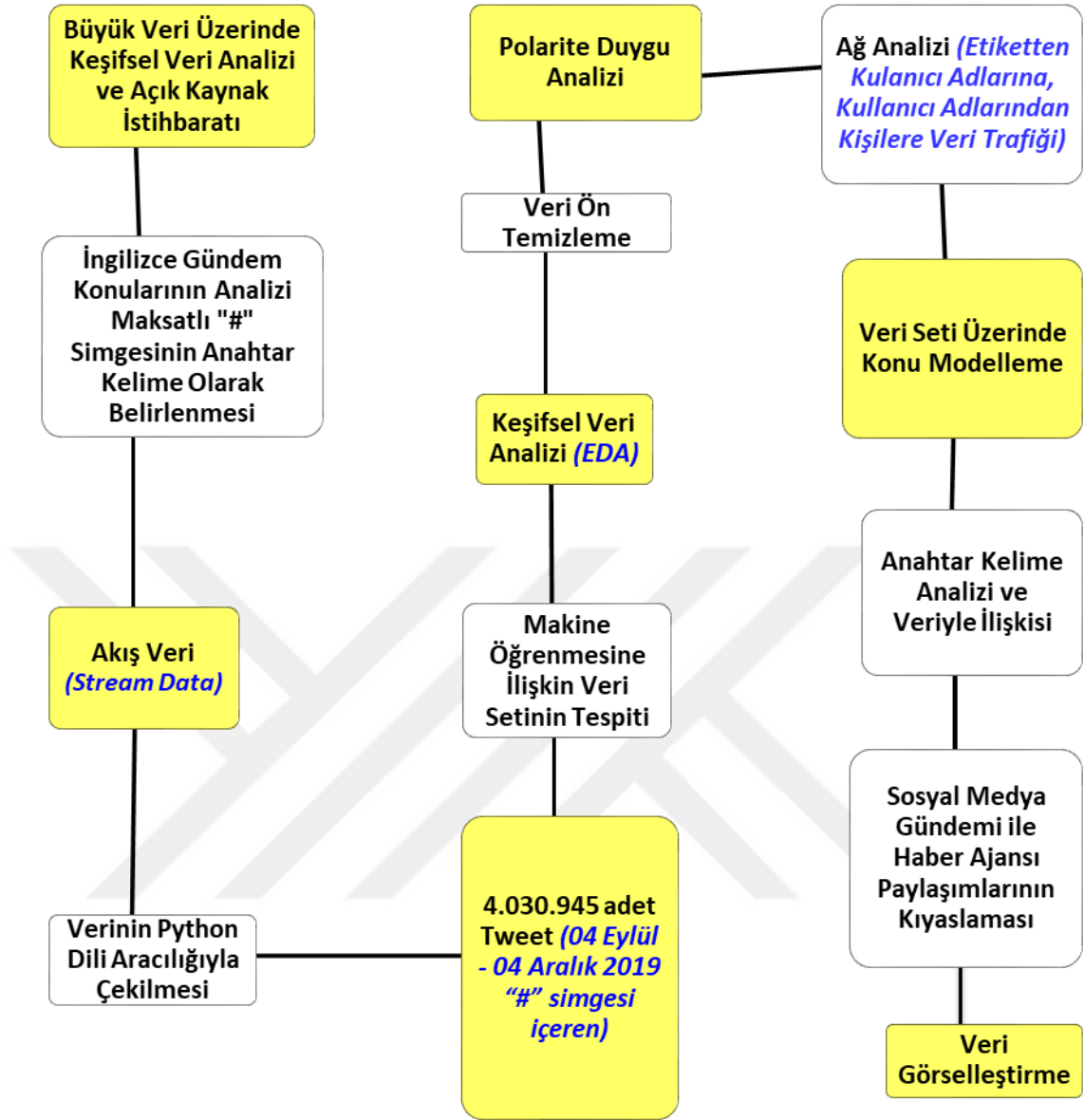
– Uygulanan makine öğrenim modellerinin hiperparametre ayarlamaları ve uygulanan değerler bulgular kısmında gösterilmiştir.

– Makine öğrenimi modellerinin doğruluk, kesinlik, anma ve F-ölçütü değerleriyle geçerliliği ölçülerek, karışıklık matrisinde de performansı gösterilmiştir.

2.5.3. Büyük Veri Üzerinde Keşifsel Veri Analizi ve Açık Kaynak İstihbaratı

Twitter örnekleminde sosyal ağlar üzerinde yapılan paylaşımlardan yola çıkarak gündem hakkında fikir edinmek, gündemi belirleyen kişi ve etiketleri bularak ağ analizi yapmak, paylaşımlardan duygu durumlarını belirlemek ve paylaşımların konu analizinin yapılması çalışmanın amacıdır. Bu çalışma “Veri madenciliği teknikleriyle sosyal ağlar üzerinde açık kaynak istihbaratı elde edilebilir.” hipotezi ve bu hipoteze ait aşağıda belirtilen araştırma sorularını cevaplandırmaya amaçlamaktadır.

- Herhangi bir konuya ilişkin atılan mesajların duygu durumu tespit edilebilir mi?
- Olumsuz mesajların (terör, hakaret, zorbalık vb. içeren) UTF sisteminde kullandığımız saat ile ilişkisi var mıdır?
- Herhangi bir konuda en etkin # ve @ etiketlerini görmek mümkün müdür?
- @ ile # işaretli etiketler arası veri ağı analizi yapmak mümkün mü?
- Denetimsiz öğrenme kapsamında konu modellemesi yapılabilir mi?
- Anahtar kelime analizi araması yapılarak, büyük veriden anahtar kelimeyle bağlantılı ilişkili veriler öğrenilebilir mi?
- @ etiketli etki gücü en yüksek hesaplar üzerine keşfedici analiz yapılabilir mi?
- Haber ajanslarından elde edilen bilgi, insanların sosyal medyada yazdıklarının ne kadarını karşılıyor?



Şekil 2.12. Büyük veri üzerinde keşifsel veri analizi işlem prosesi

Analiz kapsamında uygulanan işlem prosesi şekil 2.12’de diyagram halinde gösterilmekte olup;

- 04 Eylül 2019 - 04 Aralık 2019 tarihleri arasında İngilizce dil süzgecinde “#” simgesi içeren 4.030.945 adet tweet Twitter API aracılığıyla Python ile çekilmiştir.

- Keşifsel veri analizi yapılarak, kelimelerdeki karakter sayıları, cümlelerdeki kelime sayıları ve kelime karakter uzunlukları tespit edilerek, eğitim ve test verileri arasında anomali yaratan bir durum olup olmadığı incelenmiştir.
- Metin içinde kullanılan dile has harf karakterleri, rakam, emoji ve karakter boşlukları (kelime arasındaki boşluklar hariç) kaldırılmıştır.
- NLTK kütüphanesi kullanılarak, İngilizce durağan kelimeler cümleden kaldırılmıştır.
- Paylaşımların duygu durumları, Textblob kütüphanesi üzerinden tanımlanan kelimeler yardımıyla yapılmıştır. Olumlu kelimeler kümesi derecesine göre $X > 0$, yine olumsuz kelimeler kümesi de $X < 0$ olarak ele alınmış olup, bunların dışında yer kelimeler ise $X = 0$ nötr olarak kabul edilmiştir. Her cümlenin polarite durumu, kelime değerlerinin toplamının 0 değerine karşı konumuna göre olmuş, aynı zamanda cümlelere ait öznellik durumu da ortaya konulmuştur. Polarite değerleri kendi içerisinde teyit maksatlı karışıklık matrisi şeklinde incelenmiştir.
- Duygu durum çalışması aynı zamanda saat bazlı da incelenmiş olup, olumsuz mesajların UTF saat bazlı kullanımla ilişkili olup olmadığı incelenmiştir.
- Zamansal ve anlamsal anomali paylaşımların tespiti için Matplotlib ve Sklearn kütüphaneleri, bunların dışında kelime bulutları gösterimleri için de Wordcloud kütüphanesi kullanılmıştır.
- Konu modelleme öncesi İncelemeye konu veri seti içerisinde 4.030.945 tweet içerisinde sadece metin kısımları üzerinde ön işlem sonrası (durağan kelimelerin, url adreslerinin uzaklaştırılması vb.) 217.205.116 adet kelime elde edilmiştir. NLTK üzerinden WordNet kullanılarak, kelime kökleri bulunmuştur ve TF-IDF üzerinden vektörel düzlemde frekans değerleri hesaplanmıştır. Konu modellemesinde LDA kullanılmış olup, iterasyon değeri 5 olarak ele alınmıştır ($n\_components=10$, $learning\_method='online'$, $verbose=True$). pyLDavis kütüphanesi üzerinden interaktif görsel elde edilmiştir.
- Anahtar kelime analizinin yapılması için metnin Word2Vec metoduyla vektörel düzleme geçirilerek, 20 devirde (epoch) eğitilerek 93346 kelime elde edilmiştir. Gösterimlerde t-SNE(t-Distributed Stochastic Neighbor Embedding) kullanılmıştır.

– İngilizce “#” gündem paylaşımlarında en etkili hesaplar, en çok kullanılan mesajlar ilişkisel analiz ile bulunmuştur. Aynı zamanda etkili hesapların hangi gündem maddelerini takip ettiği veya hangi gündem maddelerinin hangi kullanıcılar tarafından takip edildiğine ilişkin örnek veri trafiği diyagramları Matplotlib Kütüphanesi aracılığıyla hazırlanmıştır.

– Gündem mesajlarında en etkili 200 kullanıcının giriş biyografilerinde yer alan bilgilere dayanarak, hangi amaçla kullanıldığı belirlenmiştir.

– Word2Vec denetimsiz ve tahmin temelli olarak kelime yapılarını vektörel düzlemde göstermek için kullanılan metin temsil yöntemi olup, bu çalışmada CBOW modeli kullanılarak, kenardaki kelimelerden faydalanılmış ve ortada yer alan kelimeler tahmin edilmeye çalışılmıştır. Özellikle suç kelimelerinde alternatif anahtar kelimelerin hazırlanmasında oldukça faydalı çıkarımlar yapılmasına olanak sağlamıştır.

– Suç önleme ve tespit çalışmalarında reaktif ve proaktif tarzda kolluğun reaksiyon göstermesinin temeli suçun olduğunu ya da olabileceğini öğrenmesine bağlıdır. Bu sebeple de her türlü bilginin yer aldığı sosyal ağlardan zaman zaman istihbari çıkarımlar da yapılarak, suç açısından kullanılabilir. Sosyal ağlarda “#” etiketiyle konuşulan konular ile aynı tarihlerde “Al Jazeera News”, “BBC Breaking News”, “BBC News (World)”, “CNN International”, “CNN”, “Reuters” ve haber paylaşımlarının tamamının birleşimi üzerinden TF-IDF kullanılarak benzerlik analizi yapılmıştır.

3. BULGULAR

Araştırma kapsamında elde edilen veriler, aşağıda belirtilen yedi ana maddede toplanmış olup, her maddede bulgular yer almaktadır.

- Türkiye özelinde kolluk teşkilatlarının paylaşımlarının incelenmesi
- İtalya özelinde kolluk teşkilatlarının paylaşımlarının incelenmesi
- Fransa özelinde kolluk teşkilatlarının paylaşımlarının incelenmesi
- İspanya özelinde kolluk teşkilatlarının paylaşımlarının incelenmesi
- Kolluk paylaşımlarının uluslararası düzeyde mukayeseli incelenmesi
- Denetimli ve yarı denetimli makine öğrenimine dayalı siber zorbalık tespiti
- Denetimsiz öğrenme kapsamında büyük veri üzerinde keşifsel veri analizi ve açık kaynak istihbaratı

3.1. Türkiye Özelinde Kolluk Teşkilatlarının Paylaşımlarının İncelenmesi

Jandarma Genel Komutanlığı ve Emniyet Genel Müdürlüğünün resmi paylaşımlarından alınan örneklem üzerinden keşfedici veri analizi yapılarak, sosyal medya kullanım alışkanlıkları üzerinden duyarlılık analizi ve söz konusu duyarlılığın makine öğrenimine dayalı tahminlemesi yapılmıştır.

3.1.1. Jandarma Genel Komutanlığı'nın Paylaşımlarının İncelenmesi

Jandarma Genel Komutanlığının resmi hesabı olan “T.C. Jandarma Gn. K.” başlıklı ve @jandarma kullanıcı adlı Twitter hesabından 09 Mart 2020 tarihinde herkese açık olarak yapılan paylaşımlar Tweepy kütüphanesi aracılığıyla Python programlama dili kullanılarak, csv formatında çekilmiştir.



Şekil 3.1. Jandarma Genel Komutanlığının resmi Twitter sayfasının ekran resmi

```
print(data.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3071 entries, 0 to 3070
Data columns (total 7 columns):
id                3071 non-null int64
created_at        3071 non-null object
text              3071 non-null object
followers_count   3071 non-null int64
retweet_count     3071 non-null int64
favourite_count   3071 non-null int64
source            3071 non-null object
dtypes: int64(4), object(3)
memory usage: 168.0+ KB
None
```

Şekil 3.2. @jandarma kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-I

	id	created_at	text	followers_count	retweet_count	favourite_count	source
0	1236689290253516808	2020-03-08 16:24:38	Yalnız değilsiniz 🇹🇷'ınJandarma Genel Komu...	348381	257	2675	Twitter Media Studio
1	1236517012337307648	2020-03-08 05:00:03	Rüzgar ettik, sevgi biçiyoruz♥'ınJandarma...	348381	1446	9096	Twitter Media Studio
2	1236350922927812609	2020-03-07 18:00:04	Yiğitlik üniformamız, 'ınKahramanlık rütbemizd...	348381	606	5962	Twitter for iPhone
3	1236320415347335173	2020-03-07 15:58:51	Fedakarlık'ınKahramanlık'ınCesaretle'ın...	348381	688	6671	Twitter for iPhone
4	1236200271010725888	2020-03-07 08:01:26	Onları yalnız bırakmıyoruz.'ın'ın#Van il #Ja...	348381	637	6161	Twitter for iPhone

Şekil 3.3. @jandarma kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler-II

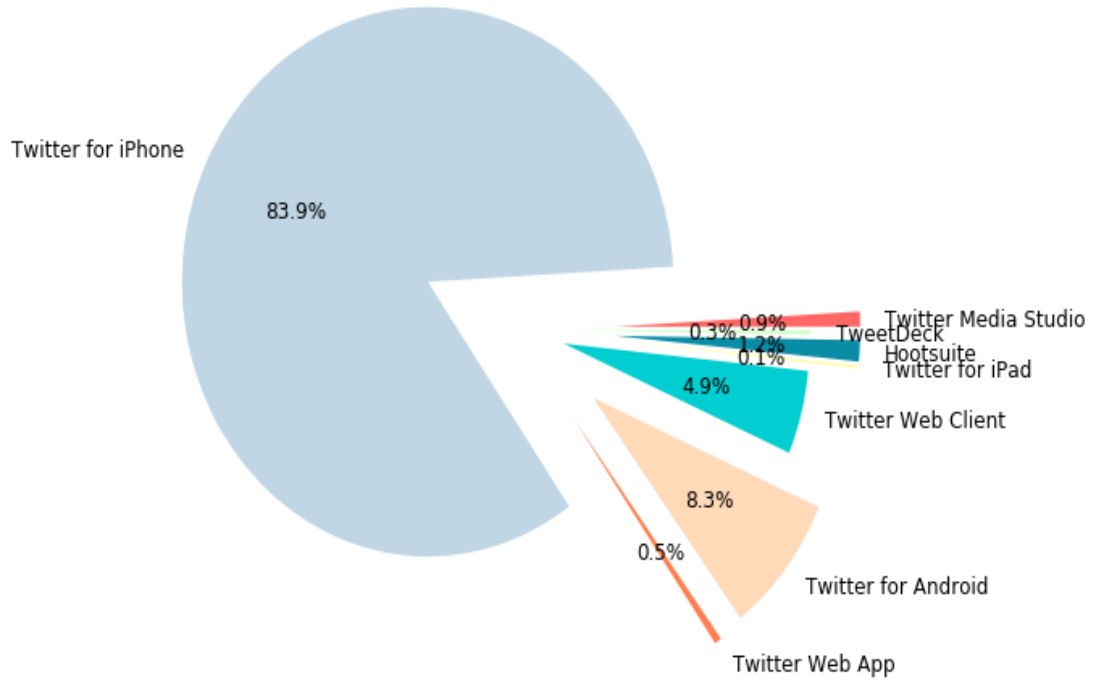
```
print(f" Data Available since {data.created_at.min()}")
print(f" Data Available upto {data.created_at.max()}")

Data Available since 2017-03-02 14:36:59
Data Available upto 2020-03-08 16:24:38
```

Şekil 3.4. @jandarma kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler–III

Pandas kütüphanesi aracılığıyla @jandarma kullanıcı adlı hesaptan çekilen veriler üzerinde ön inceleme yapıldığında;

- Verinin 3071 satır ve 7 sütundan oluştuğu (Şekil 3.1, Şekil 3.2),
- Verinin sütun isimlerinin “id”, “created_at”, “text”, “followers_count”, “retweet_count”, “favourite_count” ve “source” olduğu (Şekil 3.3),
- Veri tipleri olarak “created_at”, “text” ve “source” sütunlarının object (nesne), diğer sütunların ise int64 (sayısal - integer) olarak tanımlandığı (Şekil 3.3),
- Veride 0' dan 3070'e kadar numaralandırılmış olan 3071 adet girdi bulunduğu ve her bir sütunda boş olmayan (non-null) girdi sayısının tanımlandığı, bu değerlerden de değeri olmayan sütun değerinin bulunmadığının anlaşıldığı (Şekil 3.3),
- @jandarma adlı kullanıcıdan çekilen verilerin 02 Mart 2017'den 08 Mart 2020 tarihleri arasında yer aldığı (Şekil 3.4) anlaşılmaktadır.



Şekil 3.5. @jandarma adlı kullanıcının veri paylaşımında kullandığı kaynaklar

@jandarma adlı kullanıcının Twitter paylaşımı yaptığı cihazların kaynağı incelendiğinde; %83,9'unun Iphone, %8,3'ünün Android, %4,9'unun Twitter web istemcisi (web client), %0,5'inin Twitter web uygulaması, %0,1'inin Ipad, %0,3'ünün TweetDeck, %1,2'sinin Hootsuite, %0,9'unun da Twitter medya stüdyosu aracılığıyla olduğu şekil 3.5'de gösterilmektedir.

```
print(f" Maximum number of retweets {data.retweet_count.max()}")  
print(f" Maximum number of favorites {data.favourite_count.max()}")
```

```
Maximum number of retweets 36470  
Maximum number of favorites 73785
```

```
data.loc[data['retweet_count']==36470,['text']].values  
array([[ 'RT @RTErdogan: Türkiye, bugünü ve geleceği bakımından tarihi '  
        dtype=object)
```

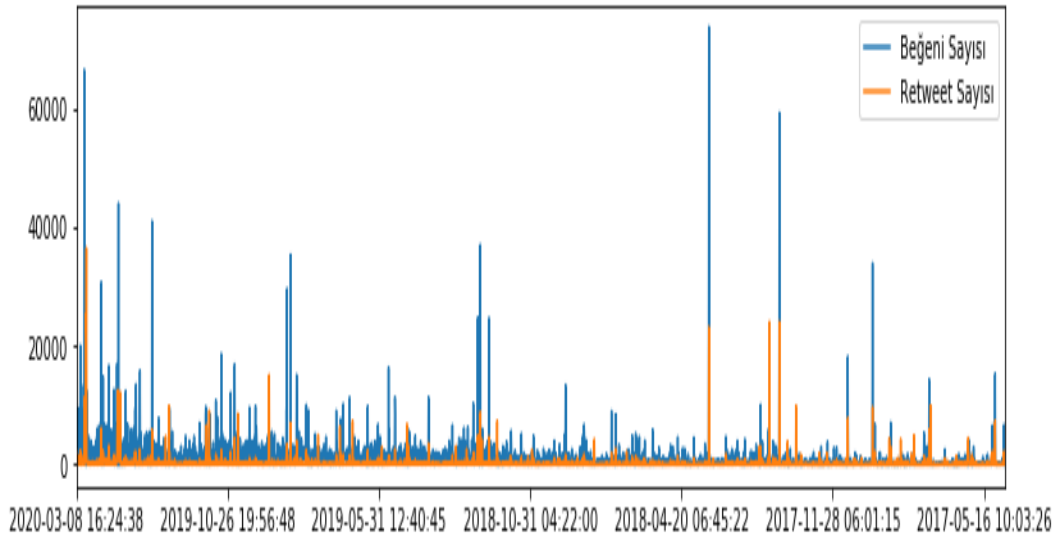
```
data.loc[data['retweet_count']==36470, ['created_at']].values  
array([[ '2020-02-29 22:15:53']], dtype=object)
```

```
data.loc[data['favourite_count']==73785,['text']].values  
array([[ 'İle uwez zısueşw! zııızeşeq nıoz" #tbt #JandarmaHavacılık '  
        dtype=object)
```

```
data.loc[data['favourite_count']==73785, ['created_at']].values  
array([[ '2018-03-29 08:19:03']], dtype=object)
```

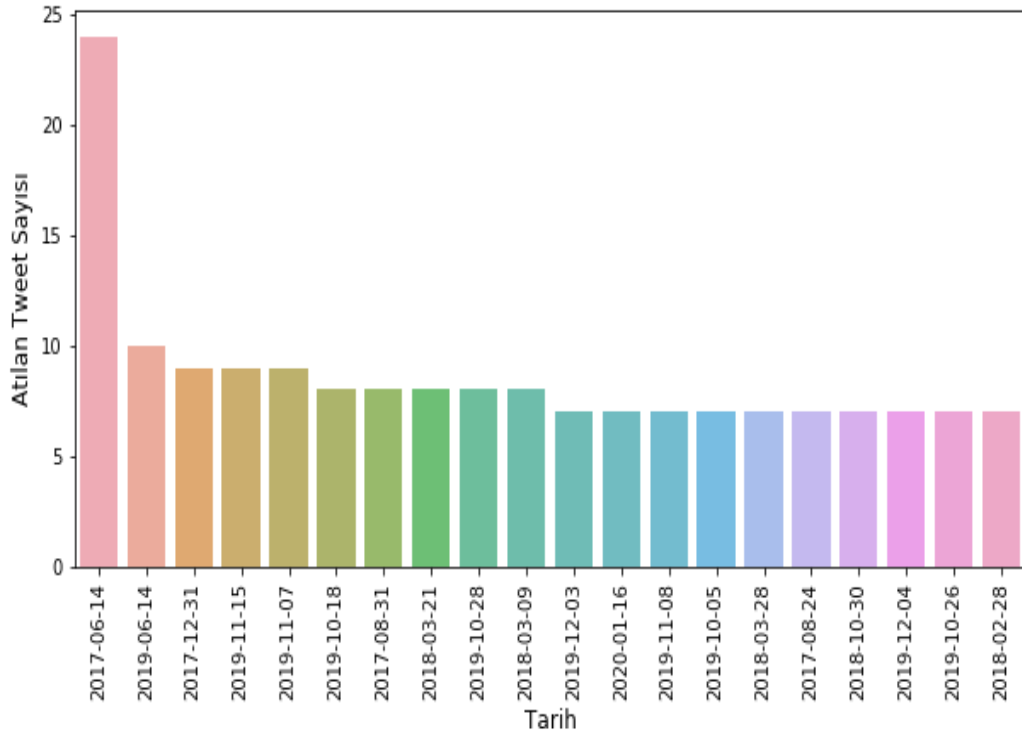
Şekil 3.6. Paylaşımların beğenilme ve retweet durumunun incelenmesi

Yapılan paylaşımlar şekil 3.6'da görüldüğü üzere incelendiğinde, 36470 tekrar paylaşım oranı ile en fazla retweet yapılan paylaşımın 29 Şubat 2020 tarihinde paylaşılan yine bir retweet mesajı olduğu, 73785 beğeni ile de en fazla beğeni alan paylaşımın 29 Mart 2018 tarihinde yapıldığı anlaşılmaktadır. Ayrıca Twitter paylaşımlarının 4.426.875 kez beğenildiği (favorites), 1.106.875 kez tekrar paylaşıldığı (retweet) ve toplam takipçi sayısının ise 348.381 olduğu belirlenmiştir.

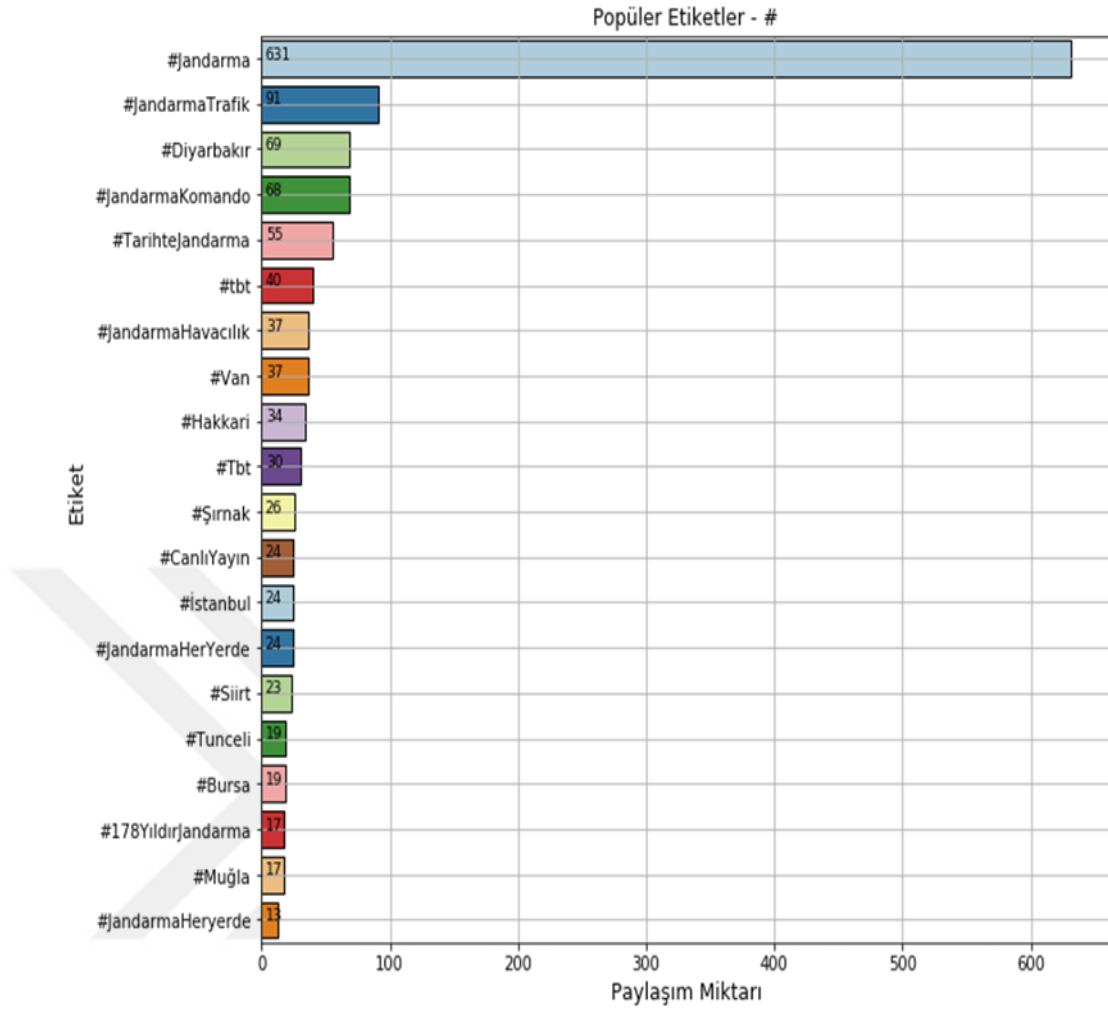


Şekil 3.7. @jandarma adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu

@jandarma adlı kullanıcının Twitter paylaşımlarının beğenilme ve tekrar paylaşılma durumlarının tarih bazlı ortak grafiği şekil 3.7’de gösterilmektedir (Takipçi sayısının tarih bazında sabit olmasından dolayı “followers_count” değeri ile tweet kimliğini içeren “id” değerinde ilişkisel analiz incelenmemiştir.).

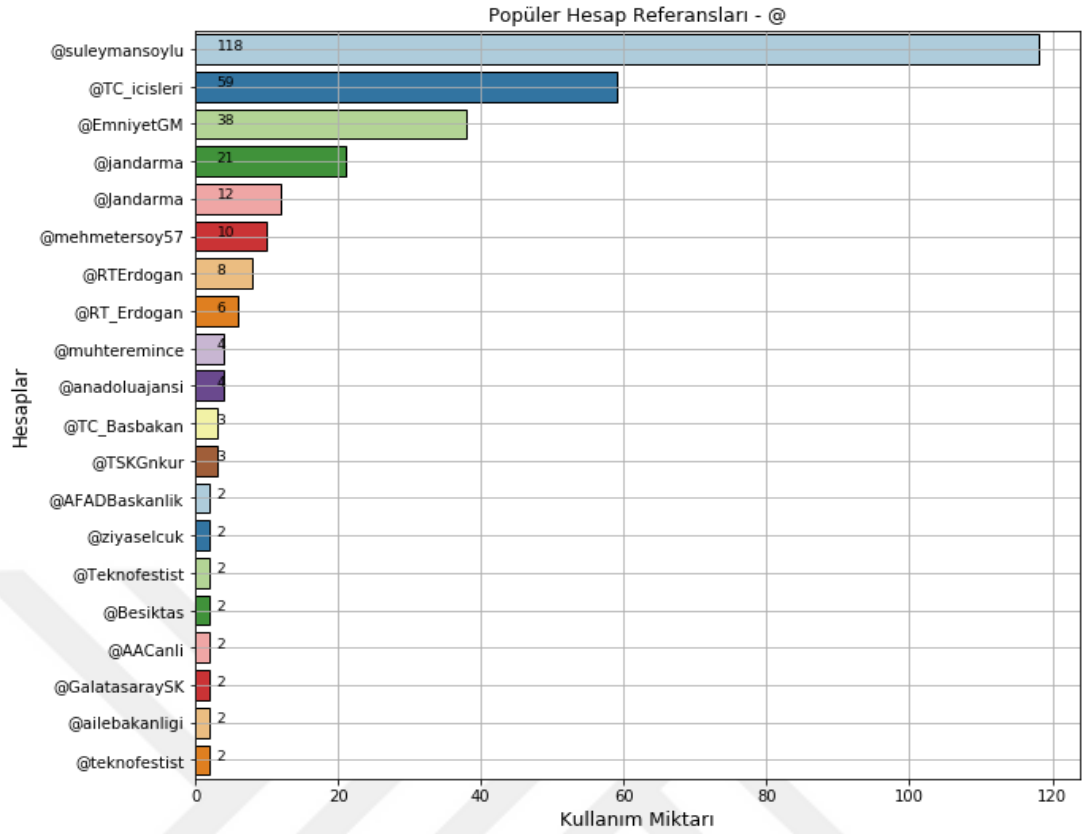


Şekil 3.8. Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı



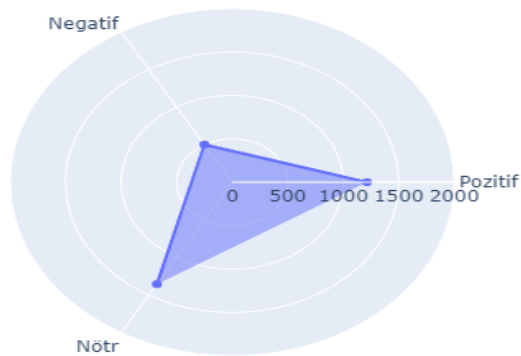
Şekil 3.11. En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)

Yapılan paylaşımlarda en fazla kullanılan sosyal medya etiketlerinin 20'sine ilişkin kullanım sıklığı ile sık kullanılan sosyal medya etiketlerine ilişkin hazırlanan kelime bulutu (Maksimum kelime girdisi 1000 olarak alınmıştır.) şekil 3.10 ve 3.11'de gösterilmektedir.



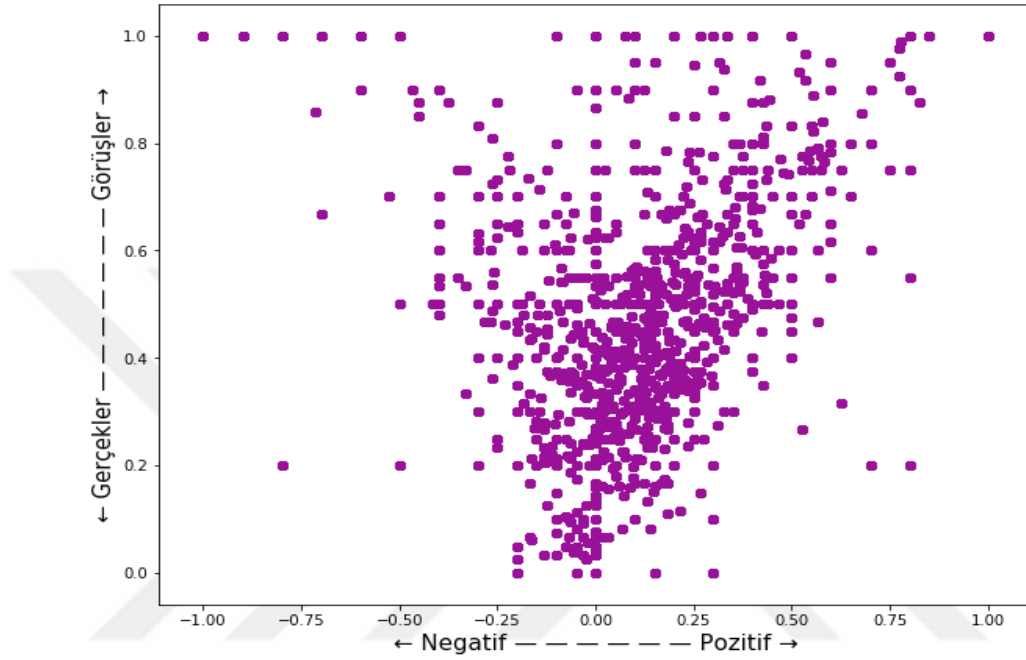
Şekil 3.12. En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)

Yapılan paylaşımlarda en fazla kullanılan sosyal medya hesaplarının 20'sine ilişkin kullanım sıklığı şekil 3.12'de gösterilmektedir. Ayrıca @jandarma kullanıcı hesabından yapılan paylaşım metinleri, metin ön işlemesine (text preprocessing) tabi tutularak tüm emojiler kaldırılmış ve Türkçe/İngilizce karakter farklılıklarından dolayı metindeki tüm harfler küçültülerek makine öğrenimine dayalı olarak Türkçe'den İngilizce'ye çeviri yapılmıştır.



Şekil 3.13. Metnin duygusal durumunun radar grafiği

Çeviri metni üzerinde yapılan duygusal dağılım (0=Nötr, 0>Negatif, 0<Pozitif) şekil 3.13’de görüldüğü üzere 1356 tweet ile nötr mesajların çoğunlukta olduğu, 1215 tweet ile pozitif mesajların nötr mesajlarını izlediği ve son olarak da 500 tweet ile negatif skorlanan mesajların bulunduğu belirlenmiştir.



Şekil 3.14. Metnin duygusal durumunun öznellik analizi

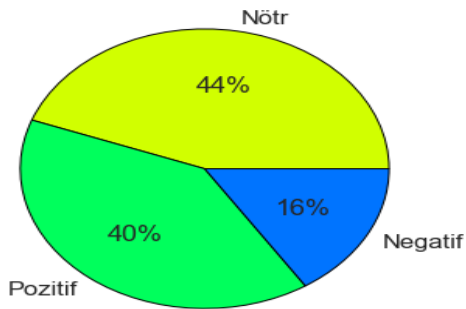
Elde edilen çeviri üzerinden kelime bazlı olarak, her kelimenin yarattığı duygusal durum ortaya konularak, her cümleye ait duygusal puan durumu, duygusal durumu ve ortaya konulan kanının öznellik ihtimali çıkarılmıştır. Metnin duygusal durumu ile öznelliğini beraber analiz ettiğimizde ise, cümlelerin içinde yer alan her kelimenin skorlanması ile cümle skoru elde edilmiş olup, ana korpusun çoğunluğunun 0.5 öznellik bandını geçmediği şekil 3.14’de gösterilmektedir. Fakat az sayıda da olsa öznelliği 1 olarak yer alan 19 tweet için, negatif, nötr veya pozitif gibi değerlendirmelerde bulunulmuştur.



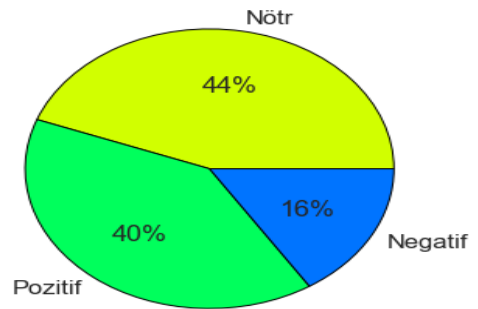
Şekil 3.17. Yapılan paylaşımlara ait kelime ağı analizi (100>Kelime Kombinasyonu)

Yapılan paylaşım metinleri BoW (Bag of words) metoduyla veri seti haline getirilmiştir. Söz konusu veri seti içerisinde de 100’den fazla kombinasyona sahip kelimelerin ağ (network) yapısı şekil 3.17’de gösterilmiştir.

Eğitim Veri Seti - Jandarma %75

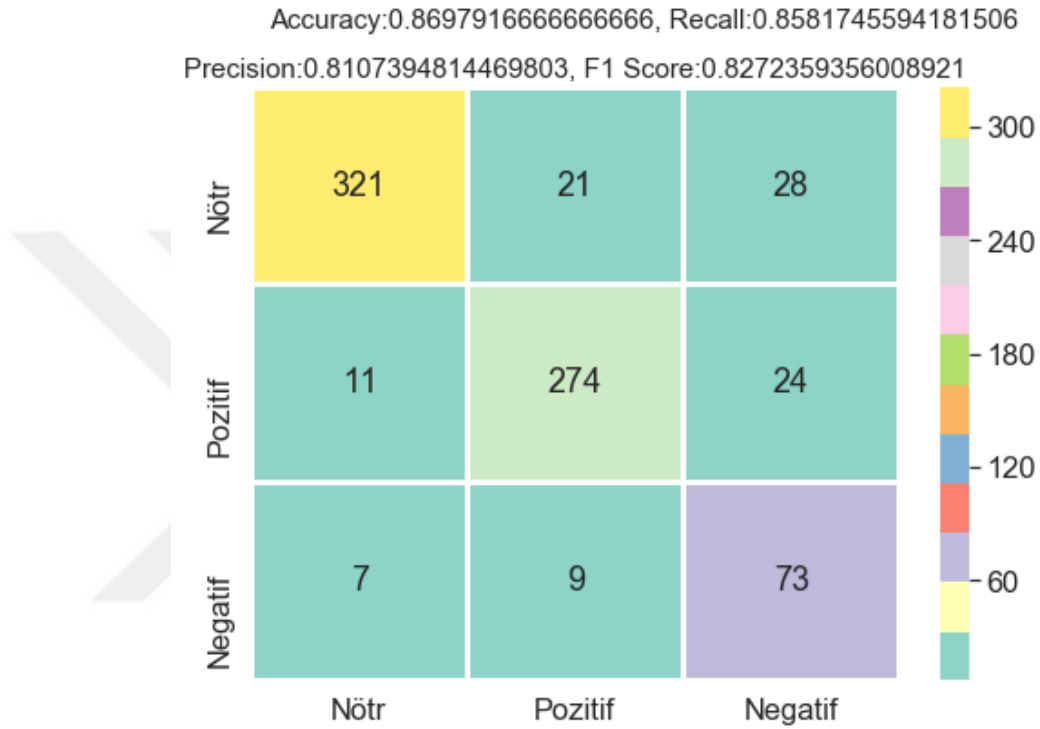


Test Veri Seti - Jandarma %25



Şekil 3.18. Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması

Textblob kütüphanesi aracılığıyla yapılan paylaşımların duygu durumları belirlenmiş olup, söz konusu tespitlerin kendi içinde tutarlı olup olmadığını incelemek adına makine öğrenimi modellerinden rastgele orman sınıflandırması (Random forest classification) ile veri seti şekil 3.18’de gösterildiği üzere test ve eğitim setine ayrılmıştır. Sınıflandırma için verinin %75’i eğitim, %25’i de test olarak ayrılmıştır.



Şekil 3.19. Test verisinin duygusal matriksi

Makine öğrenimi sonrası test verileri üzerinde çıkan sonuca ilişkin matriks şekil 3.19’da görülmekte olup, yapılan öğrenme test verileri üzerinde %86,979 oranında doğruluk oranı sağlamıştır.

3.1.2. Emniyet Genel Müdürlüğü’nün Paylaşımlarının İncelenmesi

Emniyet Genel Müdürlüğü’nün resmi hesabı olan Türk Polis Teşkilatı başlıklı ve @EmniyetGM kullanıcı adlı Twitter hesabından 09 Mart 2020 tarihinde herkese açık olarak yapılan paylaşımlar Tweepy kütüphanesi aracılığıyla Python programlama dili kullanılarak, csv formatında çekilmiştir.



Şekil 3.20. Emniyet Genel Müdürlüğünün resmi Twitter sayfasının ekran resmi

	id	created_at	text	followers_count	retweet_count	favourite_count	source
0	1236743020847853569	2020-03-08 19:58:08	RT @TrafikEgm: Ülkemizin dört bir yanında her ...	874235	61	0	Twitter for iPhone
1	1236602157283061760	2020-03-08 10:38:23	RT @korumadb: Memleketimizin dört bir yanında ...	874235	60	0	Twitter for iPhone
2	1236291997050245123	2020-03-07 14:05:55	RT @EmniyetAnkara: Tüm kadınların #8MartDünyaK...	874235	140	0	Twitter for iPhone
3	1236259771440926723	2020-03-07 11:57:52	Başta mesai arkadaşlarınızı olmak üzere tüm kad...	874235	520	4314	Twitter for iPhone
4	1236204350143414272	2020-03-07 08:17:39	Bazı sosyal medya platformlarında "Korona virü...	874235	858	3005	Twitter for iPhone

Şekil 3.21. @EmniyetGM kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler – I

```
print(data.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3228 entries, 0 to 3227
Data columns (total 7 columns):
id                3228 non-null int64
created_at        3228 non-null object
text              3228 non-null object
followers_count   3228 non-null int64
retweet_count     3228 non-null int64
favourite_count   3228 non-null int64
source            3228 non-null object
dtypes: int64(4), object(3)
memory usage: 176.7+ KB
None
```

Şekil 3.22. @EmniyetGM kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler – II

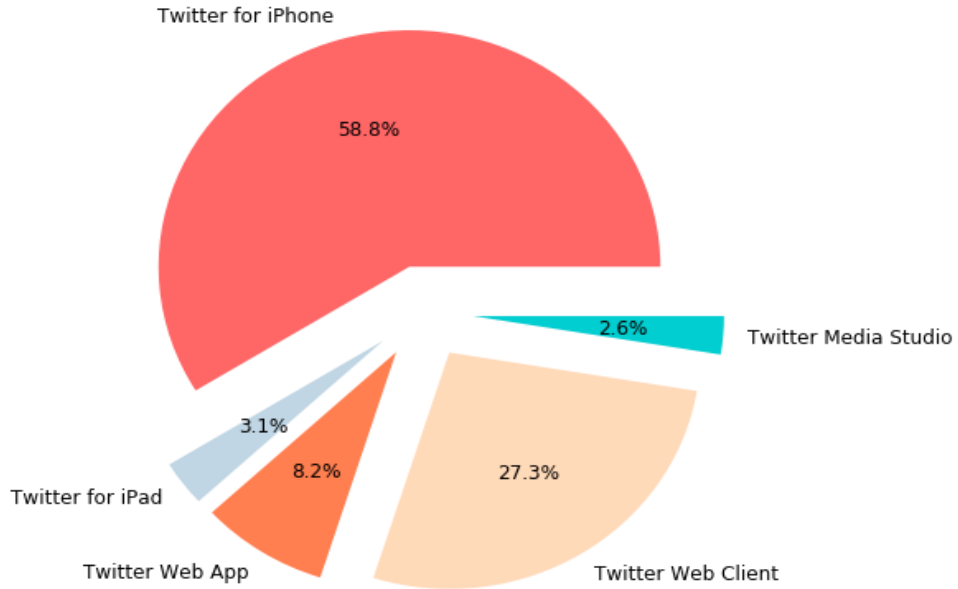
```
print(f" Data Available since {data.created_at.min()}")
print(f" Data Available upto {data.created_at.max()}")

Data Available since 2018-01-18 07:47:27
Data Available upto 2020-03-08 19:58:08
```

Şekil 3.23. @EmniyetGM kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler – III

Pandas kütüphanesi aracılığıyla @EmniyetGM kullanıcı adlı hesaptan çekilen veriler üzerinde ön inceleme yapıldığında;

- Verinin 3228 satır ve 7 sütundan oluştuğu (Şekil 3.20, Şekil 3.21),
- Verinin sütun isimlerinin “id”, “created_at”, “text”, “followers_count”, “retweet_count”, “favourite_count” ve “source” olduğu (Şekil 3.22),
- Veri tipleri olarak “created_at”, “text” ve “source” sütunlarının object (nesne), diğer sütunların ise int64 (sayısal - integer) olarak tanımlandığı (Şekil 3.22),
- Veride 0' dan 3227'ye kadar numaralandırılmış olan 3228 adet girdi bulunduğu ve her bir sütunda boş olmayan (non-null) girdi sayısının tanımlandığı, bu değerlerden de değeri olmayan sütun değerinin bulunmadığının anlaşıldığı (Şekil 3.22),
- @EmniyetGM adlı kullanıcıdan çekilen verilerin 18 Ocak 2018'den 08 Mart 2020 tarihleri arasında yer aldığı (Şekil 3.23) anlaşılmaktadır.



Şekil 3.24. @EmniyetGM adlı kullanıcının veri paylaşımında kullandığı kaynaklar

@EmniyetGM adlı kullanıcının Twitter paylaşımı yaptığı cihazların kaynağı incelendiğinde; %58,8’inin Iphone, %27,3’ünün Twitter web istemcisi (web client), %8,2’sinin Twitter web uygulaması, %3,1’inin Ipad ve %2,6’sının da Twitter medya stüdyosu aracılığıyla olduğu şekil 3.24’de gösterilmektedir.

```
print(f" Maximum number of retweets {data.retweet_count.max()}")
print(f" Maximum number of favorites {data.favourite_count.max()}")

Maximum number of retweets 36470
Maximum number of favorites 64459

data.loc[data['retweet_count']==36470,['text']].values

array([[ 'RT @RTErdogan: Türkiye, bugünü ve geleceği bakımından tarihi ve hayati bir mücadele içerisin
dtype=object)

data.loc[data['retweet_count']==36470, ['created_at']].values

array([[ '2020-02-29 20:07:51']], dtype=object)

data.loc[data['favourite_count']==64459,['text']].values

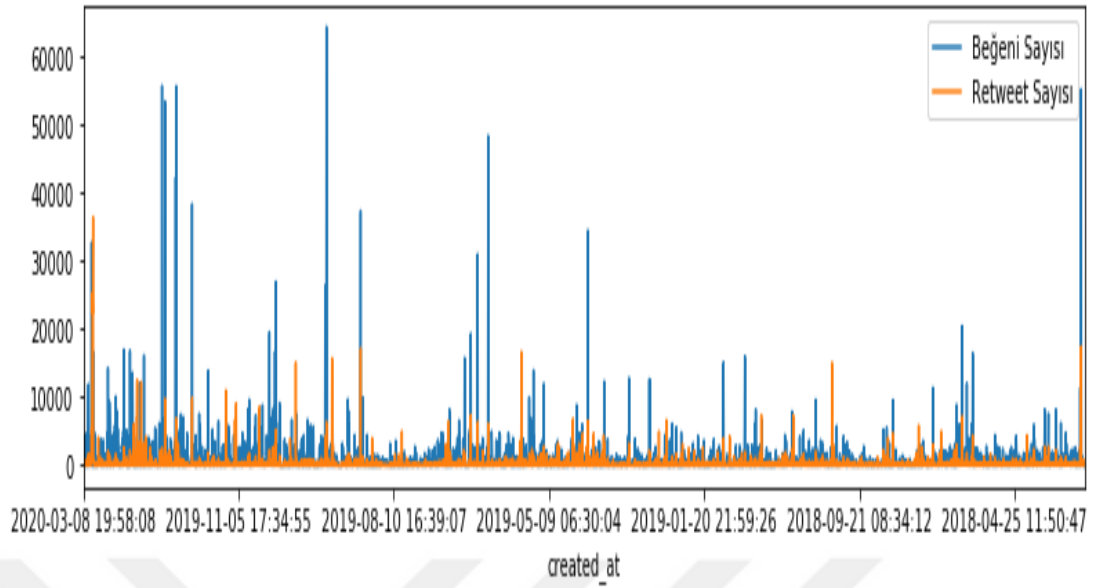
array([[ 'Bizden kaçmaz.. 😊 https://t.co/EuAJKlakGk']], dtype=object)

data.loc[data['favourite_count']==64459,['created_at']].values

array([[ '2019-09-10 13:20:00']], dtype=object)
```

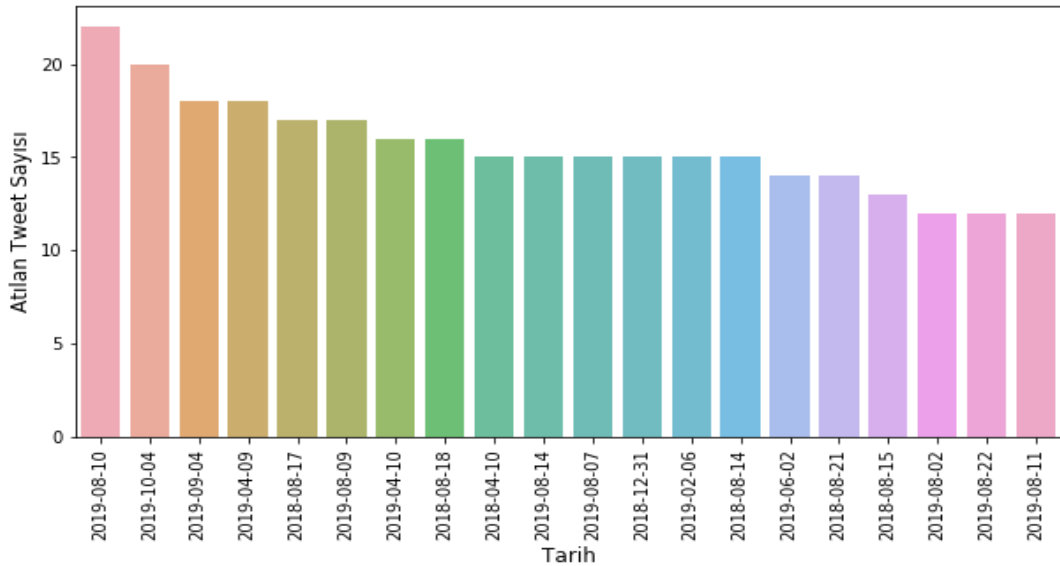
Şekil 3.25. Paylaşımların beğenilme ve retweet durumunun incelenmesi

Yapılan paylaşımlar şekil 3.25’de görüldüğü üzere incelendiğinde, 36470 tekrar paylaşım oranı ile en fazla retweet yapılan paylaşımın 29 Şubat 2020 tarihinde paylaşılan yine bir retweet mesajı olduğu, 64459 beğeni ile de en fazla beğeni alan paylaşımın 10 Mart 2019 tarihinde yapıldığı anlaşılmaktadır. Ayrıca Twitter paylaşımlarının 3.710.552 kez beğenildiği (favorites), 1.306.019 kez tekrar paylaşıldığı (retweet) ve toplam takipçi sayısının ise 874.235 olduğu belirlenmiştir.



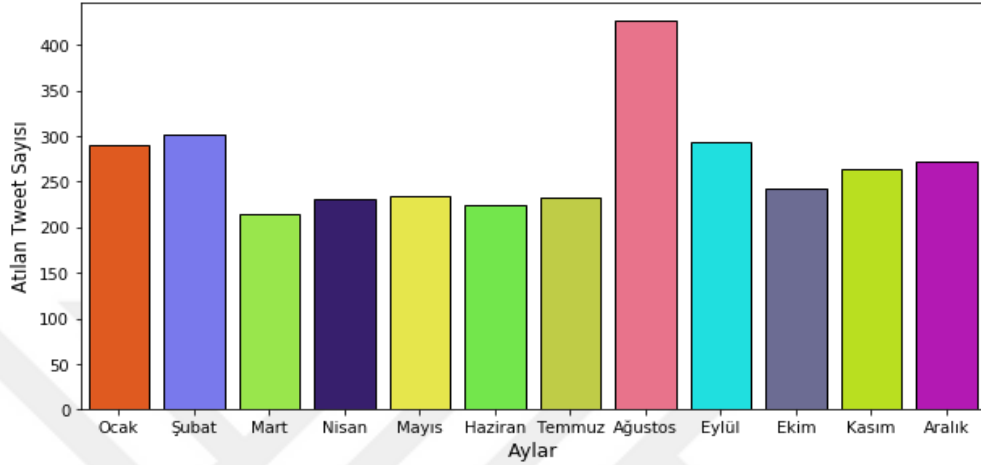
Şekil 3.26. @EmniyetGM adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu

@ EmniyetGM adlı kullanıcının Twitter paylaşımlarının beğenilme ve retweet durumlarının tarih bazlı ortak grafiği şekil 3.26’da gösterilmektedir (Takipçi sayısının tarih bazında sabit olmasından dolayı “followers_count” değeri ile tweet kimliğini içeren “id” değerinde ilişkisel analiz incelenmemiştir.).



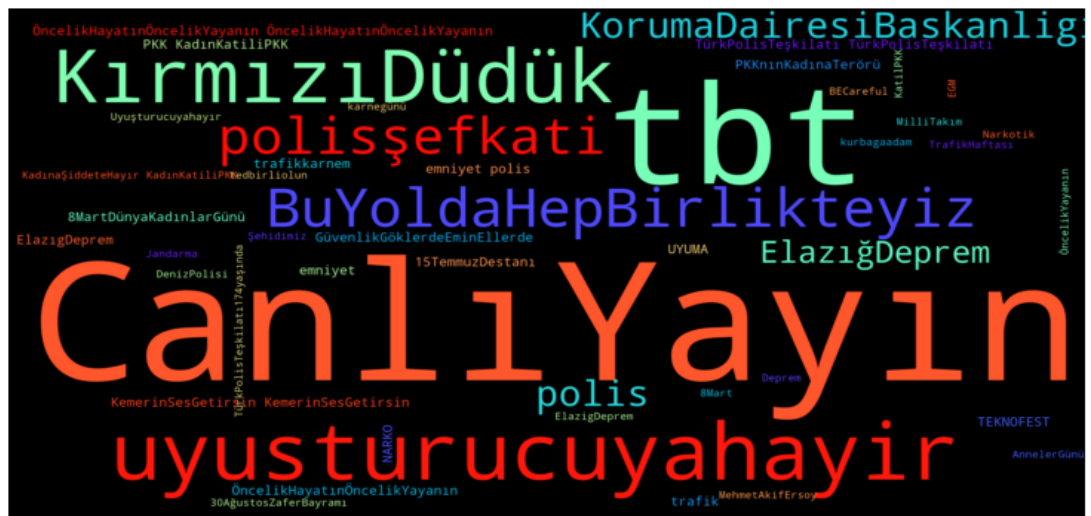
Şekil 3.27. Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı

Şekil 3.27 incelendiğinde aynı gün içerisinde sık paylaşımın yapıldığı tarihler görülmekte olup, örnek olarak en sık paylaşımın yapıldığı 10 Ağustos 2019 tarihindeki 22 adet yapılan paylaşım incelendiğinde, paylaşımın kurban bayramı mesajı, trafik kuralları hatırlatmaları ve bazı tekrar paylaşılan tweet’lerde oluştuğu görülmektedir.

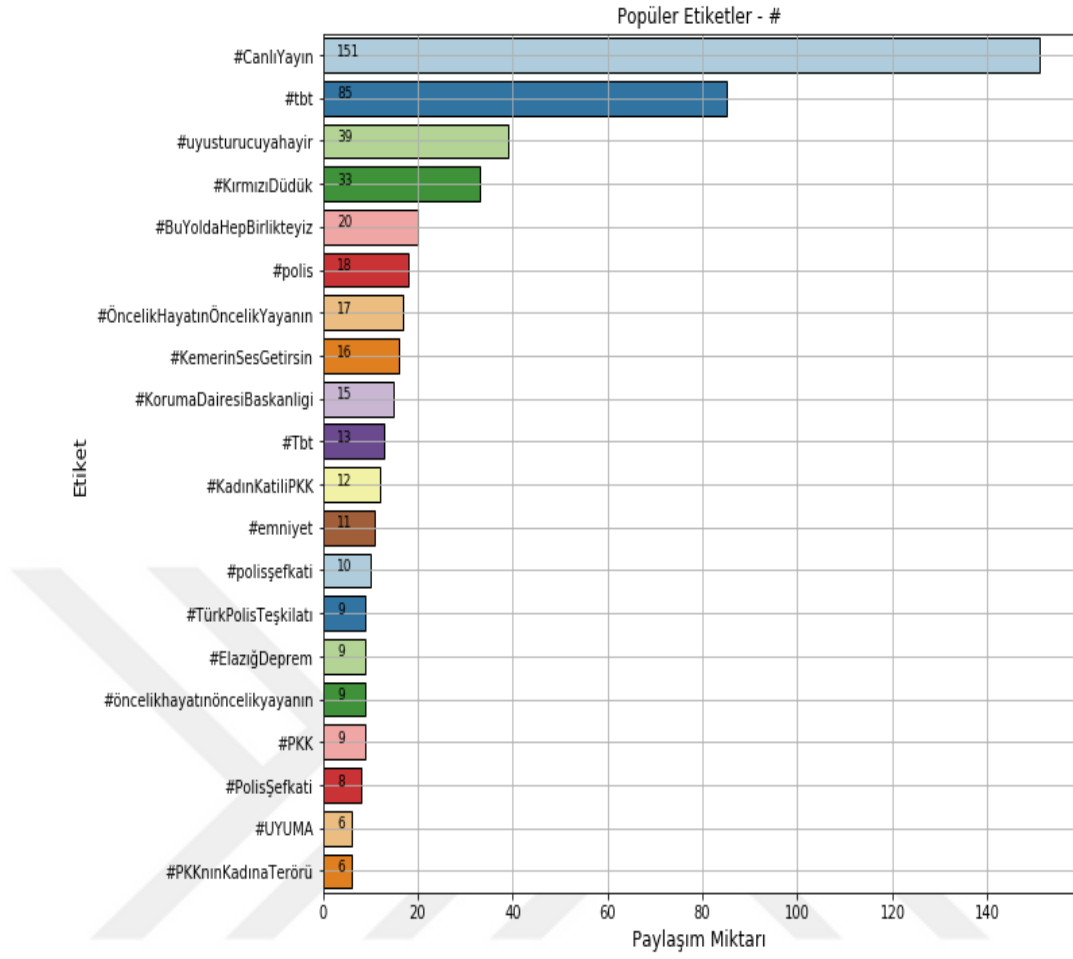


Şekil 3.28. Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı

Twitter aracılığıyla sosyal medyadan yapılan paylaşımlara şekil 3.28’de olduğu üzere aylık dağılımına bakıldığında 447 tweet ile Ağustos ayı en fazla sosyal medya paylaşımının yapıldığı ay iken, genel seyrinde paylaşımların düzenli aralıklarla yapıldığı gözlemlenmektedir.

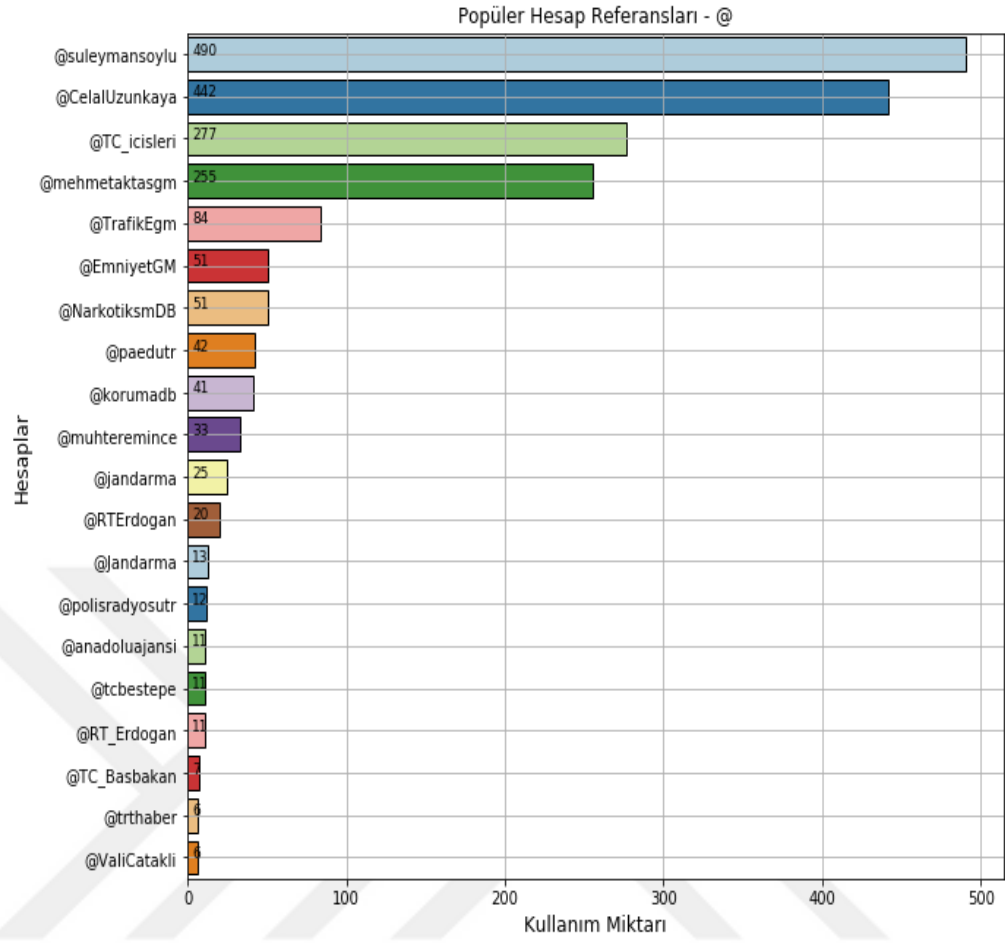


Şekil 3.29. En fazla kullanılan sosyal medya etiketlerinin (Hashtag - #) kelime bulutu



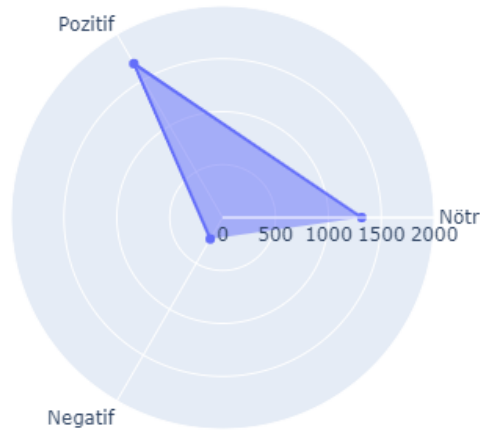
Şekil 3.30. En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)

Yapılan paylaşımlarda en fazla kullanılan sosyal medya etiketlerinin 20'sine ilişkin kullanım sıklığı ile sık kullanılan sosyal medya etiketlerine ilişkin hazırlanan kelime bulutu (Maksimum kelime girdisi 1000 olarak alınmıştır.) şekil 3.29 ve 4.30'da gösterilmektedir.



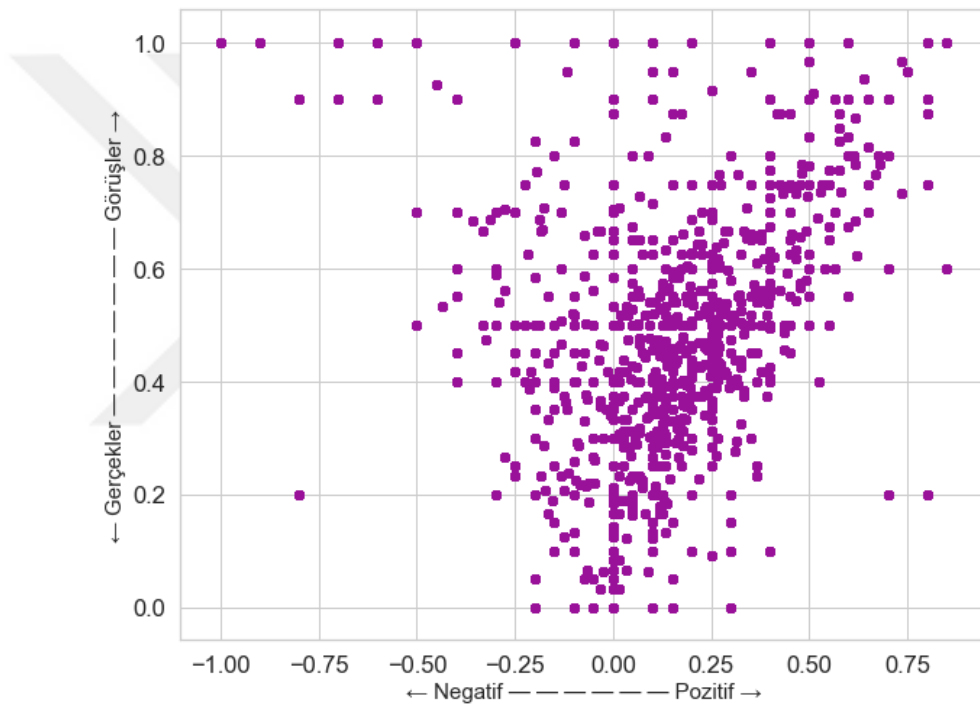
Şekil 3.31. En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)

Yapılan paylaşımlarda en fazla kullanılan sosyal medya hesaplarının 20'sine ilişkin kullanım sıklığı şekil 3.31'de gösterilmektedir.



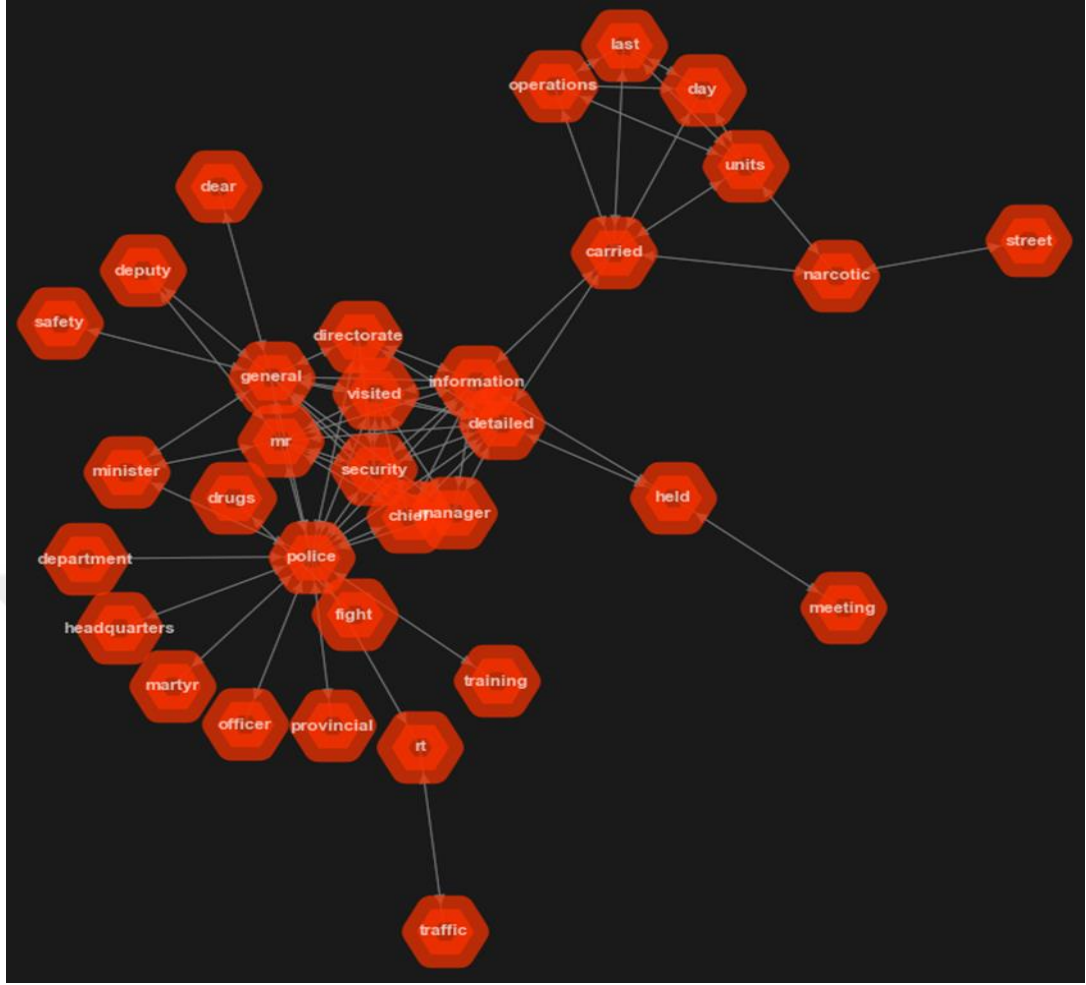
Şekil 3.32. Metnin duygusal durumunun radar grafiği

@EmniyetGM kullanıcı hesabından yapılan paylaşım metinleri, metin ön işlemesine (text preprocessing) tabi tutularak tüm emojiler kaldırılmış ve Türkçe/İngilizce karakter farklılıklarından dolayı metindeki tüm harfler küçültülerek makine öğrenimine dayalı olarak Türkçe'den İngilizce'ye çeviri yapılmıştır. Çeviri metni üzerinde yapılan duygusal dağılım (0=Nötr, 0>Negatif, 0<Pozitif) şekil 3.32'de görüldüğü üzere 1662 tweet ile pozitif mesajların çoğunlukta olduğu, 1304 tweet ile nötr mesajların pozitif mesajlarını izlediği ve son olarak da 262 tweet ile negatif skorlanan mesajların bulunduğu belirlenmiştir.



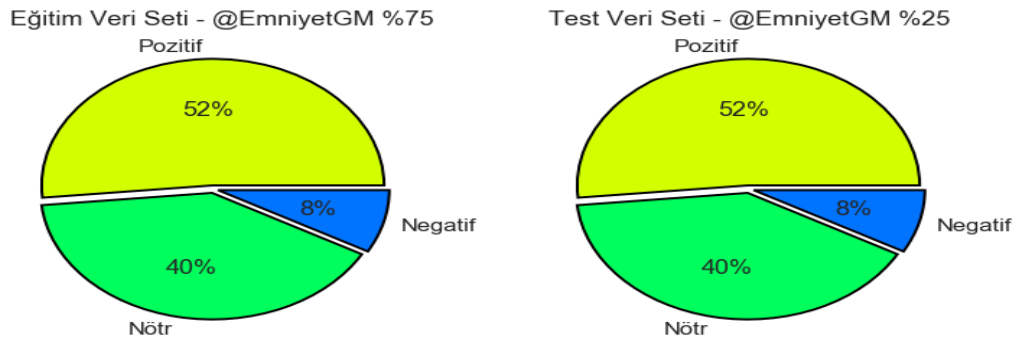
Şekil 3.33. Metnin duygusal durumunun öznellik ile analizi

Elde edilen çeviri üzerinden kelime bazlı olarak, her kelimenin yarattığı duygusal durum ortaya konularak, her cümleye ait duygusal puan durumu, duygusal durumu ve ortaya konulan kanının öznellik ihtimali çıkarılmıştır. Metnin duygusal durumu ile öznelliğini beraber analiz ettiğimizde ise, cümlelerin içinde yer alan her kelimenin skorlanması ile cümle skoru elde edilmiş olup, ana korpusun çoğunluğunun 0.5 öznellik bandını geçmediği şekil 3.33'de gösterilmektedir. Fakat az sayıda da olsa öznelliği 1 olarak yer alan 15 tweet için, negatif, nötr veya pozitif gibi değerlendirmelerde bulunulmuştur.



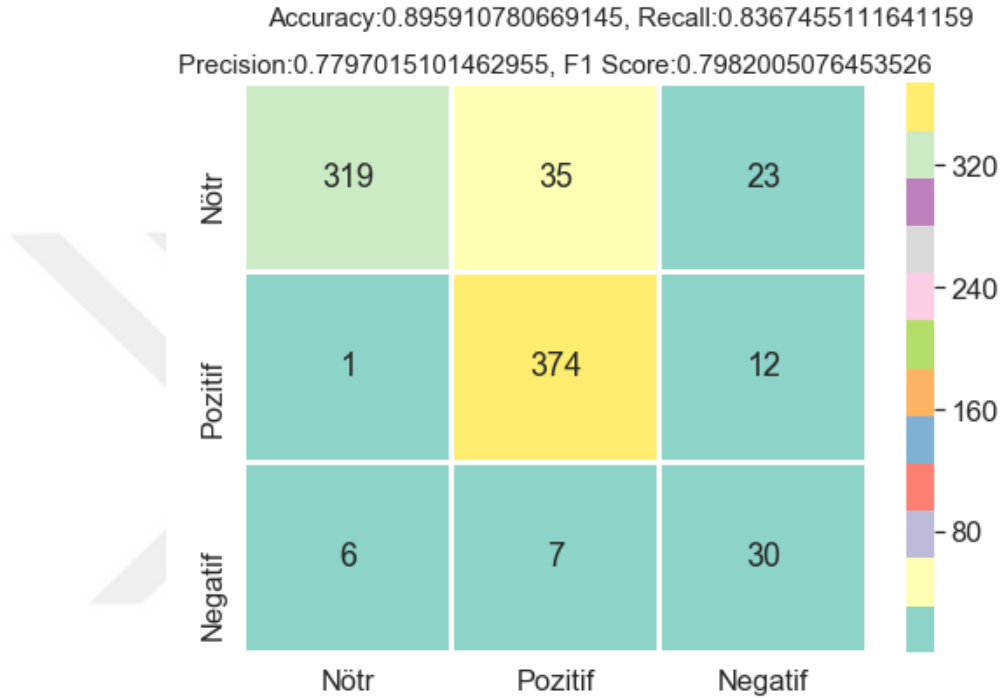
Şekil 3.36. Yapılan paylaşımlara ait kelime ağı analizi (100>Kelime Kombinasyonu)

Yapılan paylaşım metinleri BoW (Bag of words) metoduyla veriseti haline getirilmiştir. Söz konusu veriseti içerisinde de 100’den fazla kombinasyona sahip kelimelerin ağı (network) yapısı şekil 3.36’da gösterilmiştir.



Şekil 3.37. Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması

Textblob kütüphanesi aracılığıyla yapılan paylaşımların duygu durumları belirlenmiş olup, söz konusu tespitlerin kendi içinde tutarlı olup olmadığını incelemek adına makine öğrenimi modellerinden rastgele orman sınıflandırması (Random forest classification) ile veri seti şekil 3.37’de gösterildiği üzere test ve eğitim setine ayrılmıştır. Sınıflandırma için verinin %75’i eğitim, %25’i de test olarak ayrılmıştır.

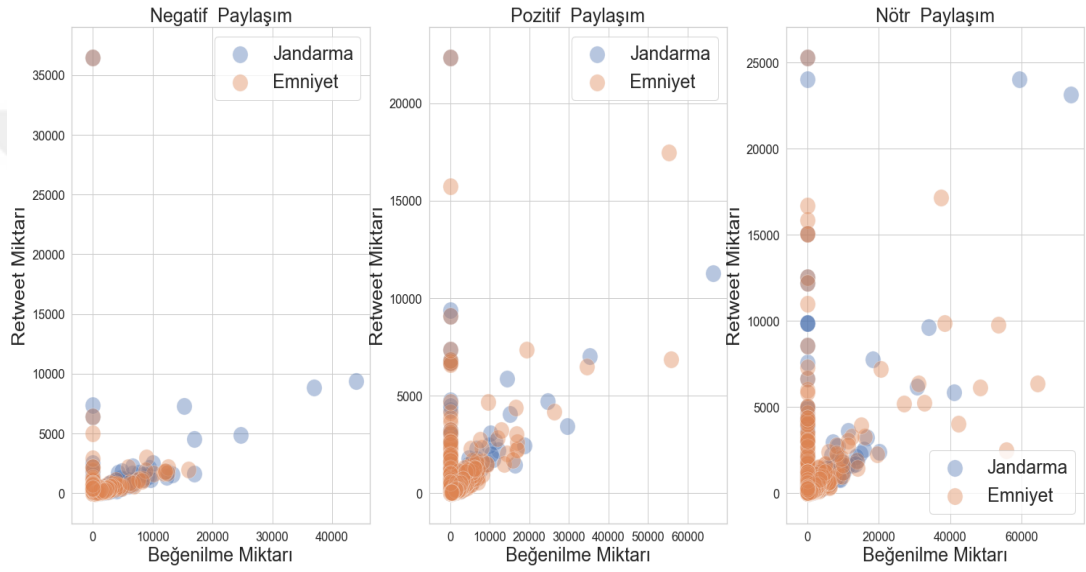


Şekil 3.38. Test verisinin duygusal matriksi

Makine öğrenimi sonrası test verileri üzerinde çıkan sonuca ilişkin matriks şekil 3.38’de görülmekte olup, yapılan öğrenme test verileri üzerinde %89.591 oranında doğruluk oranı sağlamıştır.

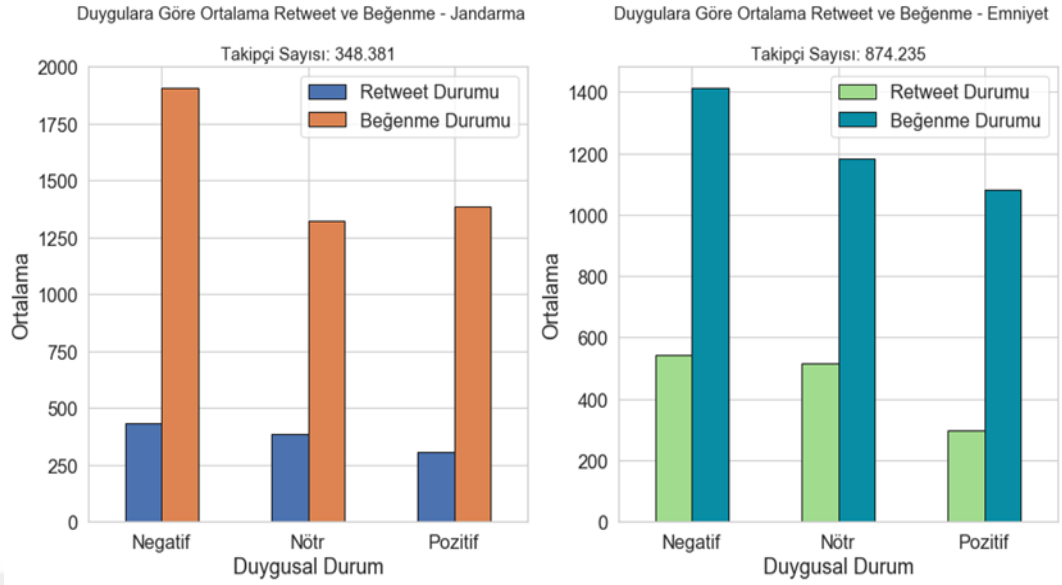
3.1.3. Jandarma ve Emniyet Teşkilatlarının Mukayeseli Ortak Analizi

Ülkemiz özelinde faaliyet gösteren kolluk teşkilatlarının topluma verdiği mesajları, mesajlarda hangi konu ve kelimeleri ağırlıkta kullandığı ile paylaştığı mesajların sosyal medyadaki etkisini mukayeseli olarak aşağıda incelenmektedir. Mukayese kapsamında korelasyon ve makine öğrenimi modellemelerinden karar destek vektörleri kullanılarak, tespitler görsel içerikle desteklenmektedir.



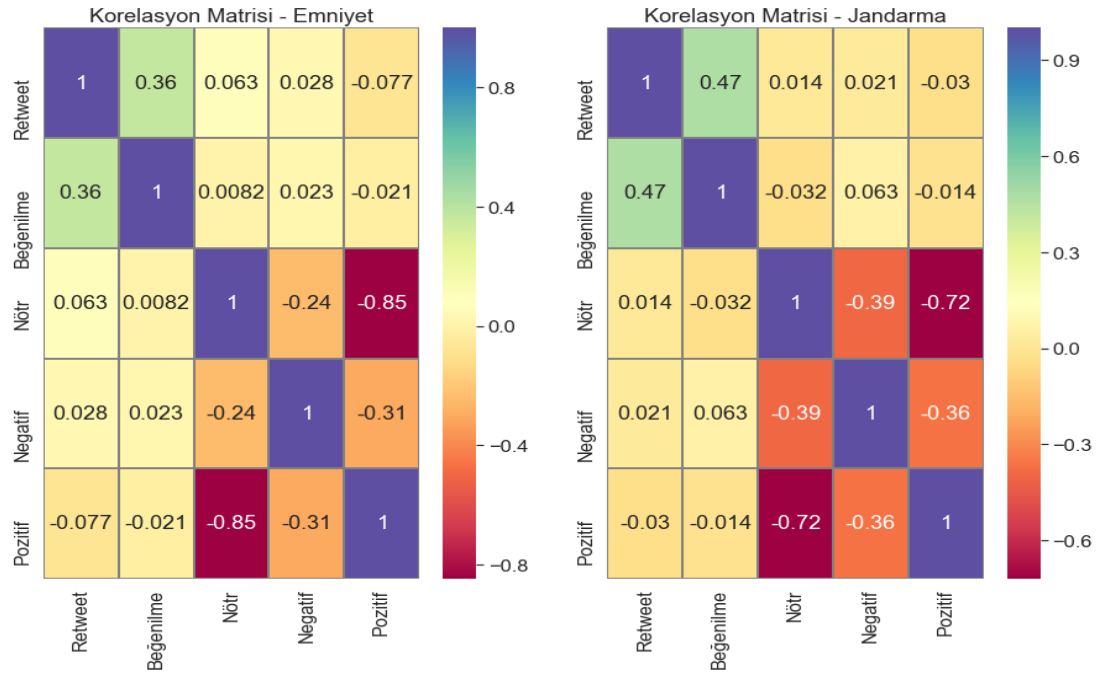
Şekil 3.39. Duygulara göre retweet ve beğenme durumu

Paylaşımlar içerisinde anomali tespiti yapmak amacıyla şekil 3.39 incelendiğinde Jandarma özelinde bazı negatif cümlelerde anomalinin fazla olduğu, Emniyet özelinde ise bazı nötr cümlelerde anomalinin fazla olduğu gözlemlenmiştir. Kurumlar arası genel dağılımın birbirine yakın olduğu ve en fazla anomali gösteren dağılımın nötr paylaşımlarda yaşandığı belirlenmiştir.



Şekil 3.40. Duyulara göre ortalama retweet ve beğenme durumu

Paylaşımların genel ortalaması şekil 3.40’da görüldüğü üzere incelendiğinde, negatif paylaşımlarda Jandarma’nın beğeni ve retweet ortalaması yüksek olup, Emniyet için de negatif paylaşımların daha fazla ortalamaya sahip olduğu gözlemlenmektedir.



Şekil 3.41. Duygusal korelasyon matrisi

Duyguların, beğeni ve retweet üzerindeki gücünün korelasyon biçimindeki gösterimi matriks biçiminde şekil 3.41’de yapılmış olup, söz konusu verilerin analizinden;

– Beğeni ve retweet arası ilişkinin Jandarma’da daha güçlü olduğu (Jandarma: 0.47, Emniyet:0.36),

– Pozitif mesajlar kapsamında yapılan paylaşımları insanların beğenmesin Jandarma’da daha yüksek ilişki düzeyinde olduğu (Jandarma: -0.014, Emniyet: -0.021),

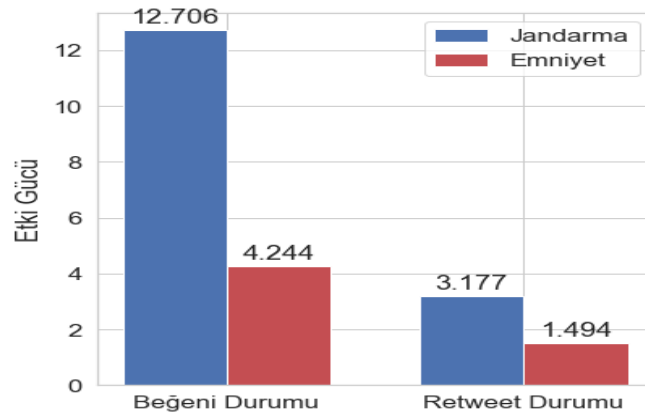
– Pozitif mesajlar kapsamında yapılan paylaşımları insanların retweet etmesinin Jandarma’da daha yüksek ilişki düzeyinde olduğu (Jandarma: -0.03, Emniyet: -0.077),

– Nötr mesajlar kapsamında yapılan paylaşımları insanların beğenmesin Emniyet’de daha yüksek ilişki düzeyinde olduğu (Jandarma: -0.032, Emniyet: 0.0082),

– Nötr mesajlar kapsamında yapılan paylaşımları insanların retweet etmesinin Emniyet’de daha yüksek ilişki düzeyinde olduğu (Jandarma: 0.014, Emniyet: 0.063),

– Negatif mesajlar kapsamında yapılan paylaşımları insanların beğenmesin Jandarma’da daha yüksek ilişki düzeyinde olduğu (Jandarma: 0.063, Emniyet: 0.023),

– Negatif mesajlar kapsamında yapılan paylaşımları insanların retweet etmesinin Emniyet’de daha yüksek ilişki düzeyinde olduğu (Jandarma: 0.021, Emniyet: 0.028) gözlemlenmiştir.



Şekil 3.42. Jandarma ve Emniyet’in etki gücü

Yapılan paylaşımların takipçileri oranında hangi derecede paylaşıldığının ve retweet edildiğinin ortaya konması maksadıyla etki gücü hesaplaması yapılmış, elde edilen sonuçlar da şekil 3.42’de gösterilmektedir. Etki gücü hesaplaması yapılırken, toplam beğeni sayısı/toplam takipçi sayısı ve toplam retweet sayısı/toplam takipçi sayısı formülasyonu kullanmıştır. Buradaki amaç her takipçinin ortalama ne kadar mesajı beğendiği ve retweet ettiği tespitiyle, mesajların ne kadar güçlü olduğunu karşılaştırmalı olarak ortaya koymaktır. Bu yapılan analiz sonucunda ise Jandarma’nın paylaşımlarının Emniyet’e kıyasla etki gücünün daha yüksek olduğu anlaşılmaktadır. Bu sebeple de aradaki farkın sebebini daha iyi anlayabilmek adına makine öğrenimine dayalı paylaşım yaptıkları konuların neler olduğunun ortaya konması gerekmektedir.



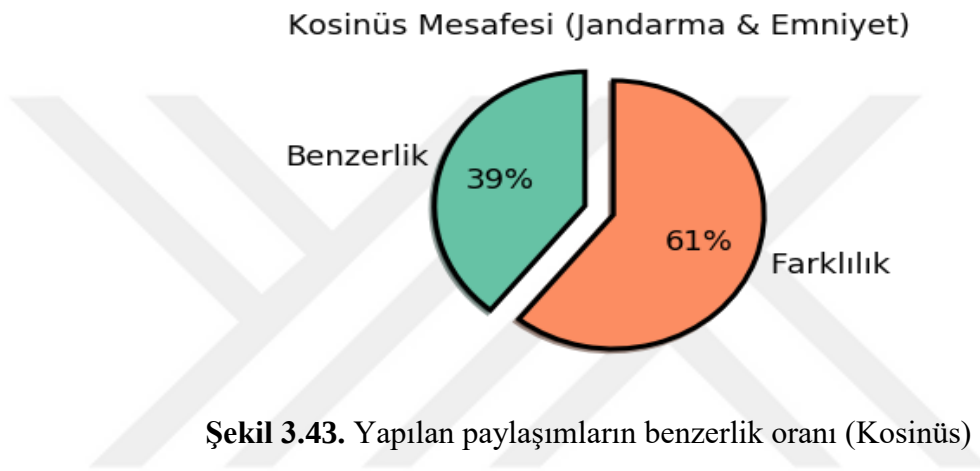
Çizelge 3.1. Jandarma ve Emniyet’in paylaşım yaptığı konuların LDA (Latent Dirichlet Allocation) analizi ve konular arası mesafe haritalandırması

Konu	Jandarma LDA Analizi	Emniyet LDA Analizi
Mesafe Haritası		
1	('operation', 629.4017413478667) ('carry', 324.47761853049184) ('terrorist', 290.9672874677739) ('command', 280.6447539946258) ('provincial', 237.4469583936999) ('gendarmerie', 152.4245373988662) ('countryside', 150.72540772378136) ('neutralize', 139.65897227848723) ('seize', 112.96306028531048) ('weapon', 110.559917276184)	('general', 766.8718240067469) ('police', 666.4975162413915) ('visit', 456.013516064948) ('security', 428.81730617072765) ('chief', 395.40378994704776) ('manager', 390.7393630950251) ('information', 323.718419647185) ('directorate', 211.6131391360382) ('dear', 158.4672949368882) ('minister', 133.07428458414483)
2	('gendarmerie', 458.0578166291945) ('command', 313.198223554641) ('training', 189.37738438053842) ('commando', 168.30272349350977) ('school', 147.42058785410504) ('center', 139.06132565986545) ('complete', 101.63748601878648) ('general', 98.08521199459211) ('day', 88.69420861101929) ('successfully', 86.07839034723983)	('hold', 206.93769927778914) ('minister', 199.12765672409915) ('information', 185.1072672893433) ('meeting', 156.11281411203913) ('ceremony', 120.59063439964558) ('police', 104.3031557107045) ('special', 89.65705573677174) ('interior', 88.79600741153342) ('operation', 84.65720911589163) ('coordination', 80.434112990089)
3	('suspect', 265.7470256299867) ('search', 154.28788522491178) ('command', 153.2151022623189) ('carry', 127.67052809645838) ('provincial', 127.64642125515802) ('operation', 119.18455886374711) ('seize', 118.97433268619753) ('gendarmerie', 108.78818685048931) ('district', 104.78712970182904) ('catch', 91.85482660149448)	('traffic', 346.0617651195887) ('accident', 88.10358932422338) ('driver', 71.21764762477481) ('control', 64.5187900749394) ('speed', 58.26732533406349) ('vehicle', 57.769851575457324) ('safe', 57.60478970851868) ('organization', 55.93673604039453) ('rule', 50.55169273718484) ('holiday', 48.81069162215694)
4	('suspect', 129.59433510624135) ('drug', 123.09980725679087) ('seize', 121.27690036671616) ('search', 114.74948218485663) ('person', 110.94255805633998) ('make', 106.42337873756954) ('vehicle', 97.23951171992388) ('application', 93.92753553133373) ('gendarmerie', 70.30869871722852) ('istanbul', 54.96992469183962)	('information', 219.3526428428303) ('carry', 138.7325138626602) ('country', 135.96717673761623) ('detailed', 82.76464625190191) ('inspection', 79.13404974650166) ('application', 66.96957509671591) ('child', 53.912082751726295) ('simultaneously', 53.67377017716502) ('people', 48.45146110132297) ('personnel', 43.20558761799919)

Çizelge 3.1. Devam

Konu	Jandarma LDA Analizi	Emniyet LDA Analizi
5	('gendarmerie', 315.0615461039764) ('general', 297.2484263439944) ('visit', 196.6546748724493) ('commander', 187.7315583260353) ('command', 142.56630098435994) ('provincial', 107.77011343641679) ('minister', 98.25650686609087) ('arif', 92.64043697074375) ('çetin', 84.60806321002833) ('martyr', 67.3322056268727)	('operation', 238.80375304769433) ('day', 186.31292828255525) ('unit', 184.91723915965355) ('drug', 164.10974513318882) ('narcotic', 155.57941206812694) ('fight', 141.94913704052354) ('carry', 139.25523291773112) ('street', 131.89669005835202) ('determine', 90.25493008498609) ('crime', 77.67718931509)
6	('gendarmerie', 263.8165101116152) ('command', 153.6507168734601) ('guard', 87.55479627353488) ('ceremony', 81.53092520534187) ('commander', 74.53227264885984) ('soldier', 72.18486430728407) ('provincial', 70.28620370431236) ('coast', 65.42744156887011) ('hold', 63.577396679099145) ('province', 59.45591629915414)	('police', 342.12924263469006) ('year', 100.65928373524845) ('nation', 78.77066570489946) ('istanbul', 77.50834074206861) ('officer', 71.52458746272933) ('thanks', 65.66049465241674) ('training', 53.57388013010187) ('student', 52.06970393501228) ('woman', 49.03444282564274) ('center', 44.565996546593496)
7	('traffic', 117.04814528179591) ('winter', 44.73558465267981) ('safe', 44.32923528582589) ('country', 41.76110415535589) ('citizen', 40.97803799746624) ('let', 39.199661083929) ('helicopter', 38.07683927129698) ('come', 34.92032377401274) ('van', 32.76794431817262) ('vehicle', 32.7148912966662)	('ankara', 102.31783381736582) ('anniversary', 97.61384713290242) ('martyr', 68.66860369881603) ('life', 46.26436163530776) ('love', 45.42594673012199) ('mustafa', 45.01212164925333) ('veteran', 44.996606772996664) ('kemal', 37.986960658804655) ('respect', 33.02985306996796) ('road', 30.754866830330403)
8	('student', 95.16347949941073) ('school', 90.1507601077504) ('happy', 74.1192107358344) ('world', 73.8255560070935) ('national', 66.07578207196474) ('brother', 65.574561667317) ('championship', 49.95227634723123) ('congratulation', 46.78489431596617) ('woman', 44.453116615700246) ('primary', 41.9054608815404)	('peace', 60.21708101837325) ('citizen', 49.08772726986628) ('use', 48.27688842066088) ('security', 48.10063830978789) ('duty', 48.02501074076333) ('today', 47.46721117755672) ('city', 43.38182282940477) ('team', 40.12190418167899) ('belt', 36.382783709395675) ('agency', 32.6394875280704)
9	('nation', 64.74425140415242) ('driver', 61.398247961514095) ('love', 52.19513209649139) ('holiday', 51.12636522828892) ('wish', 50.68159055929737) ('life', 44.283265418018814) ('accident', 36.06304699461793) ('live', 35.94804697474414) ('time', 35.47197882882861) ('future', 35.217880601114224)	('gendarmerie', 91.94570124854715) ('family', 91.15922965709738) ('child', 81.71420237594477) ('headquarters', 63.22761662581216) ('pedestrian', 58.93403165250339) ('command', 53.96616427399204) ('happy', 47.32781413936531) ('occasion', 37.410580576381285) ('commander', 36.66607032101102) ('mardin', 31.507692240026678)
10	('anniversary', 101.91551369527551) ('veteran', 97.7237410622612) ('martyr', 96.50012302180562) ('loss', 87.66819754448103) ('sorry', 80.36709183202143) ('hero', 64.89865153471108) ('year', 49.433405220557155) ('mustafa', 48.71123179359868) ('homeland', 47.62351114223641) ('kemal', 46.926669036365276)	('guard', 63.863374531726095) ('world', 52.900711385661694) ('news', 35.390937455297745) ('program', 34.806923406122) ('peaceful', 31.730539837083573) ('seize', 31.552490065839695) ('neighborhood', 30.651788942197616) ('june', 28.777177251693363) ('bazaar', 28.589051821752886) ('support', 27.552539950238952)

Çizelge 3.1’de konu modellemeye aktif olarak kullanılan LDA modellemesiyle, kurumların paylaşımları 10 konu başlığı altında incelenmiştir. 10 konu içinde de her konudan en yüksek ilişki düzeyine sahip 10 kelime örneklem olarak çizelgede gösterilmiştir. Her konunun vektörel düzeyde konular arası mesafe haritası çizelgenin başında verilmiş olup, kelimelerin yanında yer alan ifadeler de vektörel düzeyde konu içinde ilişkisini ifade etmektedir. Bu kapsamda konu bazında yapılan incelemelerde, paylaşım yapılan konuların ve vektörel düzlemdeki mesafelerinin ayrıştığı rahatlıkla görülmektedir.



Paylaşım yapılan konuların vektörel dağılımlarındaki farklılıkların, metin korpuslarının geneline yayılıp tespit edilmesi maksadıyla, öncelikle daha net bir sonucun çıkması için kelimelerin kökleri çözümlenerek (lemmatization) işlem yapılmıştır. Metin içindeki kelime (token) yapılarının her biri vektörel düzeye indirgenerek, şekil 3.43’de görüldüğü üzere kosinüs mesafesi yöntemi ile paylaşım yapılan metinlerin benzerlik oranı %39 olarak bulunmuştur.

3.2. İtalya Özelinde Kolluk Teşkilatlarının Paylaşımlarının İncelenmesi

Arma dei Carabinieri ve Polizia di Stato’nun resmi paylaşımlarından alınan örneklem üzerinden keşfedici veri analizi yapılarak, sosyal medya kullanım alışkanlıkları üzerinden duyarlılık analizi ve söz konusu duyarlılığın makine öğrenimine dayalı tahminlemesi yapılmıştır.

3.2.1. Arma dei Carabinieri'nin Paylaşımlarının İncelenmesi

İtalya'da faaliyet gösteren jandarma türü kolluk gücünün resmi hesabı olan “Arma dei Carabinieri” başlıklı ve “@_Carabinieri_” kullanıcı adlı Twitter hesabından 09 Mart 2020 tarihinde herkese açık olarak yapılan paylaşımlar Tweepy kütüphanesi aracılığıyla Python programlama dili kullanılarak, csv formatında çekilmiştir.



Şekil 3.44. Arma dei Carabinieri resmi Twitter sayfasının ekran resmi

	id	created_at	text	followers_count	retweet_count	favourite_count	source
0	1236909764841529344	2020-03-09 07:00:43	Buongiorno e buona settimana Fondo Lago Smeral...	90493	8	45	Twitter Web App
1	1236728628265783297	2020-03-08 19:00:56	RT @MinisteroSalute: #covid19Italia a tutte le...	90493	235	0	Twitter for iPhone
2	1236715573456797696	2020-03-08 18:09:04	RT @MinisteroSalute: #coronavirusitaliainL'a...	90493	124	0	Twitter for iPhone
3	1236709009710813191	2020-03-08 17:42:59	RT @DPCgov: 🚫8marzo ore 18 L'aggiornamento qu...	90493	76	0	Twitter for iPhone
4	1236683752480428033	2020-03-08 16:02:37	RT @MinisteroSalute: Covid-19, nuove regole: e...	90493	374	0	Twitter for iPhone

Şekil 3.45. “@_Carabinieri_” kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler – I

```
print(data.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3232 entries, 0 to 3231
Data columns (total 7 columns):
id                3232 non-null int64
created_at        3232 non-null object
text              3232 non-null object
followers_count   3232 non-null int64
retweet_count     3232 non-null int64
favourite_count   3232 non-null int64
source            3232 non-null object
dtypes: int64(4), object(3)
memory usage: 176.9+ KB
None
```

Şekil 3.46. “@_Carabinieri_” kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler –II

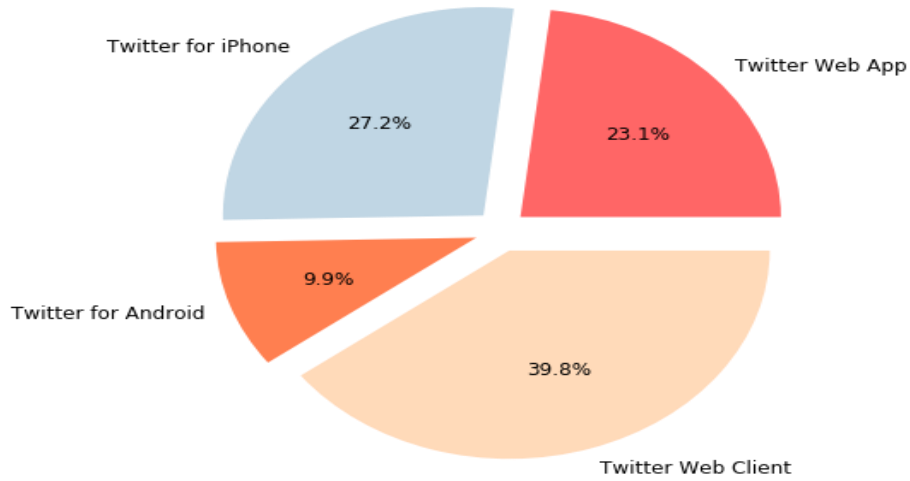
```
print(f" Data Available since {data.created_at.min()}")
print(f" Data Available upto {data.created_at.max()}")

Data Available since 2017-12-21 18:43:32
Data Available upto 2020-03-09 07:00:43
```

Şekil 3.47. “@_Carabinieri_” kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler –III

Pandas kütüphanesi aracılığıyla “@_Carabinieri_” kullanıcı adlı hesaptan çekilen veriler üzerinde ön inceleme yapıldığında;

- Verinin 3232 satır ve 7 sütundan oluştuğu (Şekil 3.45, Şekil 3.46),
- Verinin sütun isimlerinin “id”, “created_at”, “text”, “followers_count”, “retweet_count”, “favourite_count” ve “source” olduğu (Şekil 3.46),
- Veri tipleri olarak “created_at”, “text” ve “source” sütunlarının object (nesne), diğer sütunların ise int64 (sayısal - integer) olarak tanımlandığı (Şekil 3.46),
- Veride 0' dan 3231' e kadar numaralandırılmış olan 3232 adet girdi bulunduğu ve her bir sütunda boş olmayan (non-null) girdi sayısının tanımlandığı, bu değerlerden de değeri olmayan sütun değerinin bulunmadığının anlaşıldığı (Şekil 3.46),
- “@_Carabinieri_” adlı kullanıcıdan çekilen verilerin 21 Aralık 2017’den 09 Mart 2020 tarihleri arasında yer aldığı (Şekil 3.47) anlaşılmaktadır.



Şekil 3.48. “@_Carabinieri_” adlı kullanıcının veri paylaşımında kullandığı kaynaklar

“@_Carabinieri_” adlı kullanıcının Twitter paylaşımı yaptığı cihazların kaynağı incelendiğinde; %27,2’sinin Iphone, %9,9’unun Android, %39,8’inin Twitter web istemcisi (web client) ve %23,1’inin Twitter web uygulaması aracılığıyla olduğu şekil 3.48’de gösterilmektedir.

```
print(f" Maximum number of retweets {data.retweet_count.max()}")
print(f" Maximum number of favorites {data.favourite_count.max()}")

Maximum number of retweets 6524
Maximum number of favorites 15744

data.loc[data['retweet_count']==6524, ['text']].values
array([[ 'RT @Quirinale: #coronavirus: dichiarazione del Presidente #Mattarella https://
dtype=object)

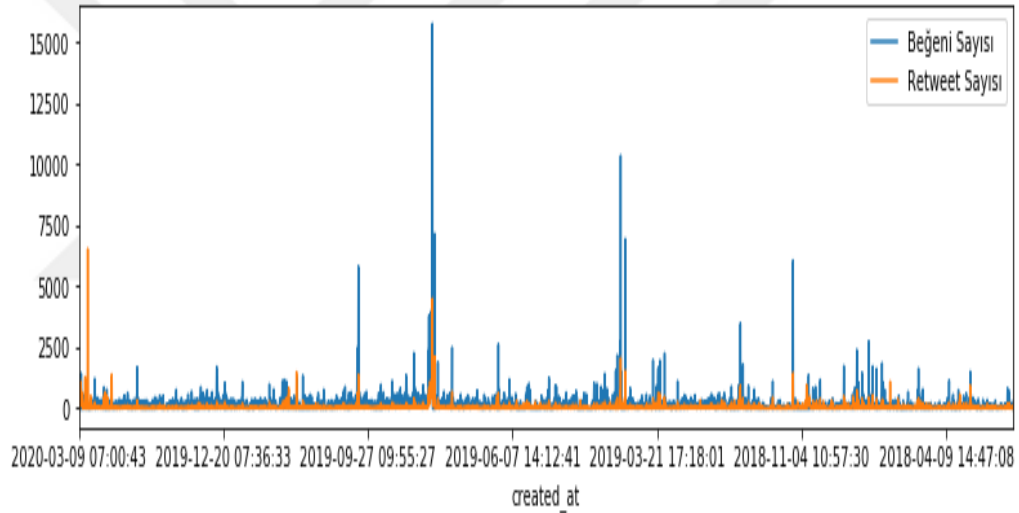
data.loc[data['retweet_count']==6524, ['created_at']].values
array([[ '2020-03-05 18:01:08']], dtype=object)

data.loc[data['favourite_count']==15744, ['text']].values
array([[ "Volanti della @poliziadistato, a sirene spiegate, davanti al Comando Generale
dtype=object)

data.loc[data['favourite_count']==15744, ['created_at']].values
array([[ '2019-07-26 11:50:39']], dtype=object)
```

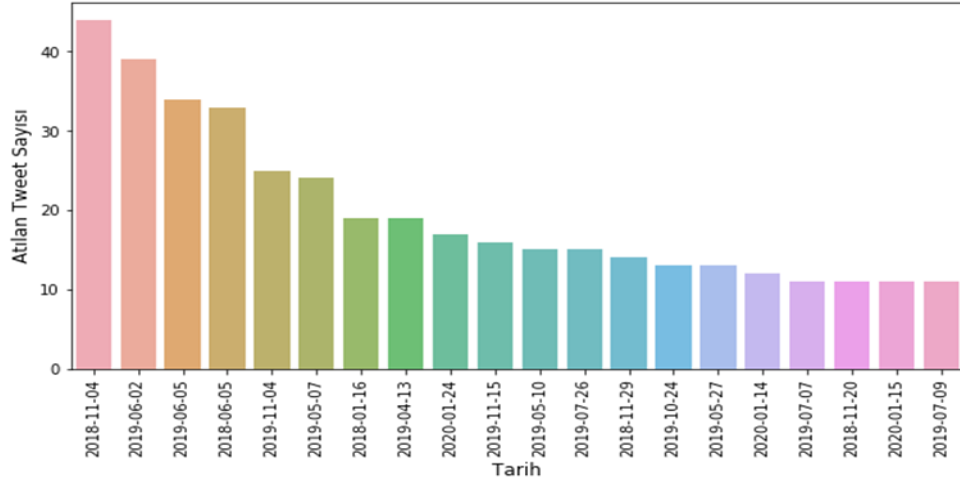
Şekil 3.49. Paylaşımların beğenilme ve tekrar paylaşılma durumunun incelenmesi

Yapılan paylaşımlar şekil 3.49’da görüldüğü üzere incelendiğinde, 6524 tekrar paylaşım oranı ile en fazla retweet yapılan paylaşımın 05 Mart 2020 tarihinde paylaşılan yine bir retweet mesajı ('RT @Quirinale: #coronavirus: dichiarazione del Presidente #Mattarella <https://t.co/QJanArqBhw>') olduğu, 73785 beğeni ile de en fazla beğeni alan paylaşımın ("Volanti della @poliziadistato, a sirene spiegate, davanti al Comando Generale dei #Carabinieri, per solidarietà all'Arma che oggi perde uno dei suoi uomini. Un momento pieno di emozione che ci sembra giusto condividere, ringraziando i colleghi della Polizia per il gesto. <https://t.co/FvnuqKiq9O>") 26 Temmuz 2019 tarihinde yapıldığı anlaşılmaktadır. Ayrıca Twitter paylaşımlarının 631.986 kez beğenildiği (favorites), 211.833 kez tekrar paylaşıldığı (retweet) ve toplam takipçi sayısının ise 90.493 olduğu belirlenmiştir.



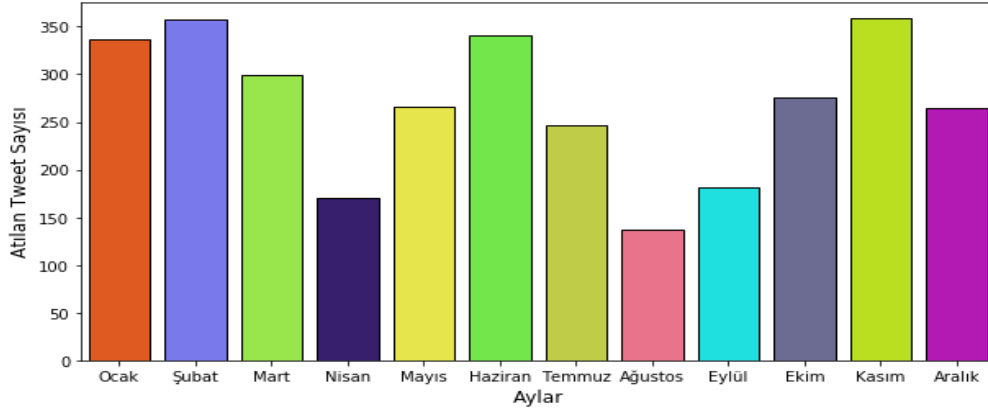
Şekil 3.50. “@_Carabinieri_” adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu

“@_Carabinieri_” adlı kullanıcının twitter paylaşımlarının beğenilme ve retweet durumlarının tarih bazlı ortak grafiği şekil 3.50’de gösterilmektedir (Takipçi sayısının tarih bazında sabit olmasından dolayı “followers_count” değeri ile tweet kimliğini içeren “id” değerinde ilişkisel analiz incelenmemiştir.).



Şekil 3.51. Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı

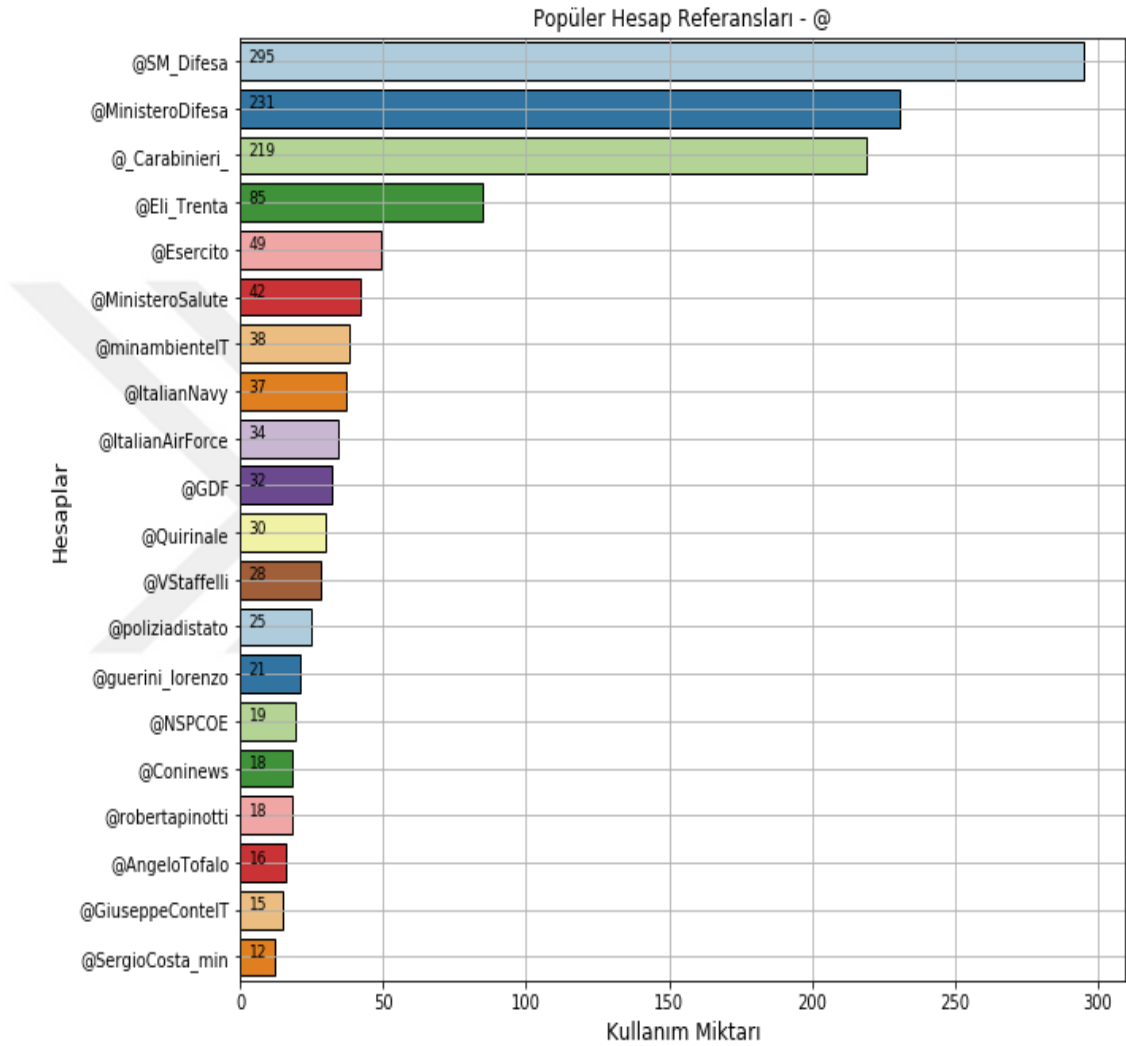
Şekil 3.51 incelendiğinde aynı gün içerisinde sık paylaşımın yapıldığı tarihler görülmekte olup, örnek olarak en sık paylaşımın yapıldığı 04 Kasım 2018 tarihindeki 44 adet yapılan paylaşım incelendiğinde, paylaşımların İtalya’nın “Ulusal Birlik Günü”(04 Kasım) ve “Silahlı Kuvvetler Günü” (04 Kasım) kapsamında mesajlar ile bazı tekrar paylaşılan tweet’lerden oluştuğu görülmektedir.



Şekil 3.52. Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı

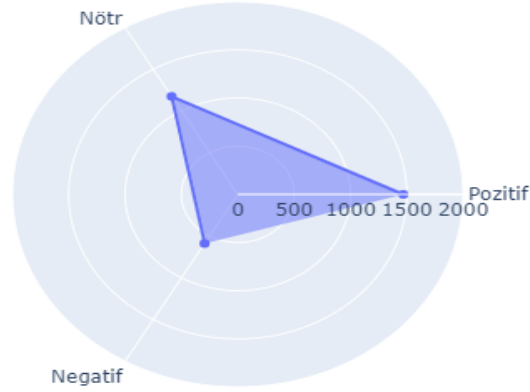
Twitter aracılığıyla sosyal medyadan yapılan paylaşımlara şekil 3.52’de olduğu üzere aylık dağılımına bakıldığında 358 Tweet ile kasım ayı ile 357 Tweet ile şubat ayı en fazla sosyal medya paylaşımının yapıldığı aylar iken, genel seyirde paylaşımların düzenli aralıklarla yapıldığı gözlemlenmektedir.

Yapılan paylaşımlarda en fazla kullanılan sosyal medya etiketlerinin 20'sine ilişkin kullanım sıklığı ile sık kullanılan sosyal medya etiketlerine ilişkin hazırlanan kelime bulutu (Maksimum kelime girdisi 1000 olarak alınmıştır.) şekil 3.53 ve 3.54'de gösterilmektedir.



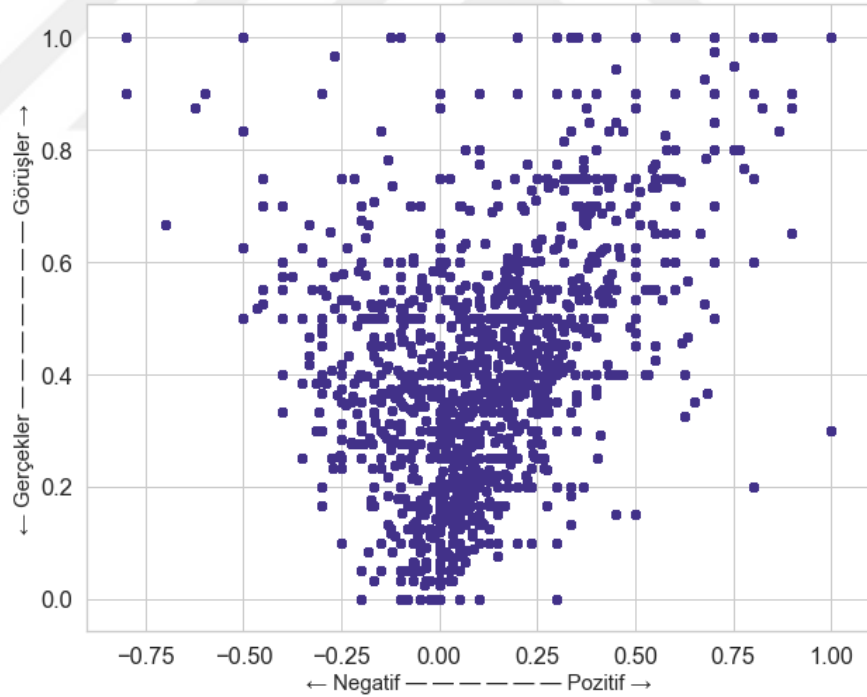
Şekil 3.55. En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)

Yapılan paylaşımlarda en fazla kullanılan sosyal medya hesaplarının 20'sine ilişkin kullanım sıklığı şekil 3.55'de gösterilmektedir. Ayrıca “@_Carabinieri_” kullanıcı hesabından yapılan paylaşım metinleri, metin ön işlemesine (text preprocessing) tabi tutularak tüm emojiler kaldırılmış ve İtalyanca/İngilizce karakter farklılıklarından dolayı metindeki tüm harfler küçültülerek makine öğrenimine dayalı olarak İtalyanca'dan İngilizce'ye çeviri yapılmıştır.



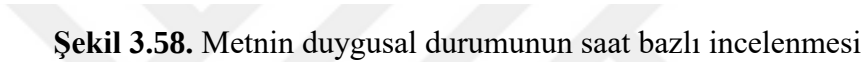
Şekil 3.56. Metnin duygusal durumunun radar grafiği

Çeviri metni üzerinde yapılan duygusal dağılım ($0=Nötr$, $0>Negatif$, $0<Pozitif$) şekil 3.56’da görüldüğü üzere 1491 tweet ile pozitif mesajların çoğunlukta olduğu, 1153 tweet ile nötr mesajların pozitif mesajları izlediği ve son olarak da 588 tweet ile negatif skorlanan mesajların bulunduğu belirlenmiştir.

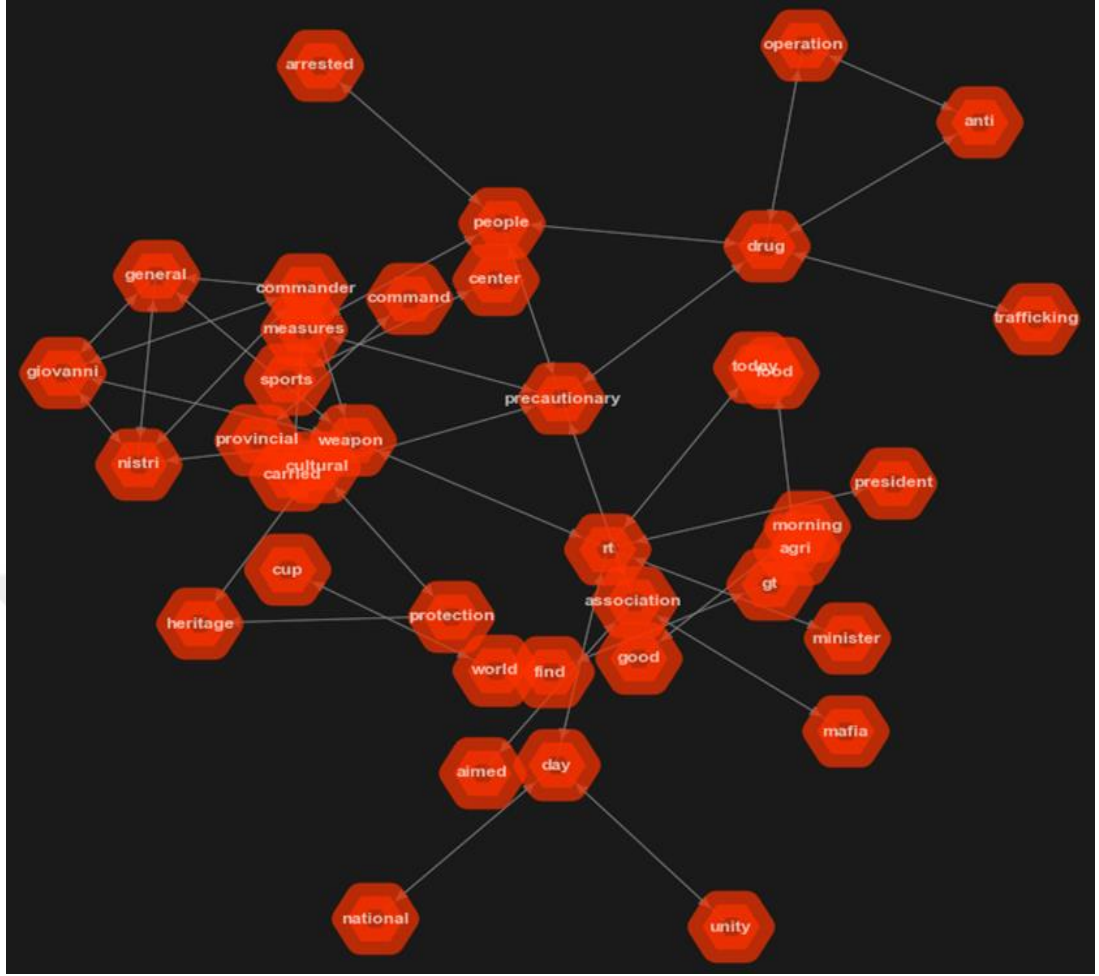


Şekil 3.57. Metnin duygusal durumunun öznellik analizi

Elde edilen çeviri üzerinden kelime bazlı olarak, her kelimenin yarattığı duygusal durum ortaya konularak, her cümleye ait duygusal puan durumu, duygusal durumu ve ortaya konulan kanının öznellik ihtimali çıkarılmıştır. Metnin duygusal

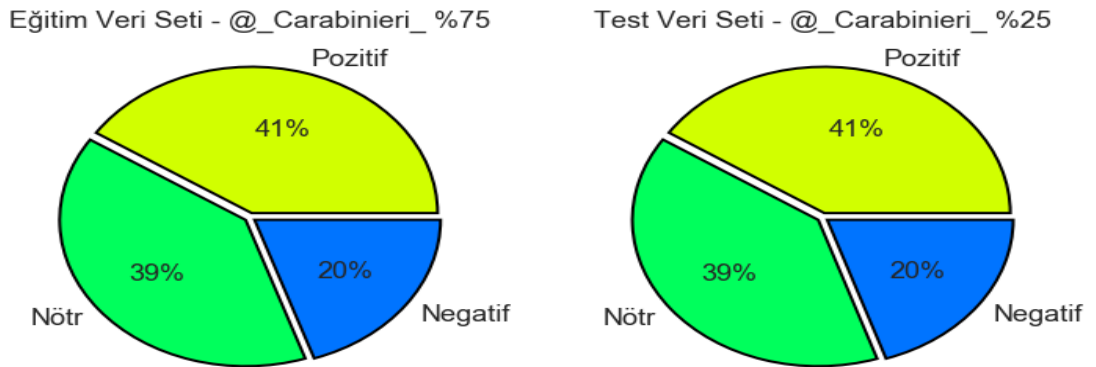
[illegible]

Yapılan paylaşımlarda en sık kullanılan ifadeler bazında (# ve @ gibi ifadeler olmaksızın) şekil 3.59'daki kelime bulutunun elde edildiği görülmektedir.



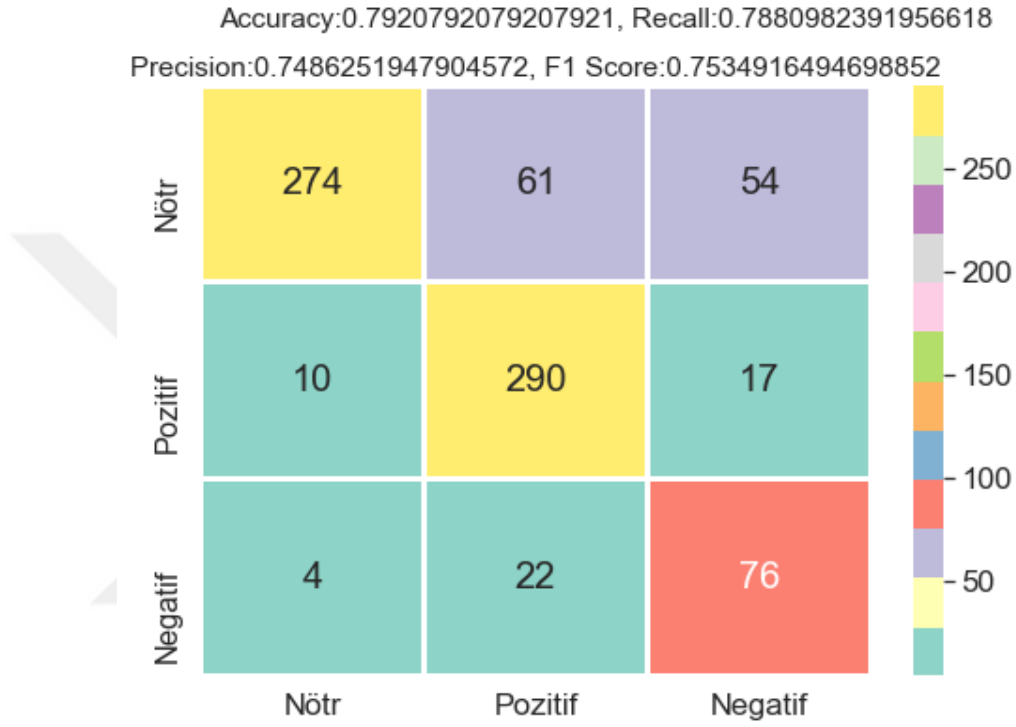
Şekil 3.60. Yapılan paylaşımlara ait kelime ağı analizi (50>Kelime Kombinasyonu)

Yapılan paylaşım metinleri BoW (Bag of words) metoduyla veri seti haline getirilmiştir. Söz konusu veri seti içerisinde de 50’den fazla kombinasyona sahip kelimelerin ağ (network) yapısı şekil 3.60’da gösterilmiştir.



Şekil 3.61. Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması

Textblob kütüphanesi aracılığıyla yapılan paylaşımların duygu durumları belirlenmiş olup, söz konusu tespitlerin kendi içinde tutarlı olup olmadığını incelemek adına makine öğrenimi modellerinden rastgele orman sınıflandırması (Random forest classification) ile veri seti şekil 3.61’de gösterildiği üzere test ve eğitim setine ayrılmıştır. Sınıflandırma için verinin %75’i eğitim, %25’i de test olarak ayrılmıştır.



Şekil 3.62. Test verisinin duygusal matriksi

Makine öğrenimi sonrası test verileri üzerinde çıkan sonuca ilişkin matriks şekil 3.62’de görülmekte olup, yapılan öğrenme test verileri üzerinde %79,208 oranında doğruluk oranı sağlamıştır.

3.2.2. Polizia di Stato’nun Paylaşımlarının İncelenmesi

İtalya’da faaliyet gösteren polis türü kolluk gücünün resmi hesabı olan “Polizia di Stato” başlıklı ve @poliziadistato kullanıcı adlı Twitter hesabından 09 Mart 2020 tarihinde herkese açık olarak yapılan paylaşımlar Tweepy kütüphanesi aracılığıyla Python programlama dili kullanılarak, csv formatında çekilmiştir.



Şekil 3.63. Polizia di Stato resmi Twitter sayfasının ekran resmi

	id	created_at	text	followers_count	retweet_count	favourite_count	source
0	1236661269782036480	2020-03-08 14:33:17	Le misure eccezionali adottate da @Palazzo_C...	122898	243	336	Twitter for Android
1	1236609685006028801	2020-03-08 11:08:18	RT @Quirinale: Presidente #Mattarella: #8marzo...	122898	1100	0	Twitter for Android
2	1236582584345395201	2020-03-08 09:20:37	#8marzo\nTanti auguri a tutte le donne, in p...	122898	145	717	Twitter for Android
3	1236547433150189568	2020-03-08 07:00:56	#coronavirusitalia \nLe importanti misure a...	122898	78	205	Twitter for Android
4	123632882432753665	2020-03-07 16:08:39	Orizzonti e #meraviglieditalia al crepuscolo c...	122898	18	81	Twitter for Android

Şekil 3.64. @poliziadistato kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler – I

```
print(data.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3246 entries, 0 to 3245
Data columns (total 7 columns):
id                3246 non-null int64
created_at        3246 non-null object
text              3246 non-null object
followers_count   3246 non-null int64
retweet_count     3246 non-null int64
favourite_count   3246 non-null int64
source            3246 non-null object
dtypes: int64(4), object(3)
memory usage: 177.6+ KB
None
```

Şekil 3.65. @poliziadistato kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler – II

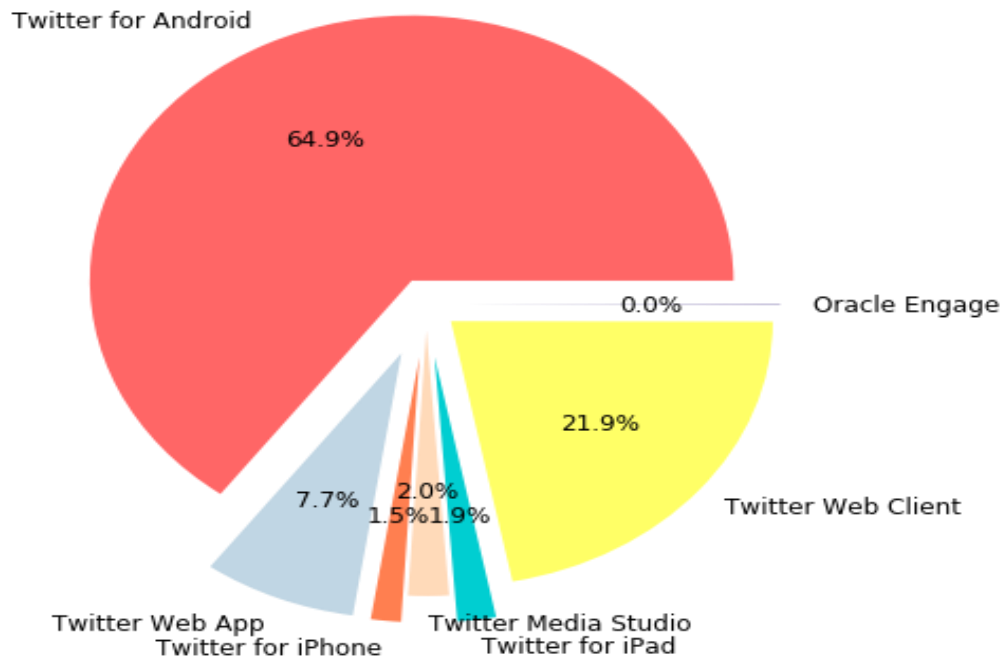
```
print(f" Data Available since {data.created_at.min()}")
print(f" Data Available upto {data.created_at.max()}")
```

Data Available since 2018-03-26 15:46:40
Data Available upto 2020-03-08 14:33:17

Şekil 3.66. @poliziadistato kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler – III

Pandas kütüphanesi aracılığıyla @poliziadistato kullanıcı adlı hesaptan çekilen veriler üzerinde ön inceleme yapıldığında;

- Verinin 3246 satır ve 7 sütundan oluştuğu (Şekil 3.64, Şekil 3.65),
- Verinin sütun isimlerinin “id”, “created_at”, “text”, “followers_count”, “retweet_count”, “favourite_count” ve “source” olduğu (Şekil 3.64),
- Veri tipleri olarak “created_at”, “text” ve “source” sütunlarının object (nesne), diğer sütunların ise int64 (sayısal - integer) olarak tanımlandığı (Şekil 3.65),
- Veride 0' dan 3245'e kadar numaralandırılmış olan 3246 adet girdi bulunduğu ve her bir sütunda boş olmayan (non-null) girdi sayısının tanımlandığı, bu değerlerden de değeri olmayan sütun değerinin bulunmadığının anlaşıldığı (Şekil 3.65),
- @poliziadistato adlı kullanıcıdan çekilen verilerin 26 Mart 2018'den 08 Mart 2020 tarihleri arasında yer aldığı (Şekil 3.66) anlaşılmaktadır.



Şekil 3.67. @poliziadistato adlı kullanıcının veri paylaşımında kullandığı kaynaklar

@poliziadistato adlı kullanıcının Twitter paylaşımı yaptığı cihazların kaynağı incelendiğinde; %1,5'inin Iphone, %64,9'unun Android, %21,9'unun Twitter web istemcisi (web client), %7,7'sinin Twitter web uygulaması, %1,9'unun Ipad, %2'sinin Twitter medya stüdyosu ve yalnız bir girişin (%0.03) de Oracle Engage aracılığıyla olduğu şekil 3.67'de gösterilmektedir.

```
print(f" Maximum number of retweets {data.retweet_count.max()}")
print(f" Maximum number of favorites {data.favourite_count.max()}")

Maximum number of retweets 6524
Maximum number of favorites 9163

data.loc[data['retweet_count']==6524,['text']].values
array(['RT @Quirinale: #coronavirus: dichiarazione del Presidente #Mattarella ht
dtype=object)

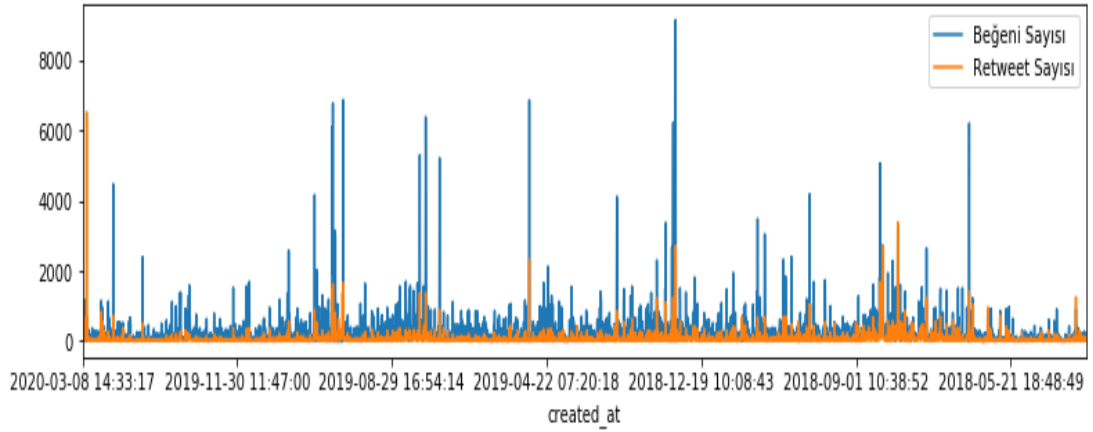
data.loc[data['retweet_count']==6524, ['created_at']].values
array(['2020-03-05 22:00:20'], dtype=object)

data.loc[data['favourite_count']==9163,['text']].values
array(['#13gennaio #Bolivia team di poliziotti italiani con polizia boliviana ha
dtype=object)

data.loc[data['favourite_count']==9163, ['created_at']].values
array(['2019-01-13 07:43:37'], dtype=object)
```

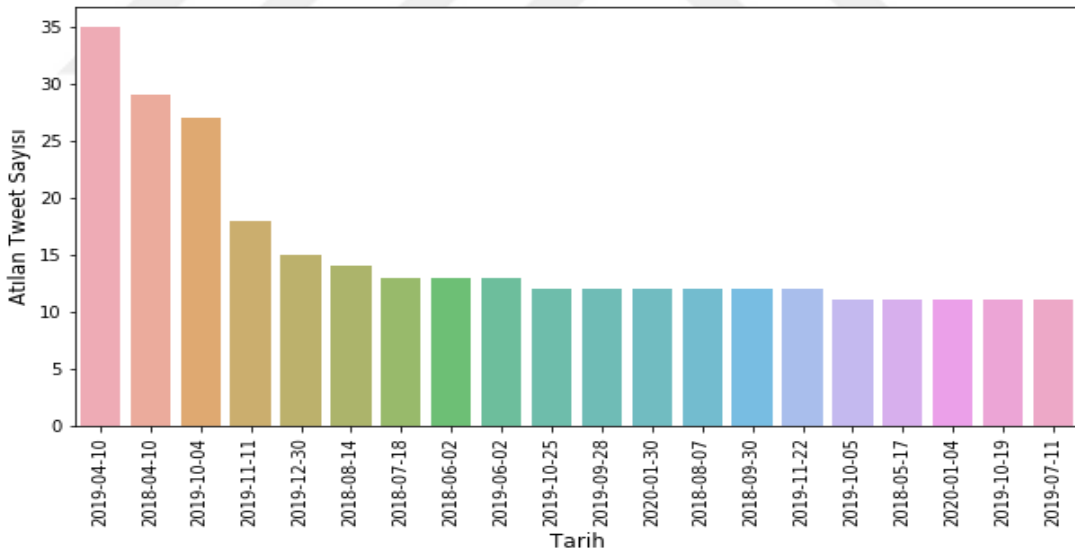
Şekil 3.68. Paylaşımların beğenilme ve retweet durumunun incelenmesi

Yapılan paylaşımlar şekil 3.68'de görüldüğü üzere incelendiğinde, 6524 tekrar paylaşım oranı ile en fazla retweet yapılan paylaşımın 05 Mart 2020 tarihinde paylaşılan yine bir retweet mesajı ('RT @Quirinale: #coronavirus: dichiarazione del Presidente #Mattarella <https://t.co/QJanArqBhw>') olduğu, 9163 beğeni ile de en fazla beğeni alan paylaşımın ("#13gennaio #Bolivia team di poliziotti italiani con polizia boliviana ha individuato e catturato #CesareBattisti\r\nFondamentale l'attività dei poliziotti Antiterrorismo italiano #Dcpp e del Servizio Cooperazione Internazionale Polizia #Scip #Criminalpol <https://t.co/CdAVZKpEY4>") 13 Ocak 2019 tarihinde yapıldığı anlaşılmaktadır. Ayrıca Twitter paylaşımlarının 848.360 kez beğenildiği (favorites), 266.834 kez tekrar paylaşıldığı (retweet) ve toplam takipçi sayısının ise 122.898 olduğu belirlenmiştir.



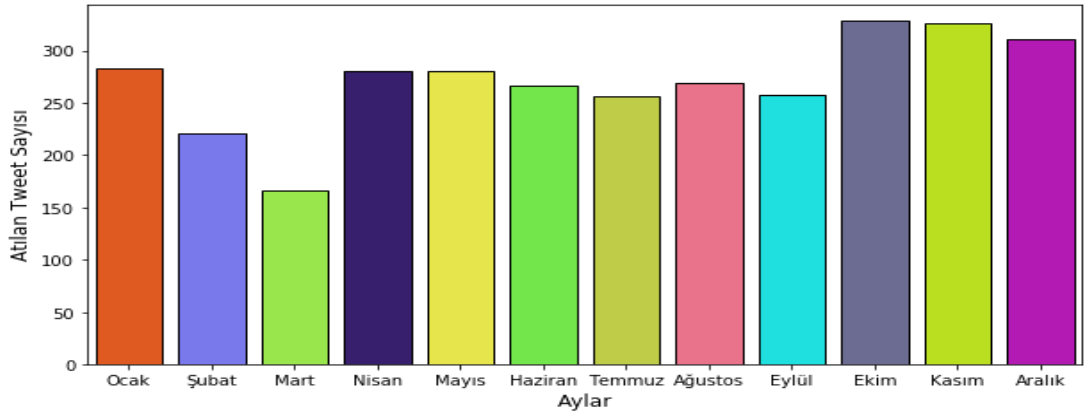
Şekil 3.69. @poliziadistato adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu

@poliziadistato adlı kullanıcının Twitter paylaşımlarının beğenilme ve retweet durumlarının tarih bazlı ortak grafiği Şekil 3.69’da gösterilmektedir (Takipçi sayısının tarih bazında sabit olmasından dolayı “followers_count” değeri ile tweet kimliğini içeren “id” değerinde ilişkisel analiz incelenmemiştir.).



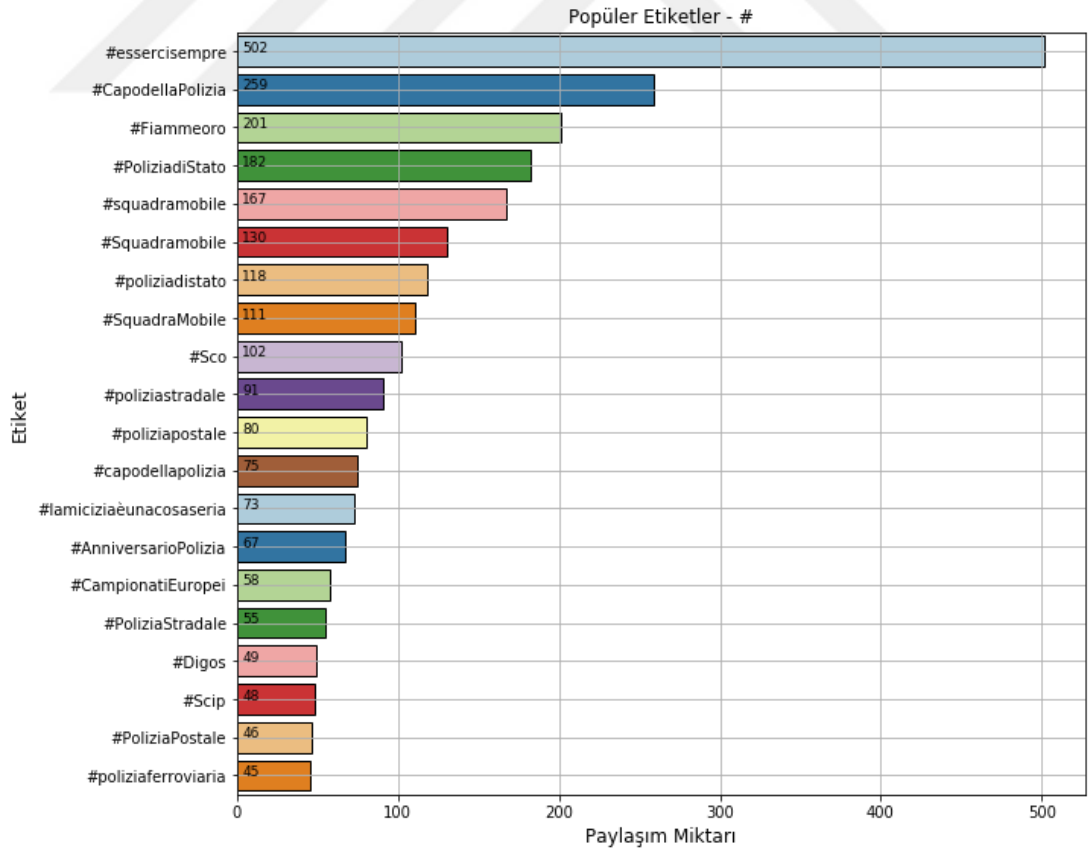
Şekil 3.70. Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı

Şekil 3.70 incelendiğinde aynı gün içerisinde sık paylaşımın yapıldığı tarihler görülmekte olup, örnek olarak en sık paylaşımın yapıldığı 04 Nisan 2019 tarihindeki 35 adet yapılan paylaşım incelendiğinde, paylaşımların kurumun 167’nci kuruluş yıldönümü kapsamında ve bazı tekrar paylaşılan tweet’lerden oluştuğu görülmektedir.



Şekil 3.71. Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı

Twitter aracılığıyla sosyal medyadan yapılan paylaşımlara şekil 3.71’de olduğu üzere aylık dağılımına bakıldığında 328 Tweet ile ekim ayı ile 326 Tweet ile kasım ayı en fazla sosyal medya paylaşımının yapıldığı aylar iken, genel seyirde paylaşımların düzenli aralıklarla yapıldığı gözlemlenmektedir.

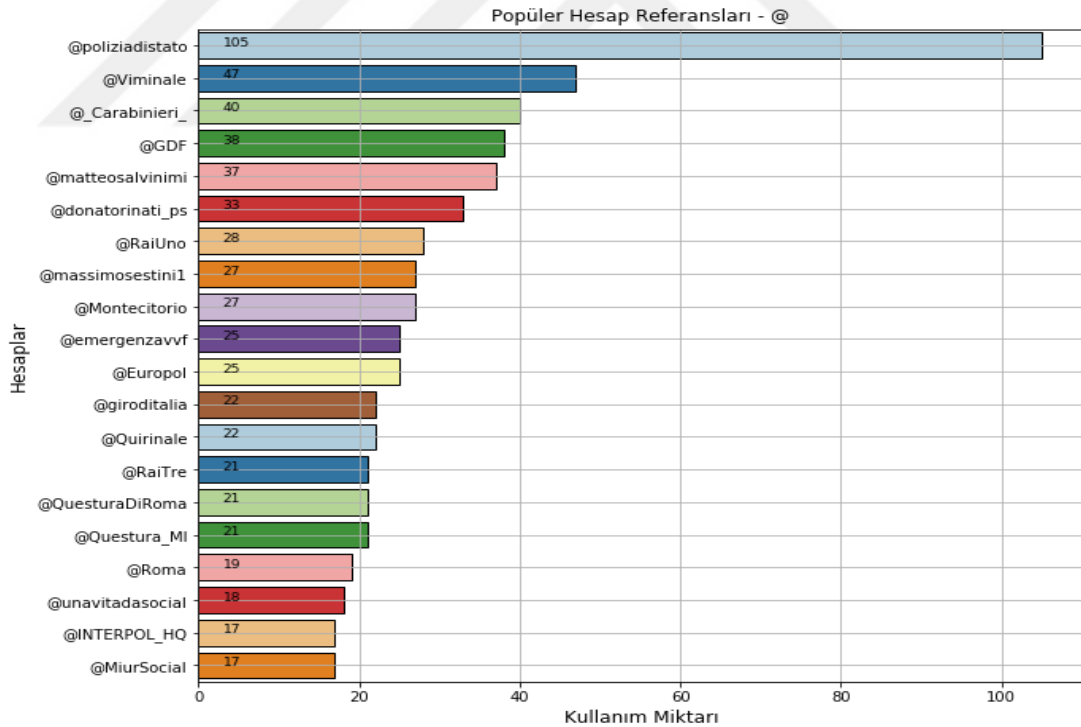


Şekil 3.72. En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)



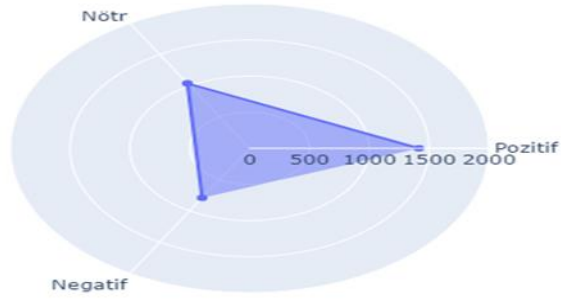
Şekil 3.73. En fazla kullanılan sosyal medya etiketlerinin (Hashtag - #) kelime bulutu

Yapılan paylaşımlarda en fazla kullanılan sosyal medya etiketlerinin 20'sine ilişkin kullanım sıklığı ile sık kullanılan sosyal medya etiketlerine ilişkin hazırlanan kelime bulutu (Maksimum kelime girdisi 1000 olarak alınmıştır.) şekil 3.72 ve 4.73'de gösterilmektedir.



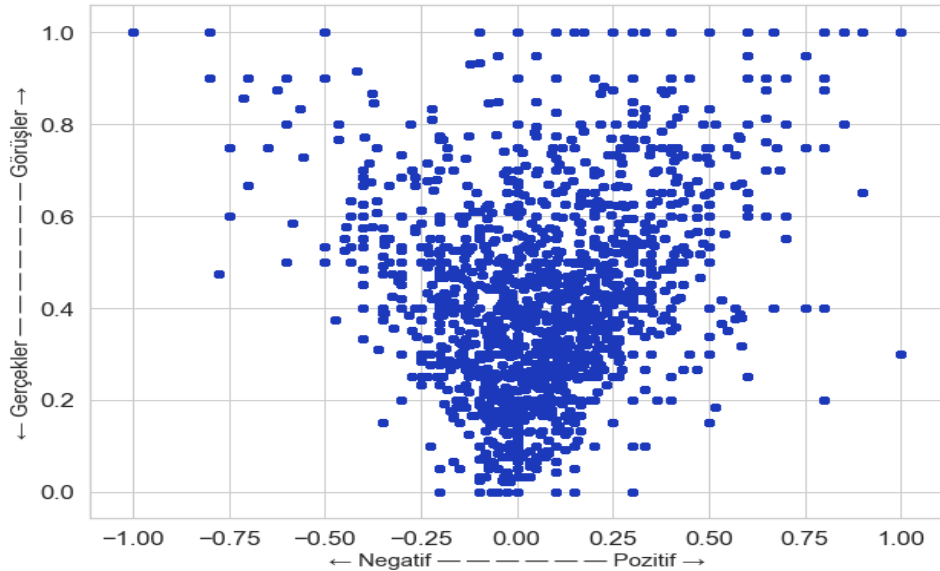
Şekil 3.74. En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)

Yapılan paylaşımlarda en fazla kullanılan sosyal medya hesaplarının 20'sine ilişkin kullanım sıklığı şekil 3.74'de gösterilmektedir.



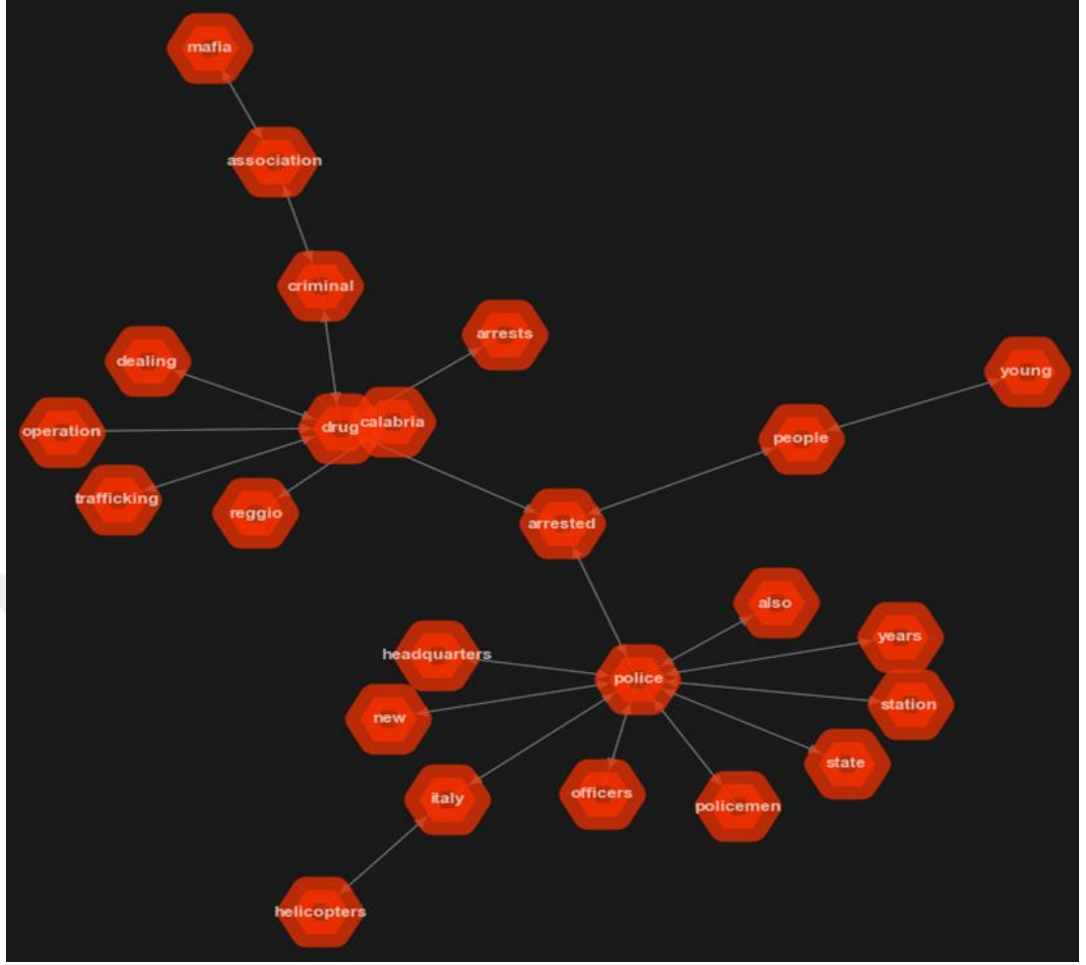
Şekil 3.75. Metnin duygusal durumunun radar grafiği

@poliziadistato kullanıcı hesabından yapılan paylaşım metinleri, metin ön işlemesine (text preprocessing) tabi tutularak tüm emojiler kaldırılmış ve İtalyanca/İngilizce karakter farklılıklarından dolayı metindeki tüm harfler küçültülerek makine öğrenimine dayalı olarak İtalyanca'dan İngilizce'ye çeviri yapılmıştır. Çeviri metni üzerinde yapılan duygusal dağılım (0=Nötr, 0>Negatif, 0<Pozitif) şekil 3.75'de görüldüğü üzere 1414 tweet ile pozitif mesajların çoğunlukta olduğu, 1043 tweet ile nötr mesajların pozitif mesajlarını izlediği ve son olarak da 789 tweet ile negatif skorlanan mesajların bulunduğu belirlenmiştir.



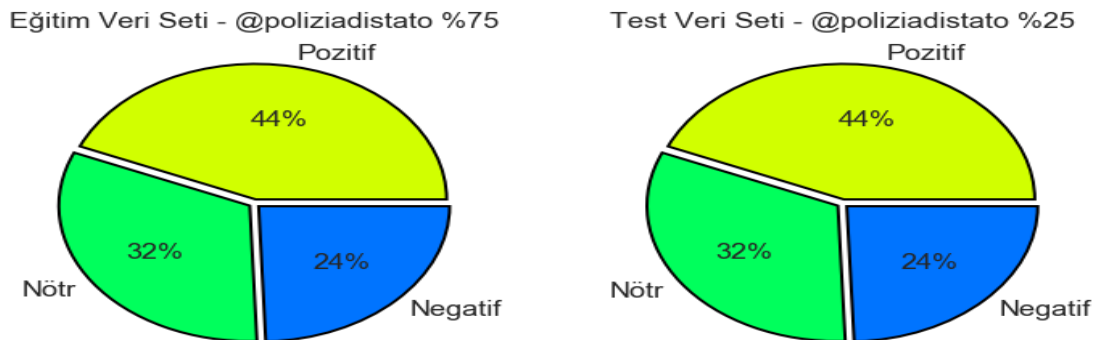
Şekil 3.76. Metnin duygusal durumunun öznellik analizi

Elde edilen çeviri üzerinden kelime bazlı olarak, her kelimenin yarattığı duygusal durum ortaya konularak, her cümleye ait duygusal puan durumu, duygusal



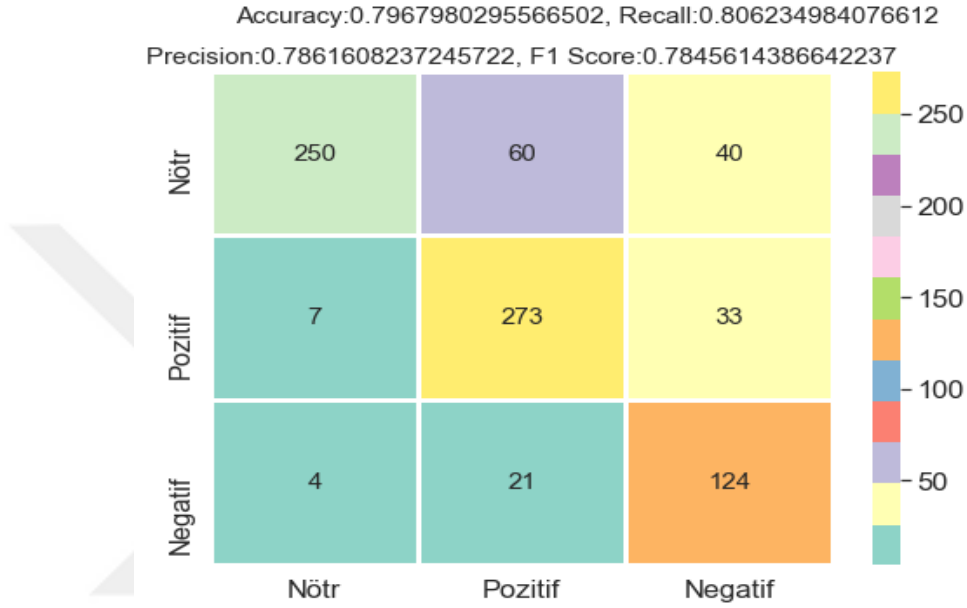
Şekil 3.79. Yapılan paylaşımlara ait kelime ağı analizi (50>Kelime Kombinasyonu)

Yapılan paylaşım metinleri BoW (Bag of words) metoduyla veri seti haline getirilmiştir. Söz konusu veri seti içerisinde de 50’den fazla kombinasyona sahip kelimelerin ağ (network) yapısı şekil 3.79’da gösterilmiştir.



Şekil 3.80. Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması

Textblob kütüphanesi aracılığıyla yapılan paylaşımların duygu durumları belirlenmiş olup, söz konusu tespitlerin kendi içinde tutarlı olup olmadığını incelemek adına makine öğrenimi modellerinden rastgele orman sınıflandırması (Random forest classification) ile veri seti şekil 3.80’de gösterildiği üzere test ve eğitim setine ayrılmıştır. Sınıflandırma için verinin %75’i eğitim, %25’i de test olarak ayrılmıştır.

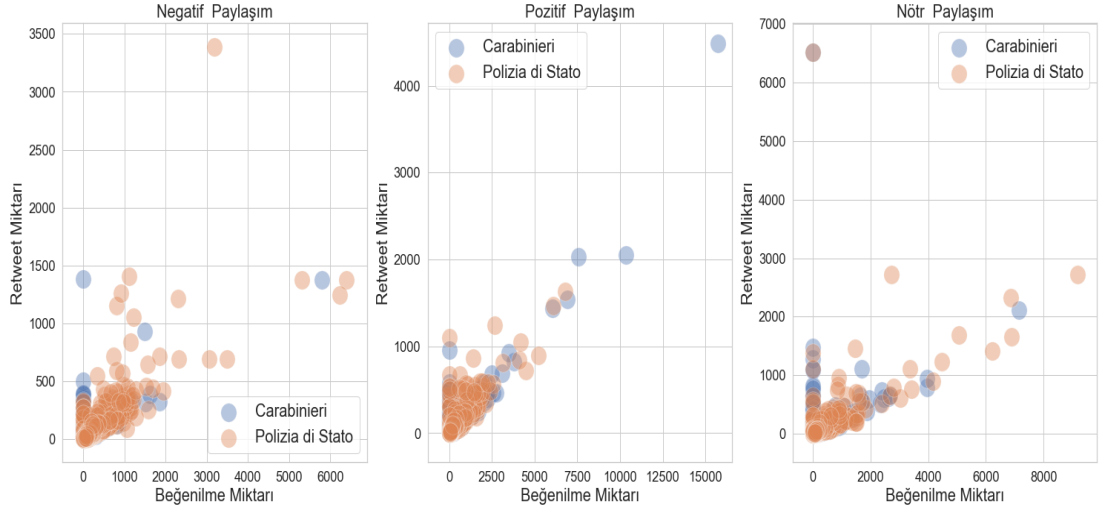


Şekil 3.81. Test verisinin duygusal matriksi

Makine öğrenimi sonrası test verileri üzerinde çıkan sonuca ilişkin matriks şekil 3.81’de görülmekte olup, yapılan öğrenme test verileri üzerinde %79.680 oranında doğruluk oranı sağlamıştır.

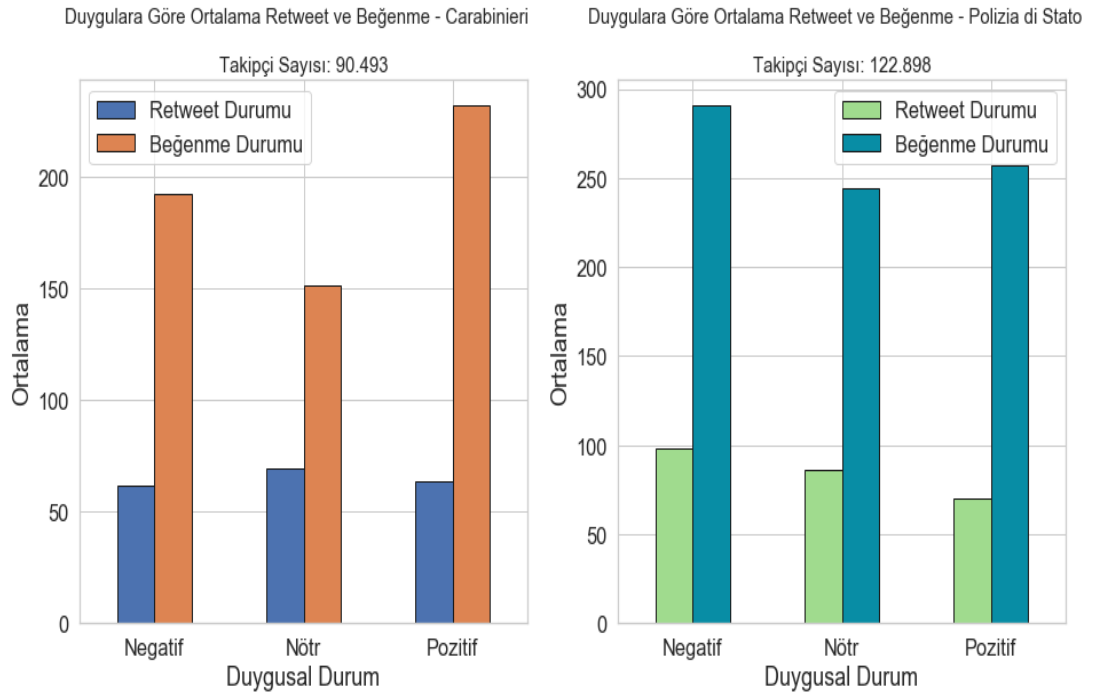
3.2.3. Arma Dei Carabinieri ve Polizia Di Stato Teşkilatlarının Mukayeseli Ortak Analizi

İtalya özelinde faaliyet gösteren kolluk teşkilatlarının topluma verdiği mesajları, mesajlarda hangi konu ve kelimeleri ağırlıkta kullandığı ile paylaştığı mesajların sosyal medyadaki etkisini mukayeseli olarak aşağıda incelenmektedir. Mukayese kapsamında korelasyon ve makine öğrenimi modellerinden karar destek vektörleri kullanılarak, tespitler görsel içerikle desteklenmektedir.



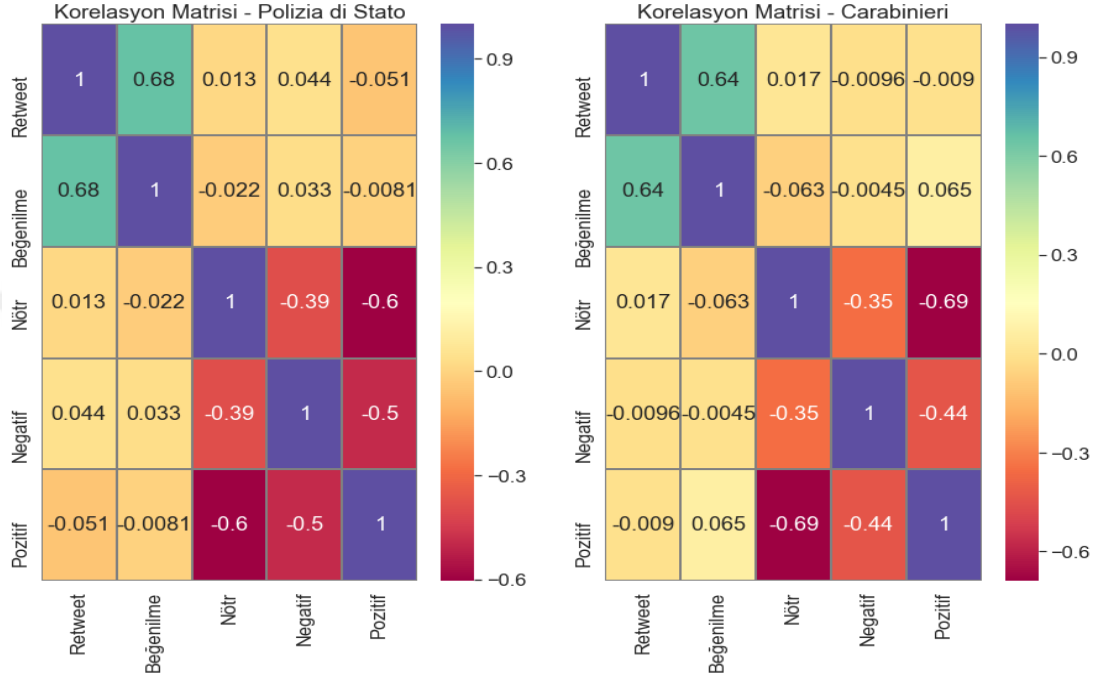
Şekil 3.82. Duygulara göre retweet ve beğenme durumu

Paylaşımlar içerisinde anomali tespiti yapmak amacıyla şekil 3.82 incelendiğinde Carabinieri özelinde bazı negatif ve nötr cümlelerde anomalisinin fazla olduğu, Polizia di Stato özelinde ise bazı pozitif cümlelerde fazla olduğu gözlemlenmiştir. Kurumlar arası genel dağılımın birbirine yakın olduğu ve en fazla anomali gösteren dağılımın nötr ve negatif paylaşımlarda yaşandığı belirlenmiştir.



Şekil 3.83. Duygulara göre ortalama retweet ve beğenme durumu

Paylaşımların genel ortalaması şekil 3.83’da görüldüğü üzere incelendiğinde, genel olarak Carabinieri’in pozitif paylaşımlarda beğeni, nötr paylaşımlarda da retweet ortalaması yüksek iken, Polizia di Stato’nun negatif paylaşımlarda ise ortalamalarının daha yüksek olduğu gözlemlenmektedir.

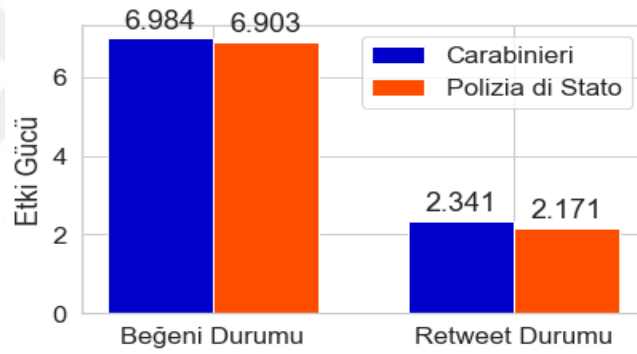


Şekil 3.84. Duygusal korelasyon matrisi

Duyguların, beğeni ve retweet üzerindeki gücünün korelasyon biçimindeki gösterimi matris biçiminde şekil 3.84’de yapılmış olup, söz konusu verilerin analizinden;

- Beğeni ve retweet arası ilişkinin Polizia di Stato’da daha güçlü olduğu (Polizia di Stato: 0.68, Carabinieri:0.64),
- Pozitif mesajlar kapsamında yapılan paylaşımları insanların beğenmesin Carabinieri’de daha yüksek ilişki düzeyinde olduğu (Polizia di Stato: -0.0081, Carabinieri:0.065),
- Pozitif mesajlar kapsamında yapılan paylaşımları insanların retweet etmesinin Carabinieri’de daha yüksek ilişki düzeyinde (Polizia di Stato: -0.051, Carabinieri:-0.009),

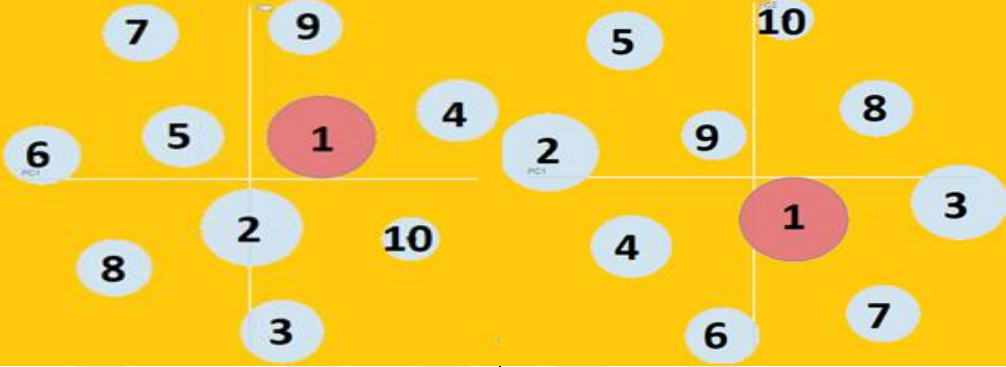
- Nötr mesajlar kapsamında yapılan paylaşımları insanların beğenmesin Polizia di Stato’da daha yüksek ilişki düzeyinde olduğu (Polizia di Stato: -0.022, Carabinieri:-0.063),
- Nötr mesajlar kapsamında yapılan paylaşımları insanların retweet etmesinin Carabinieri’de daha yüksek ilişki düzeyinde olduğu (Polizia di Stato: 0.013, Carabinieri:0.017),
- Negatif mesajlar kapsamında yapılan paylaşımları insanların beğenmesin Polizia di Stato’da daha yüksek ilişki düzeyinde olduğu (Polizia di Stato: 0.033, Carabinieri:-0.0045),
- Negatif mesajlar kapsamında yapılan paylaşımları insanların retweet etmesinin Polizia di Stato’da daha yüksek ilişki düzeyinde olduğu (Polizia di Stato: 0.044, Carabinieri:-0.0096) gözlemlenmiştir.



Şekil 3.85. Carabinieri ve Polizia di Stato’nun etki gücü

Yapılan paylaşımların takipçileri oranında hangi derecede paylaşıldığının ve retweet edildiğinin ortaya konması maksadıyla etki gücü hesaplaması yapılmış, elde edilen sonuçlar da şekil 3.85’de gösterilmektedir. Etki gücü hesaplaması yapılırken, toplam beğeni sayısı/toplam takipçi sayısı ve toplam retweet sayısı/toplam takipçi sayısı formülasyonu kullanmıştır. Buradaki amaç her takipçinin ortalama ne kadar mesajı beğendiği ve retweet ettiği tespitiyle, mesajların ne kadar güçlü olduğunu karşılaştırmalı olarak ortaya koymaktır. Bu yapılan analiz sonucunda ise Carabinieri’nin paylaşımlarının Policia di Stato’ya kıyasla etki gücünün kısmi daha yüksek olduğu anlaşılmaktadır. Bu sebeple de aradaki farkın sebebini daha iyi anlayabilmek adına makine öğrenimine dayalı paylaşım yaptıkları konuların neler olduğunun ortaya konması gerekmektedir.

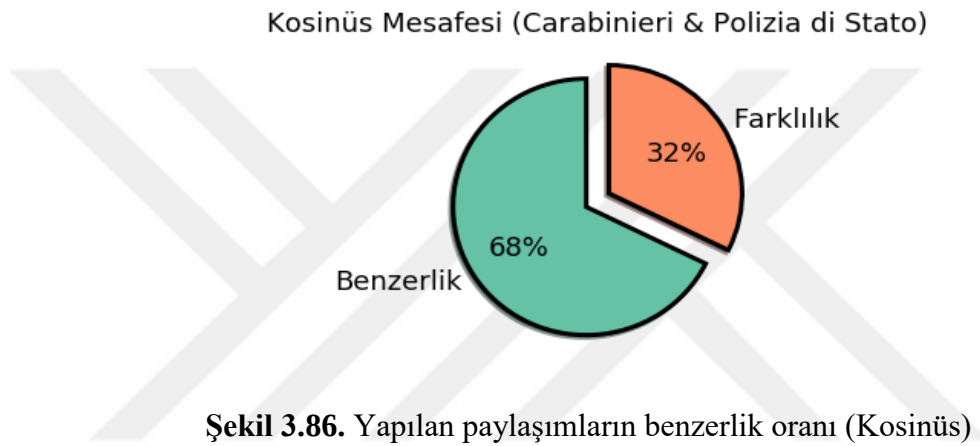
Çizelge 3.2. Carabinieri ve Polizia di Stato’nun paylaşım yaptığı konuların LDA (Latent Dirichlet Allocation) analizi ve konular arası mesafe haritalandırması

Konu	Carabinieri LDA Analizi	Polizia Di Stato LDA Analizi
Mesafe Haritası		
1	('arrest', 248.37397825112166) ('people', 223.50625832803894) ('drug', 178.67261150703087) ('carry', 136.8682944246112) ('precautionary', 135.41575448995735) ('measure', 123.30490830456682) ('company', 122.85684584560246) ('association', 104.84118889511055) ('station', 94.63326590582393) ('seize', 93.60028136122067)	('arrest', 430.6502921987093) ('drug', 293.9036337341669) ('operation', 184.39522954749827) ('criminal', 170.64374110604882) ('investigation', 148.1404275519702) ('association', 140.21917352223883) ('mafia', 124.90712387653683) ('deal', 103.65085411569471) ('robbery', 98.75514100242654) ('crime', 96.47242697170738)
2	('general', 303.44566136671756) ('weapon', 282.20487802442506) ('commander', 274.11034764988284) ('good', 176.3523996422223) ('nistri', 139.9281998031724) ('giovanni', 137.75244681260045) ('visit', 113.96317279013341) ('gen', 108.19246054951415) ('carabinieri', 102.97998598811948) ('minister', 101.23953731164347)	('year', 168.0777304741184) ('thousand', 153.89476950156532) ('police', 134.80540141811275) ('italian', 120.44409010046023) ('school', 102.2999265866241) ('train', 94.31040652425543) ('new', 85.80797178368445) ('station', 81.54177690337221) ('headquarters', 81.17692138117225) ('identify', 79.38122823990362)
3	('protection', 86.81509204285462) ('minister', 85.81990023946767) ('heritage', 79.0681004167757) ('cultural', 78.31144632492759) ('force', 78.16357126532705) ('event', 71.93677103606046) ('police', 65.59900140326039) ('command', 62.13235814427319) ('collaboration', 55.68766587227297) ('work', 53.45604118867707)	('police', 227.1651664883268) ('policeman', 142.942090381614) ('work', 98.50378754769012) ('year', 88.26676456405215) ('state', 80.71977213986699) ('project', 71.93732047302211) ('country', 67.14784891678494) ('day', 66.1050045922444) ('crime', 62.25918329826101) ('meet', 61.083829091904306)
4	('italian', 133.43735465393556) ('world', 102.22096162352923) ('carabiniere', 97.70978184524176) ('center', 84.61755182400327) ('team', 82.89055472650979) ('sport', 82.56893373677613) ('athlete', 59.04475613847288) ('congratulation', 58.57247267580358) ('cup', 49.78298499806616) ('championship', 46.149067587573064)	('victim', 103.34941203091492) ('men', 91.63492808042898) ('life', 74.31364600560549) ('people', 69.90203683583584) ('accident', 66.00372812569695) ('road', 62.90882872073936) ('safety', 61.04484921027263) ('year', 60.123208104559104) ('memory', 58.72755179820465) ('drive', 53.00350615683325)

Çizelge 3.2. Devam

Konu	Jandarma LDA Analizi	Emniyet LDA Analizi
5	('car', 74.2323825109179) ('waste', 72.69016841775044) ('control', 68.65902624778658) ('place', 67.35834528447889) ('gold', 62.048464215256125) ('medal', 58.73845222535468) ('seize', 58.50272178373626) ('del', 57.4432329606545) ('choose', 52.46961493988297) ('area', 39.92321505775399)	('italy', 164.995463715319) ('car', 100.39729774515092) ('arrest', 91.42393438872986) ('vehicle', 91.24979249764844) ('theft', 73.47167636792554) ('steal', 68.25623678000355) ('good', 65.23733193798746) ('stop', 47.39738093094984) ('citizen', 47.27271278582139) ('use', 46.86092615881875)
6	('weapon', 116.52074887443356) ('order', 96.74879342044648) ('family', 71.21479634661327) ('activity', 61.9597212877149) ('regiment', 58.41426444549891) ('police', 53.77460532371712) ('new', 53.58641161825598) ('mission', 53.07853815033408) ('train', 51.89603904586074) ('discover', 48.46913572851335)	('security', 58.72364968461304) ('case', 53.48329293269614) ('center', 51.221599195269924) ('social', 51.01482181715425) ('network', 50.491220360356394) ('public', 50.026661951296646) ('investigation', 50.02294448767876) ('sign', 49.932602048696765) ('agreement', 47.654339334475985) ('prevention', 42.23073654502919)
7	('day', 141.44944716095716) ('year', 96.04463641031029) ('thanks', 85.87295888793624) ('wish', 82.61162875336827) ('today', 62.35357248633629) ('military', 60.874088916122574) ('great', 56.678864365220925) ('life', 51.2209002077548) ('commitment', 50.518242614845285) ('make', 48.17401351299217)	('child', 85.387672822904) ('traffic', 62.737686334526785) ('road', 62.63247266921668) ('home', 43.624505538033254) ('agent', 38.33131719802307) ('section', 37.402591776207274) ('family', 36.153036543078294) ('city', 35.344633469405395) ('blood', 34.90411331115451) ('trieste', 32.55937785163538)
8	('operation', 167.97514313815287) ('food', 77.58370630952868) ('anti', 58.47712328332176) ('health', 46.048078488671564) ('weapon', 44.127664524915865) ('territory', 43.548459492370654) ('international', 40.951971928812014) ('agri', 38.72576376723813) ('environmental', 38.034971086309106) ('ndrangheta', 35.780074291987695)	('medal', 122.57204427419356) ('win', 94.03895108220674) ('gold', 85.29594156750683) ('man', 77.72417565975915) ('bronze', 67.59726225271747) ('team', 64.50113667851659) ('silver', 57.52052651506443) ('champion', 54.17266257797533) ('final', 52.702122011866685) ('podium', 50.24460238938278)
9	('weapon', 117.68320383324195) ('anniversary', 81.02513382552118) ('historical', 72.79578771427533) ('helicopter', 51.869502179079014) ('ceremony', 47.828604043217716) ('open', 46.128269108079174) ('young', 41.35897331034257) ('live', 40.94151308987802) ('squadron', 40.23034656147328) ('want', 39.76203670371724)	('today', 158.5170409286343) ('rome', 113.38897761291526) ('concert', 55.74012775650618) ('piazza', 42.91805326394807) ('naples', 39.65870822525099) ('journalist', 39.09182131042804) ('san', 37.09356424639031) ('celebrate', 36.59737174131375) ('direction', 36.152740808743204) ('gabrielli', 35.452383536189316)
10	('condolence', 51.10013448822861) ('express', 51.00670339432092) ('european', 50.72839385409303) ('road', 33.83754138932433) ('chief', 32.94100946508626) ('nature', 30.364763753523306) ('come', 28.589570625037837) ('accident', 28.53503275702075) ('environment', 28.380288450612788) ('action', 25.253934864764034)	('president', 52.55151109764961) ('travel', 47.622158507694024) ('follow', 47.611863496279135) ('shot', 45.962082552436954) ('history', 43.51256832548134) ('free', 42.73780989191156) ('today', 40.57450674655993) ('helicopter', 39.86915733147086) ('presentation', 33.84378413249903) ('image', 33.39736711610289)

Çizelge 3.2’de konu modellemeye aktif olarak kullanılan LDA modellemesiyle, kurumların paylaşımları 10 konu başlığı altında incelenmiştir. 10 konu içinde de her konudan en yüksek ilişki düzeyine sahip 10 kelime örneklem olarak çizelgede gösterilmiştir. Her konunun vektörel düzeyde konular arası mesafe haritası çizelgenin başında verilmiş olup, kelimelerin yanında yer alan ifadeler de vektörel düzeyde konu içinde ilişkisini ifade etmektedir. Bu kapsamda konu bazında yapılan incelemelerde, paylaşım yapılan konuların ve vektörel düzlemdeki mesafelerinin ayrıştığı rahatlıkla görülmektedir.



Paylaşım yapılan konuların vektörel dağılımlarındaki farklılıkların, metin korpuslarının geneline yayılıp tespit edilmesi maksadıyla, öncelikle daha net bir sonucun çıkması için kelimelerin kökleri çözümlenerek (lemmatization) işlem yapılmıştır. Metin içindeki kelime (token) yapılarının her biri vektörel düzeye indirgenerek, şekil 3.43’de görüldüğü üzere kosinüs mesafesi yöntemi ile paylaşım yapılan metnin benzerlik oranı %68 olarak bulunmuştur.

3.3. Fransa Özelinde Kolluk Teşkilatlarının Paylaşımlarının İncelenmesi

Gendarmerie Nationale ve Police Nationale’nin resmi paylaşımlarından alınan örneklem üzerinden keşfedici veri analizi yapılarak, sosyal medya kullanım alışkanlıkları üzerinden duyarlılık analizi ve söz konusu duyarlılığın makine öğrenimine dayalı tahminlemesi yapılmıştır.

3.3.1. Gendarmerie Nationale'nin Paylaşımlarının İncelenmesi

Fransa'da faaliyet gösteren jandarma türü kolluk gücünün resmi hesabı olan “Gendarmerie Nationale” başlıklı ve “@Gendarmerie” kullanıcı adlı Twitter hesabından 09 Mart 2020 tarihinde herkese açık olarak yapılan paylaşımlar Tweepy kütüphanesi aracılığıyla Python programlama dili kullanılarak, csv formatında çekilmiştir.



Şekil 3.87. Gendarmerie Nationale resmi Twitter sayfasının ekran resmi

	id	created_at	text	followers_count	retweet_count	favourite_count	source
0	1236567253295992837	2020-03-08 08:19:42	#JournéeDesDroitsDesFemmes/r/nDepuis 75 ans, l...	450089	256	918	Twitter Media Studio
1	1236270484301455361	2020-03-07 12:40:26	#securite /r/n🍷 Terminer un repas sans dessert...	450089	108	289	Twitter Web App
2	1235967889372581890	2020-03-06 16:38:02	#BellesAffaires La gendarmerie, les @douane_fr...	450089	39	140	Twitter for iPhone
3	1235890321965355009	2020-03-06 11:29:49	#Coronavirus/r/nIl existe des gestes simples p...	450089	64	129	Twitter Media Studio
4	1235822662519107584	2020-03-06 07:00:57	#Cybersécurité 🛡️ Ne laissez pas les #pourriels...	450089	33	75	Twitter Web App

Şekil 3.88. @Gendarmerie kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler – I

```
print(data.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3239 entries, 0 to 3238
Data columns (total 7 columns):
id                3239 non-null int64
created_at        3239 non-null object
text              3239 non-null object
followers_count   3239 non-null int64
retweet_count     3239 non-null int64
favourite_count   3239 non-null int64
source            3239 non-null object
dtypes: int64(4), object(3)
memory usage: 177.3+ KB
None
```

Şekil 3.89. @Gendarmerie kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler –II

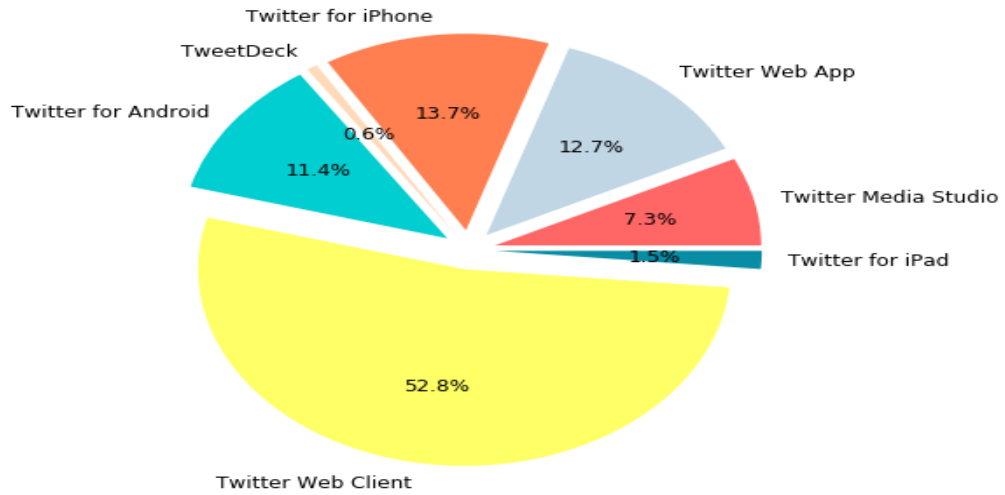
```
print(f" Data Available since {data.created_at.min()}")
print(f" Data Available upto {data.created_at.max()}")

Data Available since 2017-10-21 07:46:08
Data Available upto 2020-03-08 08:19:42
```

Şekil 3.90. @Gendarmerie kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler –III

Pandas kütüphanesi aracılığıyla @Gendarmerie kullanıcı adlı hesaptan çekilen veriler üzerinde ön inceleme yapıldığında;

- Verinin 3239 satır ve 7 sütundan oluştuğu (Şekil 3.88, Şekil 3.89),
- Verinin sütun isimlerinin “id”, “created_at”, “text”, “followers_count”, “retweet_count”, “favourite_count” ve “source” olduğu (Şekil 3.89),
- Veri tipleri olarak “created_at”, “text” ve “source” sütunlarının object (nesne), diğer sütunların ise int64 (sayısal - integer) olarak tanımlandığı (Şekil 3.89),
- Veride 0' dan 3238'e kadar numaralandırılmış olan 3239 adet girdi bulunduğu ve her bir sütunda boş olmayan (non-null) girdi sayısının tanımlandığı, bu değerlerden de değeri olmayan sütun değerinin bulunmadığının anlaşıldığı (Şekil 3.46),
- @Gendarmerie adlı kullanıcıdan çekilen verilerin 21 Ekim 2017'den 09 Mart 2020 tarihleri arasında yer aldığı (Şekil 3.90) anlaşılmaktadır.



Şekil 3.91. @Gendarmerie adlı kullanıcının veri paylaşımında kullandığı kaynaklar

@Gendarmerie adlı kullanıcının Twitter paylaşımı yaptığı cihazların kaynağı incelendiğinde; %13,7'sinin Iphone, %11,4'ünün Android, %52,8'inin Twitter web istemcisi (web client), %12,7'sinin Twitter web uygulaması, %1,5'inin Ipad, %0,6'sının TweetDeck, %7,3'ünün de Twitter medya stüdyosu aracılığıyla olduğu şekil 3.91'de gösterilmektedir.

```
print(f" Maximum number of retweets {data.retweet_count.max()}")
print(f" Maximum number of favorites {data.favourite_count.max()}")

Maximum number of retweets 38654
Maximum number of favorites 55914

data.loc[data['retweet_count']==38654,['text']].values

array([[ 'RT @gerardcollomb: Le lieutenant-colonel Arnaud Beltrame nous a quittés.\r\nMort p
n\'oubliera son héroïsme, sa...']],
      dtype=object)

data.loc[data['retweet_count']==38654, ['created_at']].values

array([[ '2018-03-24 05:09:22']], dtype=object)

data.loc[data['favourite_count']==55914,['text']].values

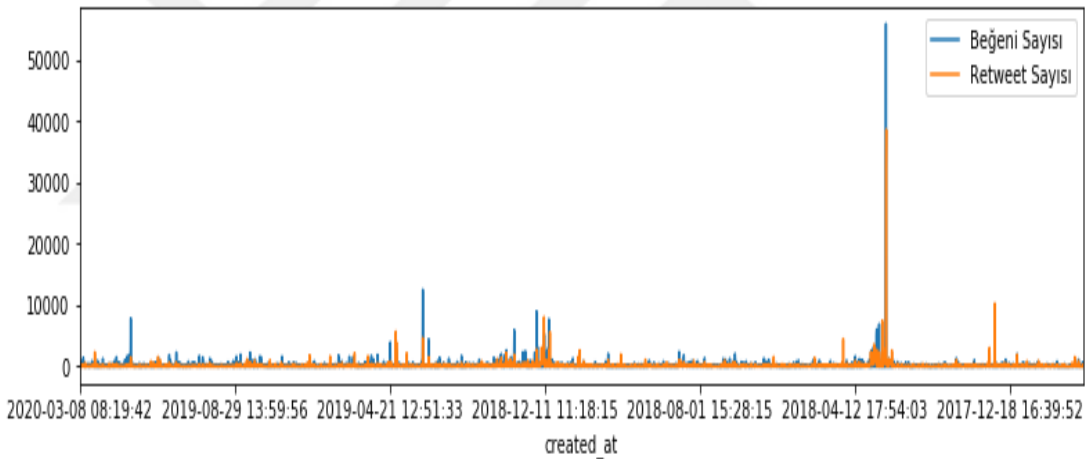
array([[ "#DGGN C'est avec une très vive émotion que je souhaite rendre solennellement homm
Beltrame décédé cette nuit. Je m'incline devant le courage, le sens du sacrifice et l'exemp
a vie pour la liberté des otages. https://t.co/GnzmXWuFOX"]],
      dtype=object)

data.loc[data['favourite_count']==55914,['created_at']].values

array([[ '2018-03-24 06:53:23']], dtype=object)
```

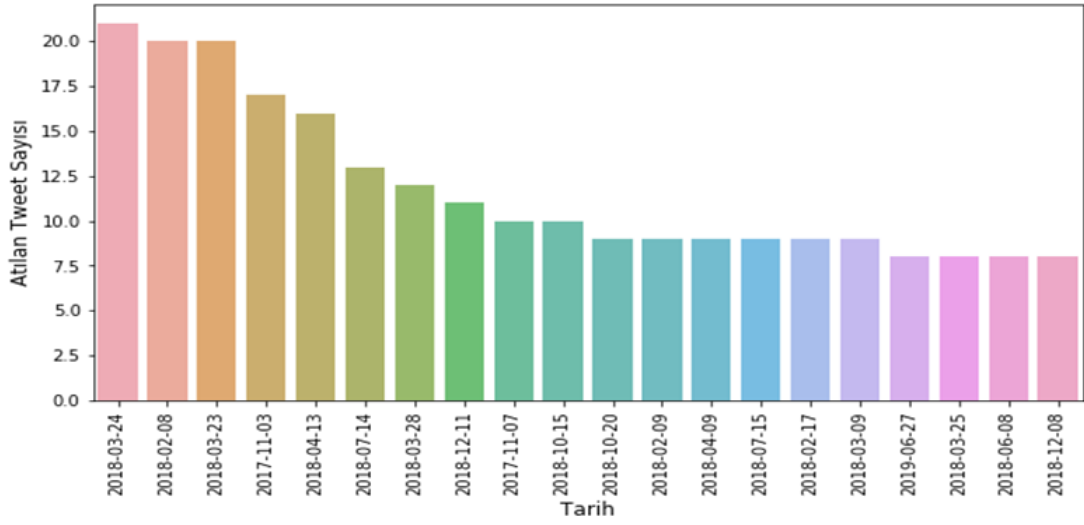
Şekil 3.92. Paylaşımların beğenilme ve tekrar paylaşılma durumunun incelenmesi

Yapılan paylaşımlar şekil 3.92’de görüldüğü üzere incelendiğinde, 38654 tekrar paylaşım oranı ile en fazla retweet yapılan paylaşımın 24 Mart 2018 tarihinde paylaşılan yine bir retweet mesajı ('RT @gerardcollomb: Le lieutenant-colonel Arnaud Beltrame nous a quittés.\r\nMort pour la patrie.\r\nJamais la France n’oubliera son héroïsme, sa...') olduğu, 55914 beğeni ile de en fazla beğeni alan paylaşımın ('#DGGN C'est avec une très vive émotion que je souhaite rendre solennellement hommage à l'héroïsme de notre camarade A.Beltrame décédé cette nuit. Je m'incline devant le courage, le sens du sacrifice et l'exemplarité de cet officier qui a donné sa vie pour la liberté des otages. <https://t.co/GnzmXWuFOX>') 24 Mart 2018 tarihinde yapıldığı anlaşılmaktadır. Ayrıca Twitter paylaşımlarının 691.814 kez beğenildiği (favorites), 561.630 kez tekrar paylaşıldığı (retweet) ve toplam takipçi sayısının ise 450.089 olduğu belirlenmiştir.



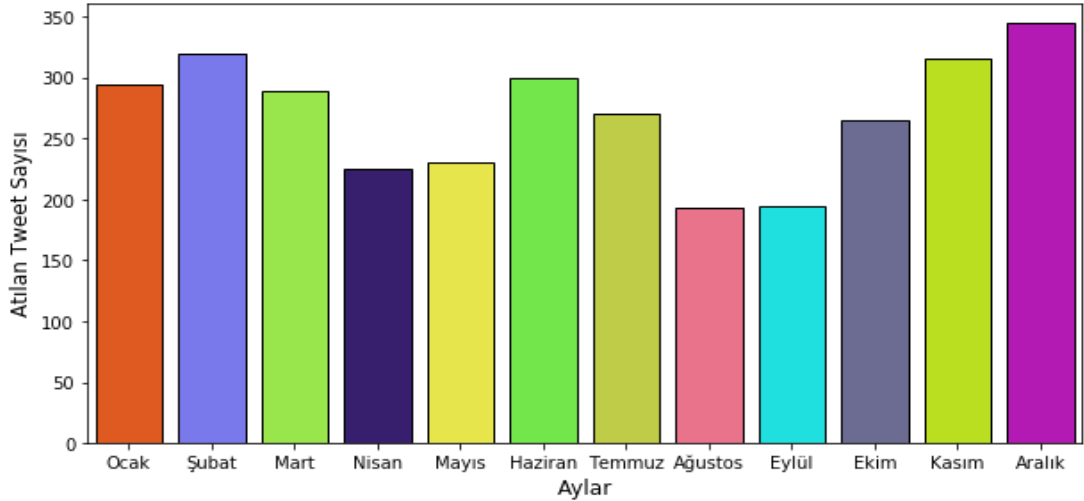
Şekil 3.93. @Gendarmerie adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu

@Gendarmerie adlı kullanıcının Twitter paylaşımlarının beğenilme ve retweet durumlarının tarih bazlı ortak grafiği şekil 3.93’de gösterilmektedir (Takipçi sayısının tarih bazında sabit olmasından dolayı “followers_count” değeri ile tweet kimliğini içeren “id” değeri ilişkisel analiz incelenmemiştir.).

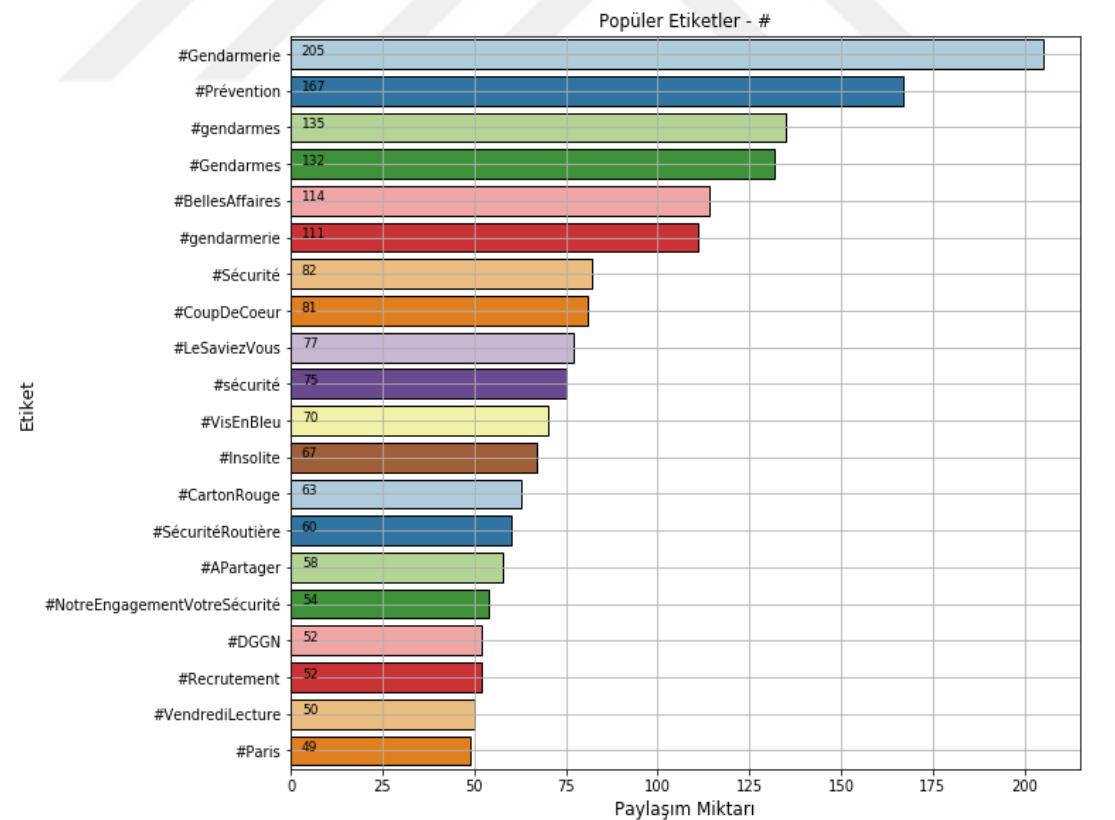


Şekil 3.94. Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı

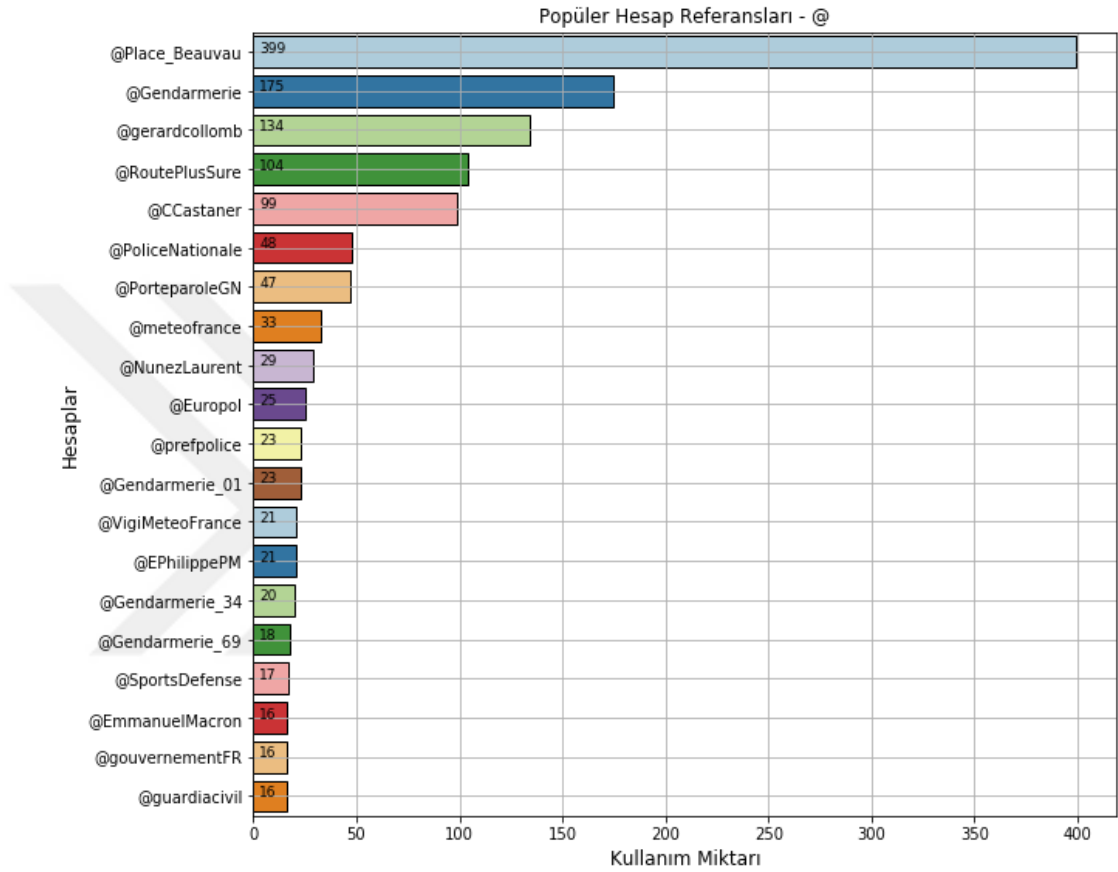
Şekil 3.94 incelendiğinde aynı gün içerisinde sık paylaşımın yapıldığı tarihler görülmekte olup, örnek olarak en sık paylaşımın yapıldığı 24 Mart 2018 tarihindeki 21 adet yapılan paylaşım incelendiğinde, paylaşımların Fransa'daki İŞİD saldırısında rehineyle yer değiştiren kahraman jandarma personelinin ölümü kapsamında mesajlar ile bazı tekrar paylaşılan tweet'lerden oluştuğu görülmektedir.



Şekil 3.95. Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı

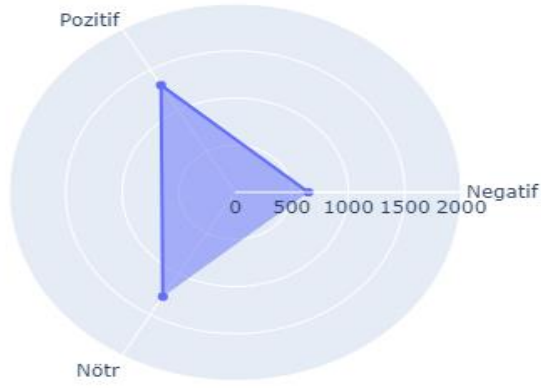
[illegible]

Yapılan paylaşımlarda en fazla kullanılan sosyal medya etiketlerinin 20'sine ilişkin kullanım sıklığı ile sık kullanılan sosyal medya etiketlerine ilişkin hazırlanan kelime bulutu (Maksimum kelime girdisi 1000 olarak alınmıştır.) şekil 3.96 ve 3.97'de gösterilmektedir.



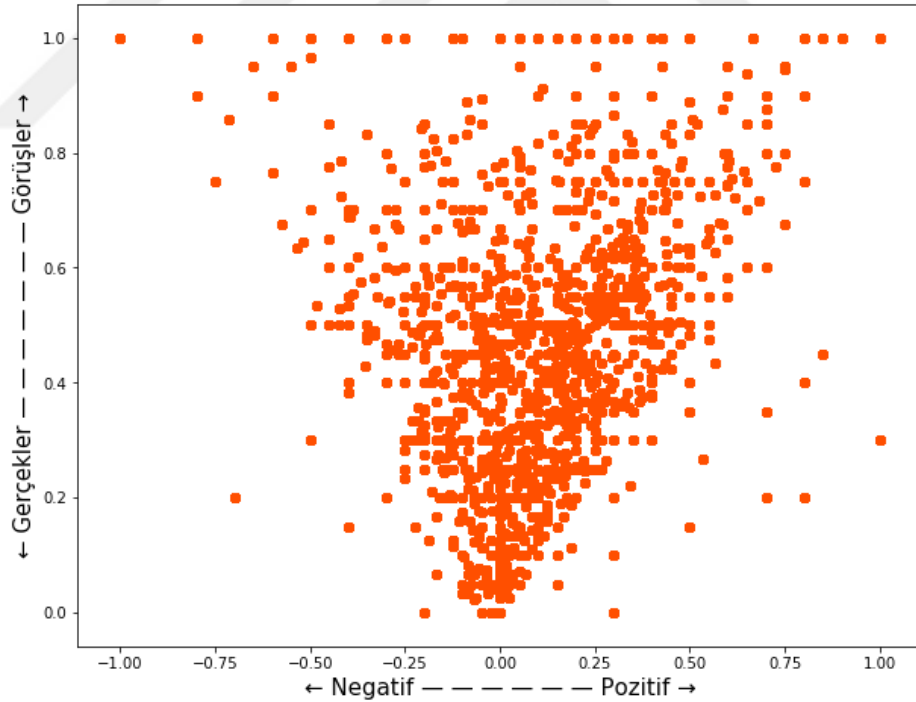
Şekil 3.98. En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)

Yapılan paylaşımlarda en fazla kullanılan sosyal medya hesaplarının 20'sine ilişkin kullanım sıklığı şekil 3.98'de gösterilmektedir. Ayrıca @Gendarmerie kullanıcı hesabından yapılan paylaşım metinleri, metin ön işlemesine (text preprocessing) tabi tutularak tüm emojiler kaldırılmış ve Fransızca/İngilizce karakter farklılıklarından dolayı metindeki tüm harfler küçültülerek makine öğrenimine dayalı olarak Fransızca'dan İngilizce'ye çeviri yapılmıştır.



Şekil 3.99. Metnin duygusal durumunun radar grafiği

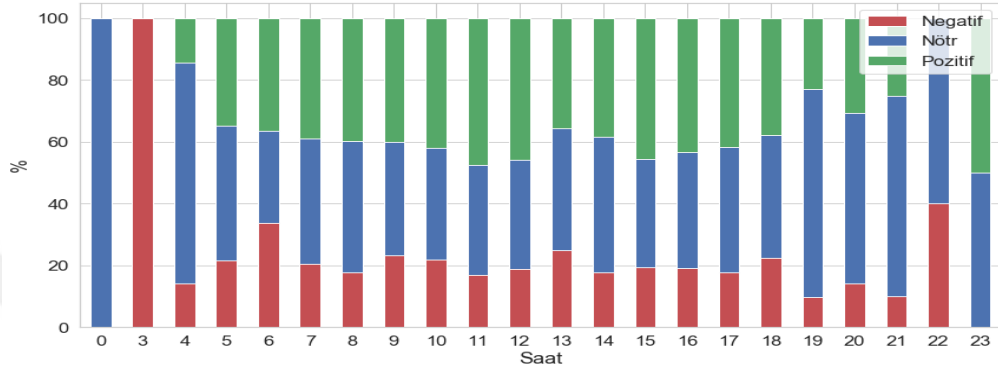
Çeviri metni üzerinde yapılan duygusal dağılım ($0=Nötr$, $0>Negatif$, $0<Pozitif$) şekil 3.99’da görüldüğü üzere 1320 tweet ile pozitif mesajların çoğunlukta olduğu, 1277 tweet ile nötr mesajların pozitif mesajları izlediği ve son olarak da 642 tweet ile negatif skorlanan mesajların bulunduğu belirlenmiştir.



Şekil 3.100. Metnin duygusal durumunun öznellik analizi

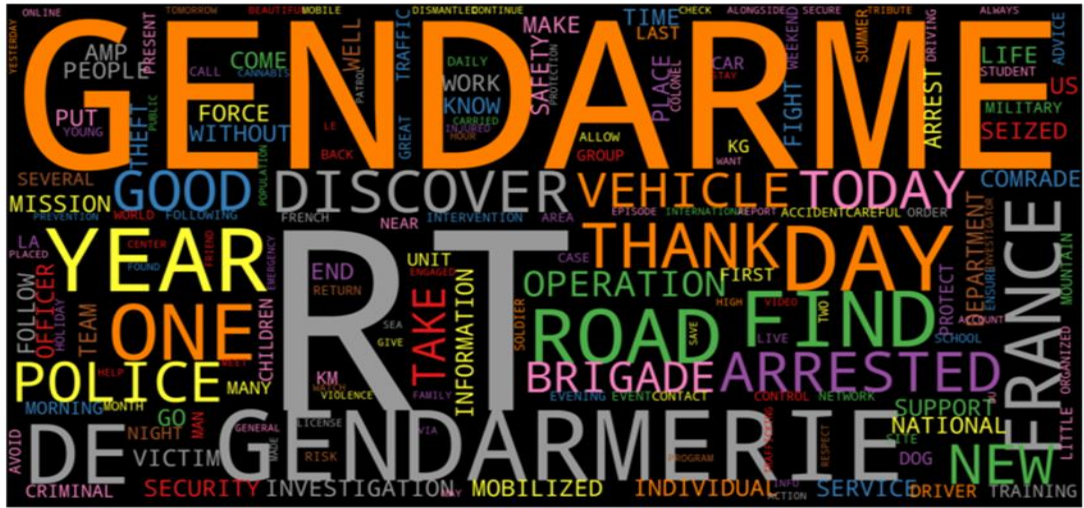
Elde edilen çeviri üzerinden kelime bazlı olarak, her kelimenin yarattığı duygusal durum ortaya konularak, her cümleye ait duygusal puan durumu, duygusal durumu ve ortaya konulan kanının öznellik ihtimali çıkarılmıştır. Metnin duygusal

durumu ile öznelliğini beraber analiz ettiğimizde ise, cümlelerin içinde yer alan her kelimenin skorlanması ile cümle skoru elde edilmiş olup, ana korpusun çoğunluğunun 0.5 öznellik bandını geçmediği şekil 3.100’de gösterilmektedir. Fakat az sayıda da olsa öznelliği 1 olarak yer alan 25 tweet için, negatif, nötr veya pozitif gibi değerlendirmelerde bulunulmuştur.



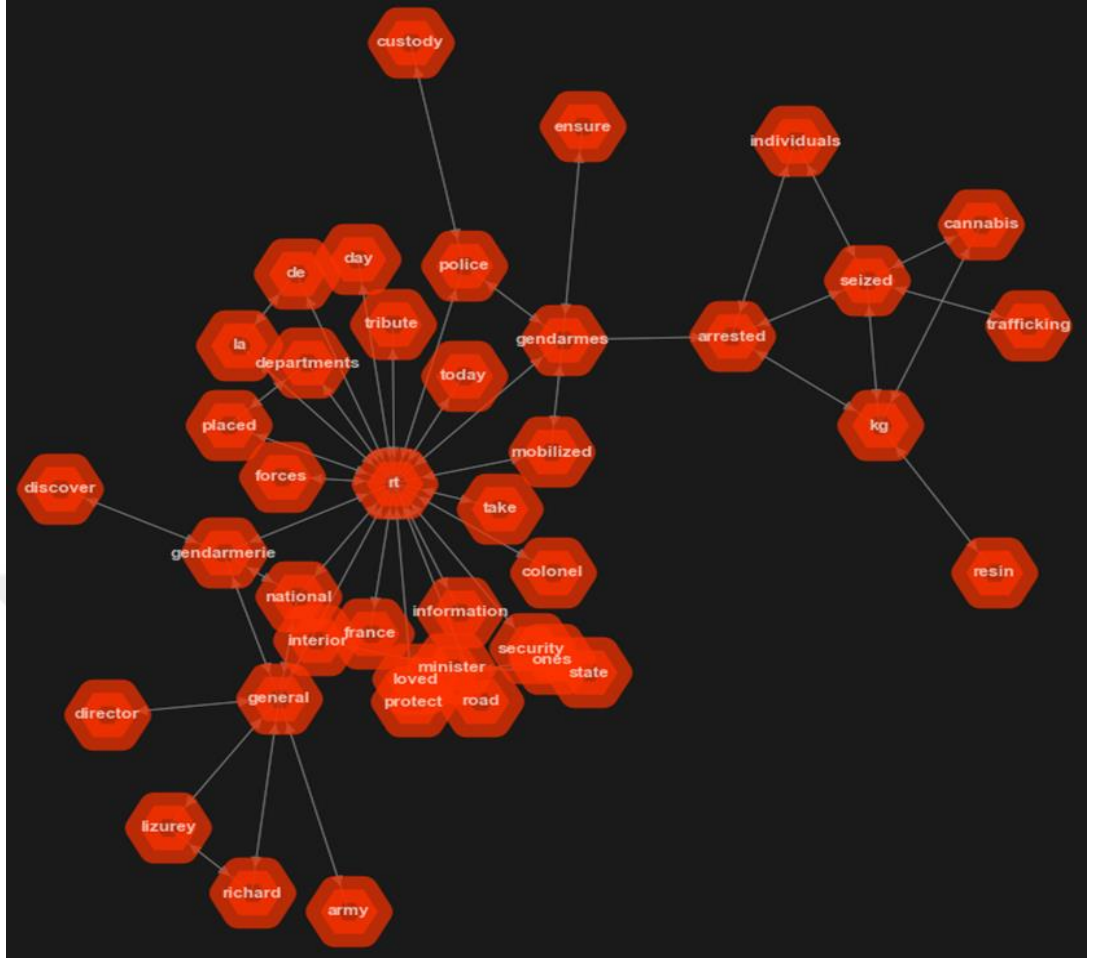
Şekil 3.101. Metnin duygusal durumunun saat bazlı incelenmesi

Paylaşımların yapıldığı zaman dilimi ile duygusal durumu arasındaki ilişki şekil 3.101’de gösterilmekte olup, incelenen korpusun yaklaşık 2 yılı geçkin süreyi kapsamasından dolayı zamanın da duygusal duruma etkisi olduğu söylenebilmektedir.



Şekil 3.102. Paylaşımlar arasında sık kullanılan ifadelerle ilgili kelime bulutu

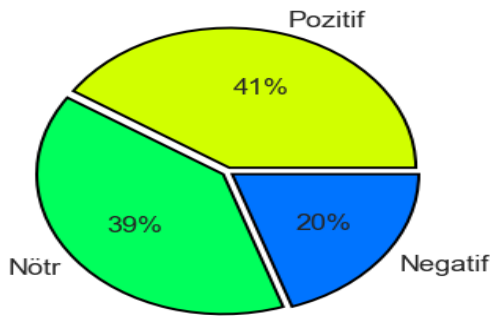
Yapılan paylaşımlarda en sık kullanılan ifadeler bazında (# ve @ gibi ifadeler olmaksızın) şekil 3.02’deki kelime bulutunun elde edildiği görülmektedir.



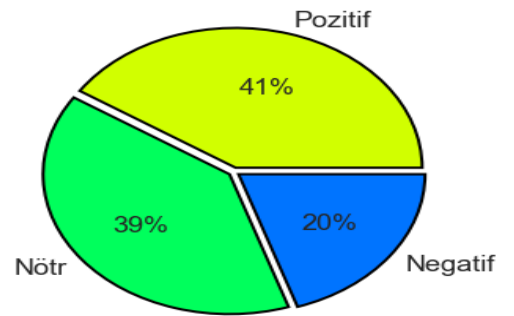
Şekil 3.103. Yapılan paylaşımlara ait kelime ağı analizi (25>Kelime Kombinasyonu)

Yapılan paylaşım metinleri BoW (Bag of words) metoduyla veri seti haline getirilmiştir. Söz konusu veri seti içerisinde de 25’den fazla kombinasyona sahip kelimelerin ağ (network) yapısı şekil 3.103’de gösterilmiştir.

Eğitim Veri Seti - @Gendarmerie %75

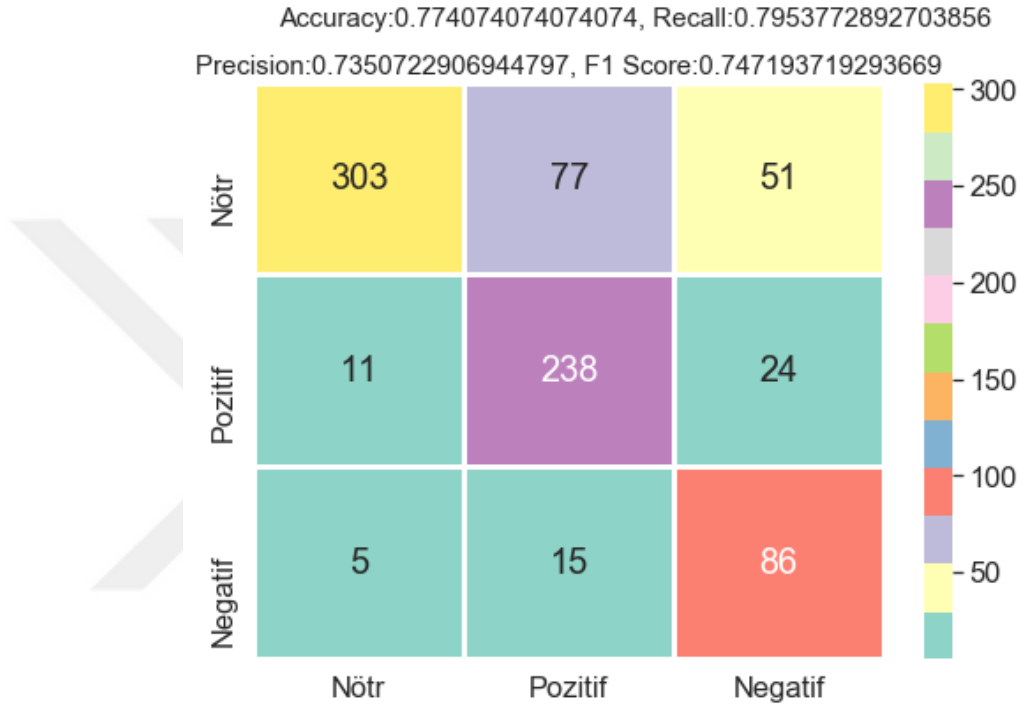


Test Veri Seti - @Gendarmerie %25



Şekil 3.104. Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması

Textblob kütüphanesi aracılığıyla yapılan paylaşımların duygu durumları belirlenmiş olup, söz konusu tespitlerin kendi içinde tutarlı olup olmadığını incelemek adına makine öğrenimi modellerinden rastgele orman sınıflandırması (Random forest classification) ile veri seti şekil 3.104’de gösterildiği üzere test ve eğitim setine ayrılmıştır. Sınıflandırma için verinin %75’i eğitim, %25’i de test olarak ayrılmıştır.



Şekil 3.105. Test verisinin duygusal matriksi

Makine öğrenimi sonrası test verileri üzerinde çıkan sonuca ilişkin matriks şekil 3.105’de görülmekte olup, yapılan öğrenme test verileri üzerinde %77.407 oranında doğruluk oranı sağlamıştır.

3.3.2. Police Nationale’nin Paylaşımlarının İncelenmesi

Fransa’da faaliyet gösteren polis türü kolluk gücünün resmi hesabı olan “Police Nationale” başlıklı ve @PoliceNationale kullanıcı adlı Twitter hesabından 09 Mart 2020 tarihinde herkese açık olarak yapılan paylaşımlar Tweepy kütüphanesi aracılığıyla Python programlama dili kullanılarak, csv formatında çekilmiştir.



Şekil 3.106. Police Nationale resmi Twitter sayfasının ekran resmi

	id	created_at	text	followers_count	retweet_count	favourite_count	source
0	1236730117012873221	2020-03-08 19:06:51	@HmlEmpire Bonsoir, les enquêteurs de la plate...	495541	25	231	Twitter for iPad
1	1236729967339331584	2020-03-08 19:06:16	@pinto_emelo Bonsoir, les enquêteurs de la pla...	495541	155	1531	Twitter for iPad
2	1236729846379790340	2020-03-08 19:05:47	@Comme1Timal Bonsoir, les enquêteurs de la pla...	495541	450	4566	Twitter for iPad
3	1236562182139260928	2020-03-08 07:59:33	Des policiers spécialisés luttent chaque jour ...	495541	41	136	Twitter for iPad
4	1236562161964548097	2020-03-08 07:59:28	Nous souhaitons surtout faire passer un messag...	495541	103	310	Twitter for iPad

Şekil 3.107. @PoliceNationale kullanıcı adlı hesaptan çekilen veriye ait bilgiler- I

```
print(data.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3141 entries, 0 to 3140
Data columns (total 7 columns):
id                3141 non-null int64
created_at        3141 non-null object
text              3141 non-null object
followers_count   3141 non-null int64
retweet_count     3141 non-null int64
favourite_count   3141 non-null int64
source            3141 non-null object
dtypes: int64(4), object(3)
memory usage: 171.9+ KB
None
```

Şekil 3.108. @PoliceNationale kullanıcı adlı hesaptan çekilen veriye ait bilgiler-II

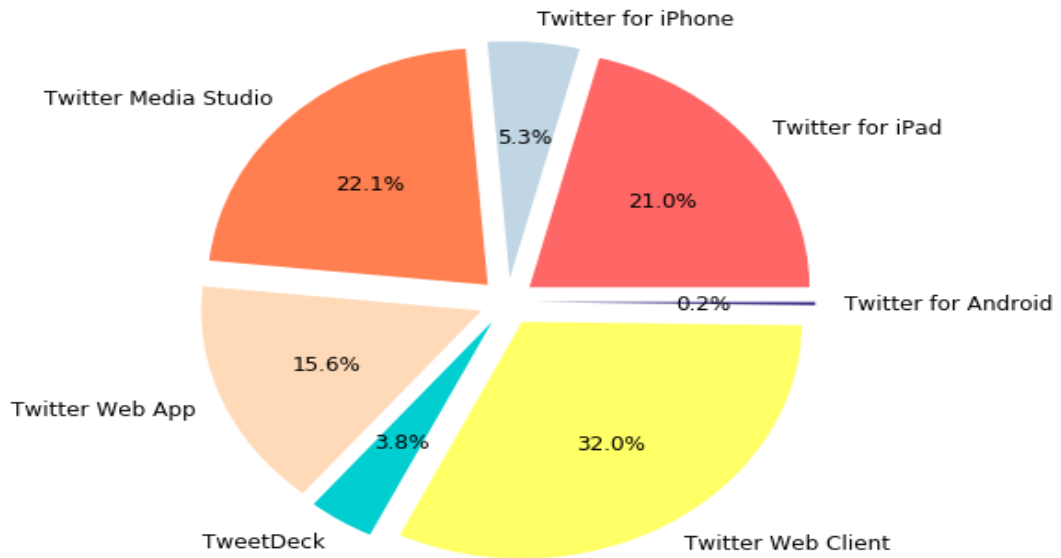
```
print(f" Data Available since {data.created_at.min()}")
print(f" Data Available upto {data.created_at.max()}")
```

Data Available since 2018-09-21 14:11:31
Data Available upto 2020-03-08 19:06:51

Şekil 3.109. @PoliceNationale kullanıcı adlı hesaptan çekilen veriye ait bilgiler- III

Pandas kütüphanesi aracılığıyla @PoliceNationale kullanıcı adlı hesaptan çekilen veriler üzerinde ön inceleme yapıldığında;

- Verinin 3141 satır ve 7 sütundan oluştuğu (Şekil 3.107, Şekil 3.108),
- Verinin sütun isimlerinin “id”, “created_at”, “text”, “followers_count”, “retweet_count”, “favourite_count” ve “source” olduğu (Şekil 3.108),
- Veri tipleri olarak “created_at”, “text” ve “source” sütunlarının object (nesne), diğer sütunların ise int64 (sayısal - integer) olarak tanımlandığı (Şekil 3.108),
- Veride 0' dan 3140'a kadar numaralandırılmış olan 3141 adet girdi bulunduğu ve her bir sütunda boş olmayan (non-null) girdi sayısının tanımlandığı, bu değerlerden de değeri olmayan sütun değerinin bulunmadığının anlaşıldığı (Şekil 3.108),
- @PoliceNationale adlı kullanıcıdan çekilen verilerin 21 Ağustos 2018'den 08 Mart 2020 tarihleri arasında yer aldığı (Şekil 3.109) anlaşılmaktadır.



Şekil 3.110. @PoliceNationale adlı kullanıcının veri paylaşımında kullandığı kaynaklar

@PoliceNationale adlı kullanıcının Twitter paylaşımı yaptığı cihazların kaynağı incelendiğinde; %5,3'ünün Iphone, %0,2'sinin Android, %32'sinin Twitter web istemcisi (web client), %15,6'sının Twitter web uygulaması, %21'inin Ipad, %22.1'inin Twitter medya stüdyosu ve %3.8'in de TweetDeck aracılığıyla olduğu şekil 3.110'da gösterilmektedir

```
print(f" Maximum number of retweets {data.retweet_count.max()}")
print(f" Maximum number of favorites {data.favourite_count.max()}")

Maximum number of retweets 17617
Maximum number of favorites 26755

data.loc[data['retweet_count']==17617,['text']].values
array([[ 'RT @EducationFrance: Une photo c'est perso, la partager c'est harceler
personne sans son con...' ]],
      dtype=object)

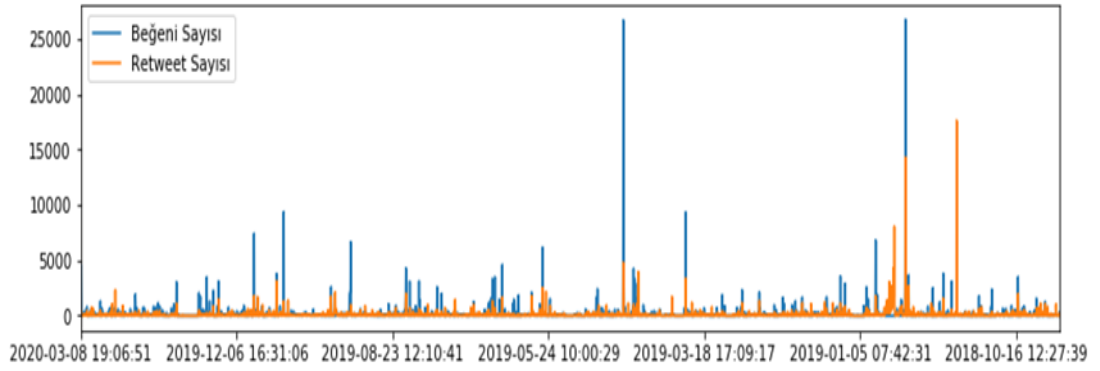
data.loc[data['retweet_count']==17617, ['created_at']].values
array([[ '2018-11-08 08:02:37' ]], dtype=object)

data.loc[data['favourite_count']==26755,['text']].values
array([[ 'Des appels à la violence contre des policiers circulent sur les réseau
des hommes et leurs proches qui sont visés alors que leur engagement pour assur
dtype=object)

data.loc[data['favourite_count']==26755,['created_at']].values
array([[ '2018-12-05 11:29:02' ]], dtype=object)
```

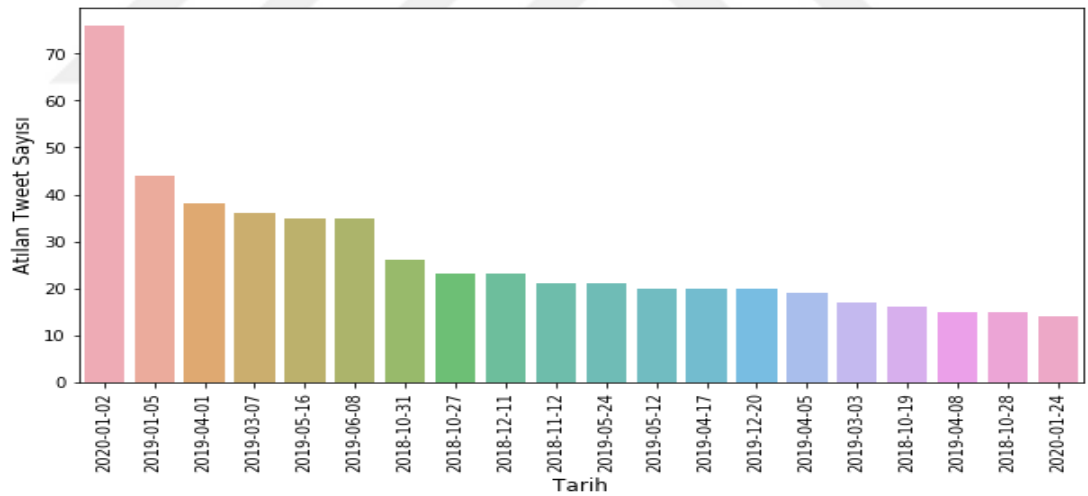
Şekil 3.111. Paylaşımların beğenilme ve retweet durumunun incelenmesi

Yapılan paylaşımlar şekil 3.111'de görüldüğü üzere incelendiğinde, 17617 tekrar paylaşım oranı ile en fazla retweet yapılan paylaşımın 08 Kasım 2018 tarihinde paylaşılan bir retweet mesajı (['RT @EducationFrance: Une photo c'est perso, la partager c'est harceler. Partager des images à caractère sexuel d'une personne sans son con...']) olduğu, 26755 beğeni ile de en fazla beğeni alan paylaşımın ('Des appels à la violence contre des policiers circulent sur les réseaux sociaux. Avec ces mots, ce sont des femmes, des hommes et leurs proches qui sont visés alors que leur engagement pour assurer la paix publique est total.') 11 Ağustos 2018 tarihinde yapıldığı anlaşılmaktadır. Ayrıca Twitter paylaşımlarının 530.223 kez beğenildiği (favorites), 308.659 kez tekrar paylaşıldığı (retweet) ve toplam takipçi sayısının ise 495.541 olduğu belirlenmiştir.



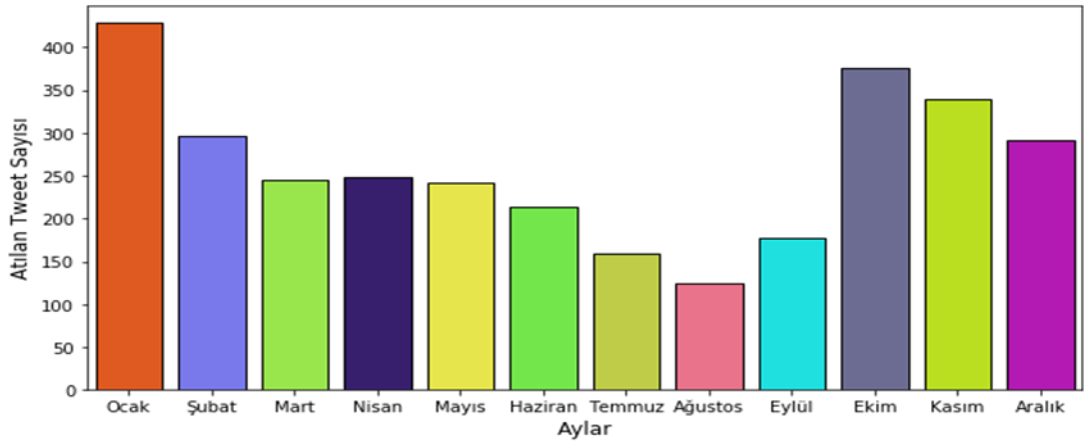
Şekil 3.112. @PoliceNationale adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu

@PoliceNationale adlı kullanıcının Twitter paylaşımlarının beğenilme ve retweet durumlarının tarih bazlı ortak grafiği şekil 3.112’de gösterilmektedir (Takipçi sayısının tarih bazında sabit olmasından dolayı “followers_count” değeri ile tweet kimliğini içeren “id” değerinde ilişkisel analiz incelenmemiştir.).



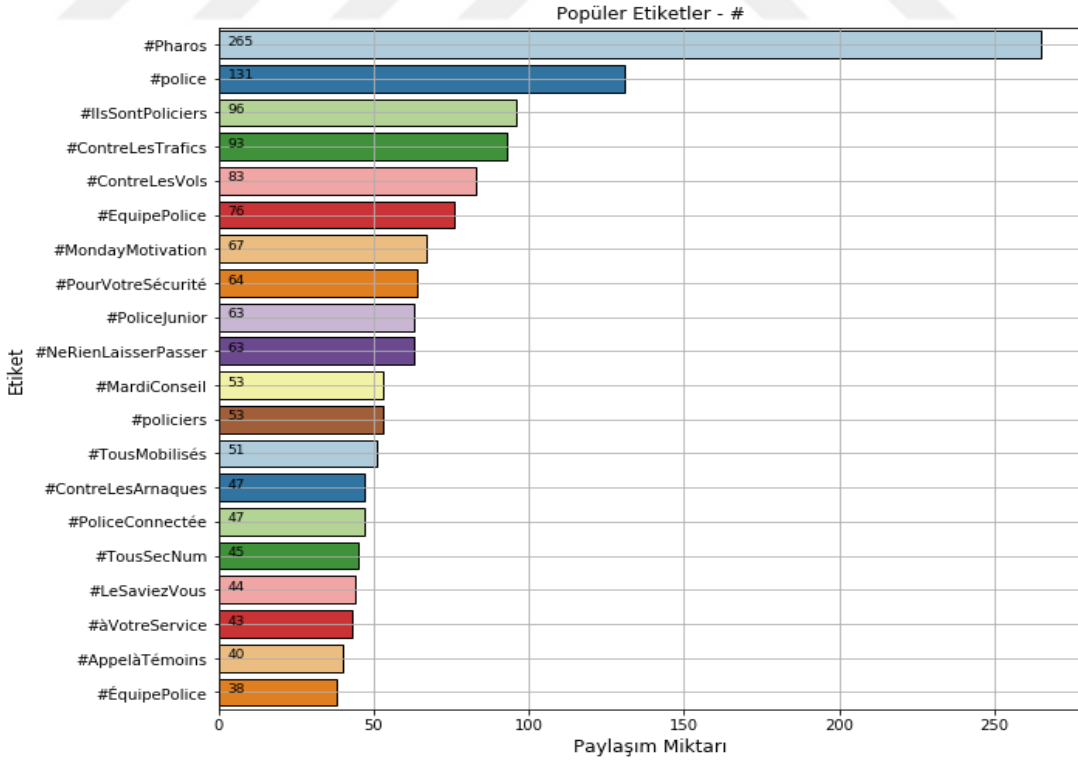
Şekil 3.113. Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı

Şekil 3.113 incelendiğinde aynı gün içerisinde sık paylaşımın yapıldığı tarihler görülmekte olup, örnek olarak en sık paylaşımın yapıldığı 02 Ocak 2020 tarihindeki 76 adet yapılan paylaşım incelendiğinde, paylaşımların bir grup gencin, başka bir genci dövdüğü video kapsamında ve bazı tekrar paylaşılan tweet’lerden oluştuğu görülmektedir.



Şekil 3.114. Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı

Twitter aracılığıyla sosyal medyadan yapılan paylaşımlara şekil 3.114’de olduğu üzere aylık dağılımına bakıldığında 428 Tweet ile ocak ayı en fazla sosyal medya paylaşımının yapıldığı ay iken, genel seyrinde paylaşımların düzenli aralıklarla yapıldığı gözlemlenmektedir.

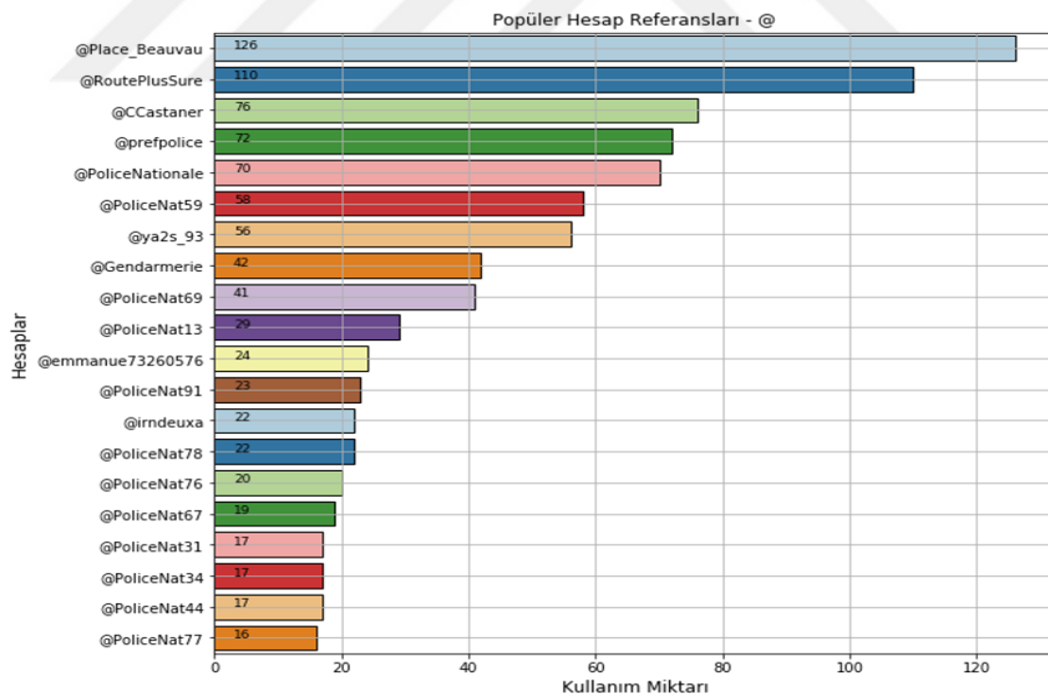


Şekil 3.115. En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)



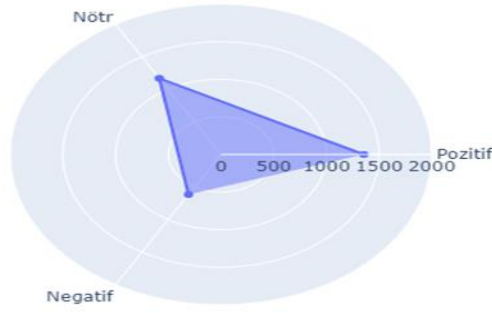
Şekil 3.116. En fazla kullanılan sosyal medya etiketlerinin (Hashtag - #) kelime bulutu

Yapılan paylaşımlarda en fazla kullanılan sosyal medya etiketlerinin 20'sine ilişkin kullanım sıklığı ile sık kullanılan sosyal medya etiketlerine ilişkin hazırlanan kelime bulutu (Maksimum kelime girdisi 1000 olarak alınmıştır.) şekil 3.115 ve 3.116'da gösterilmektedir.



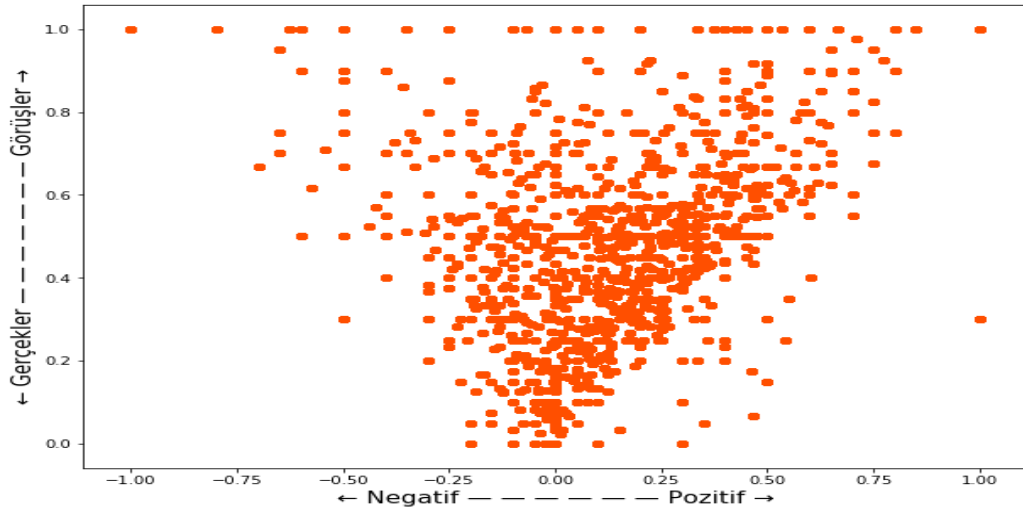
Şekil 3.117. En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)

Yapılan paylaşımlarda en fazla kullanılan sosyal medya hesaplarının 20'sine ilişkin kullanım sıklığı Şekil 3.117'de gösterilmektedir.



Şekil 3.118. Metnin duygusal durumunun radar grafiği

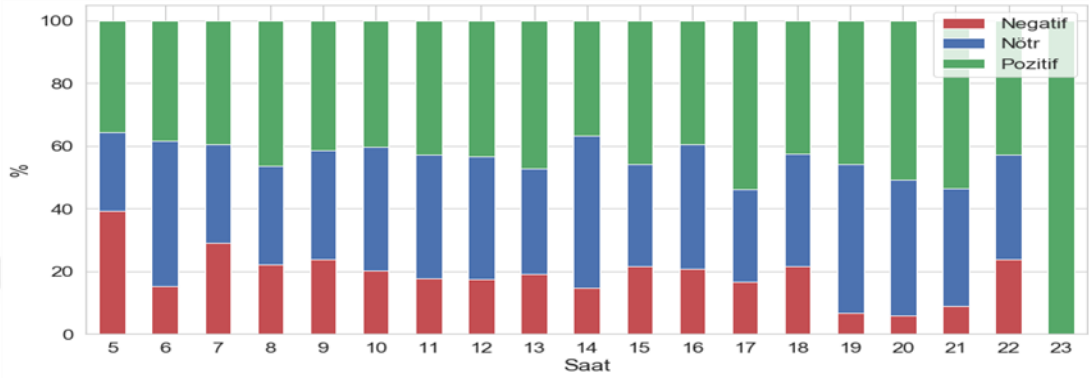
@poliziadistato kullanıcı hesabından yapılan paylaşım metinleri, metin ön işlemesine (text preprocessing) tabi tutularak tüm emojiler kaldırılmış ve Fransızca/İngilizce karakter farklılıklarından dolayı metindeki tüm harfler küçültülerek makine öğrenimine dayalı olarak Fransızca'dan İngilizce'ye çeviri yapılmıştır. Çeviri metni üzerinde yapılan duygusal dağılım ($0=Nötr$, $0>Negatif$, $0<Pozitif$) şekil 3.118'de görüldüğü üzere 1351 tweet ile pozitif mesajların çoğunlukta olduğu, 1173 tweet ile nötr mesajların pozitif mesajlarını izlediği ve son olarak da 617 tweet ile negatif skorlanan mesajların bulunduğu belirlenmiştir.



Şekil 3.119. Metnin duygusal durumunun öznellik analizi

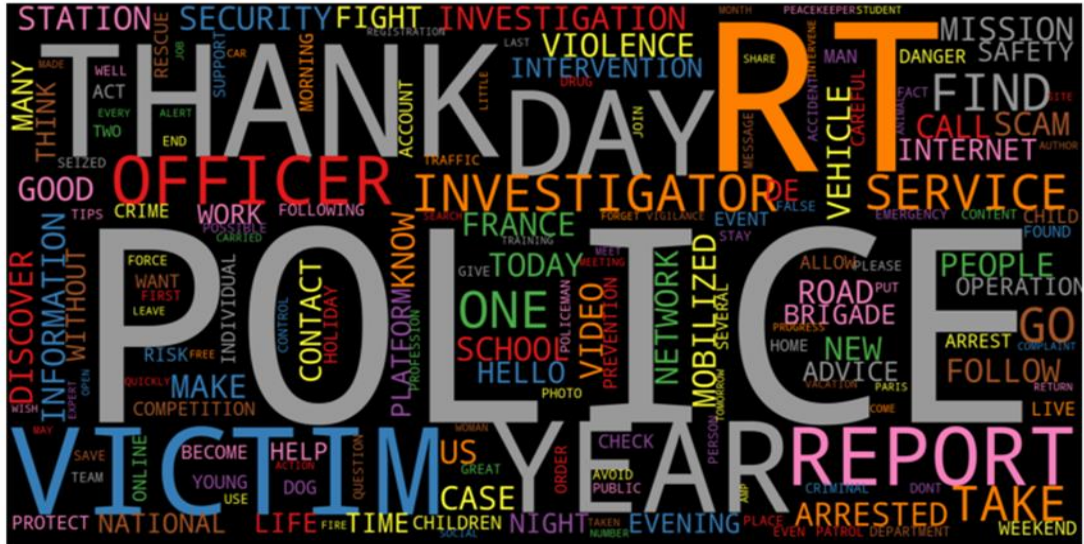
Elde edilen çeviri üzerinden kelime bazlı olarak, her kelimenin yarattığı duygusal durum ortaya konularak, her cümleye ait duygusal puan durumu, duygusal durumu ve ortaya konulan kanının öznellik ihtimali çıkarılmıştır. Metnin duygusal

durumu ile öznelliğini beraber analiz ettiğimizde ise, cümlelerin içinde yer alan her kelimenin skorlanması ile cümle skoru elde edilmiş olup, ana korpusun çoğunluğunun 0.5 öznellik bandını geçmediği şekil 3.119’da gösterilmektedir. Fakat az sayıda da olsa öznelliği 1 olarak yer alan 25 tweet için, negatif, nötr veya pozitif gibi değerlendirmelerde bulunulmuştur.



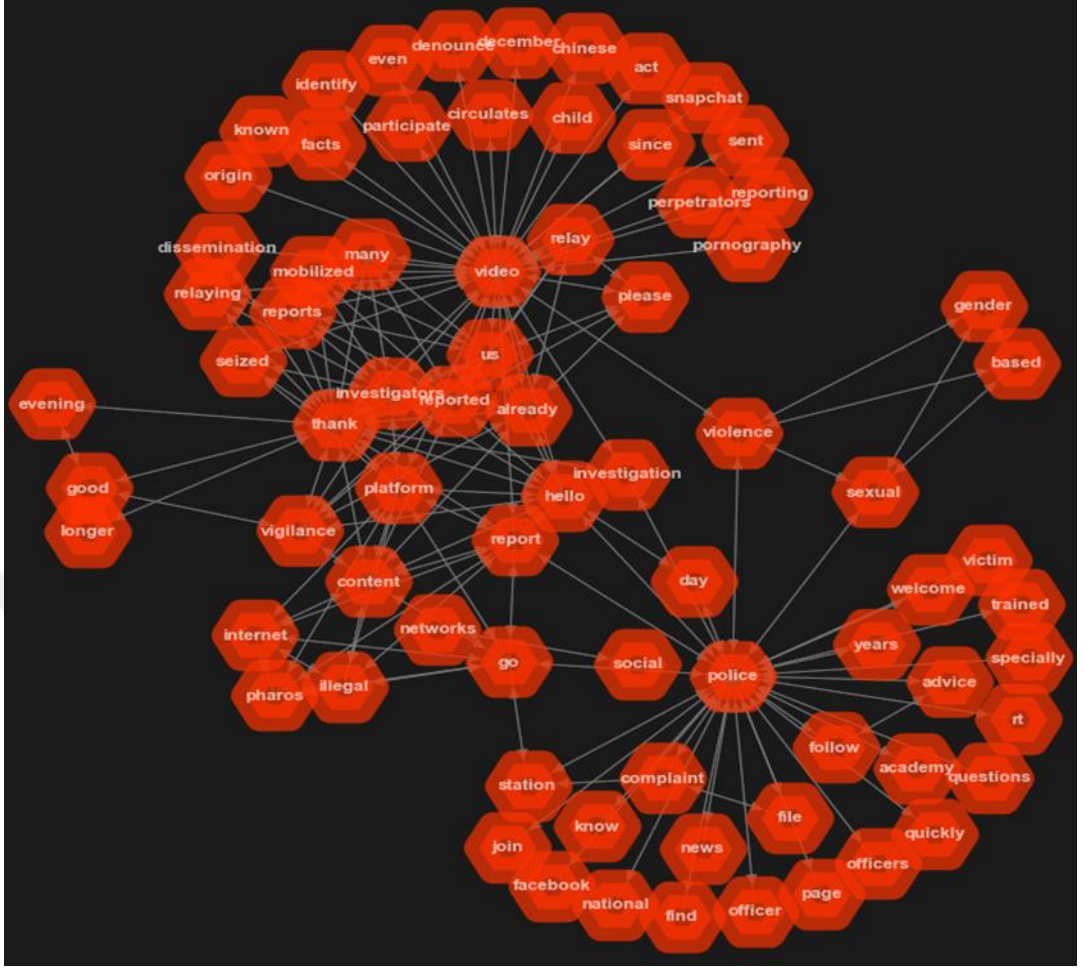
Şekil 3.120. Metnin duygusal durumunun saat bazlı incelenmesi

Paylaşımların yapıldığı zaman dilimi ile duygusal durumu arasındaki ilişki şekil 3.120’de gösterilmekte olup, incelenen korpusun yaklaşık 18 aya yakın süreyi kapsamasından dolayı zamanın da duygusal duruma etkisi olduğu söylenebilmektedir.



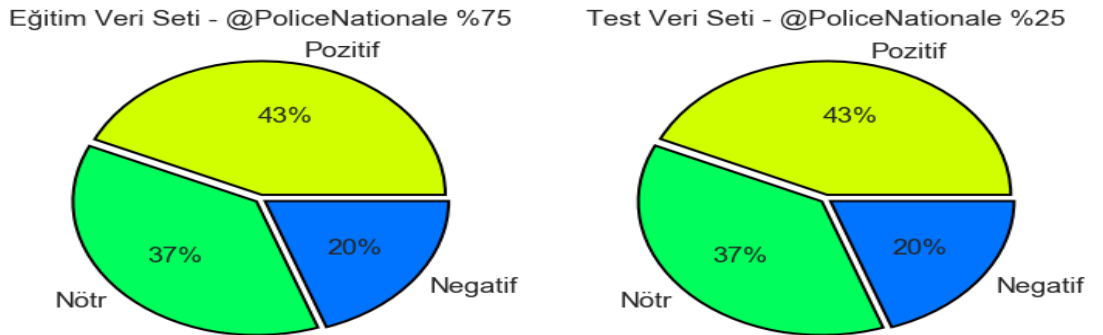
Şekil 3.121. Paylaşımlar arasında sık kullanılan ifadelerle ilgili kelime bulutu

Yapılan paylaşımlarda en sık kullanılan ifadeler bazında (# ve @ gibi ifadeler olmaksızın) şekil 3.121’deki kelime bulutunun elde edildiği görülmektedir.



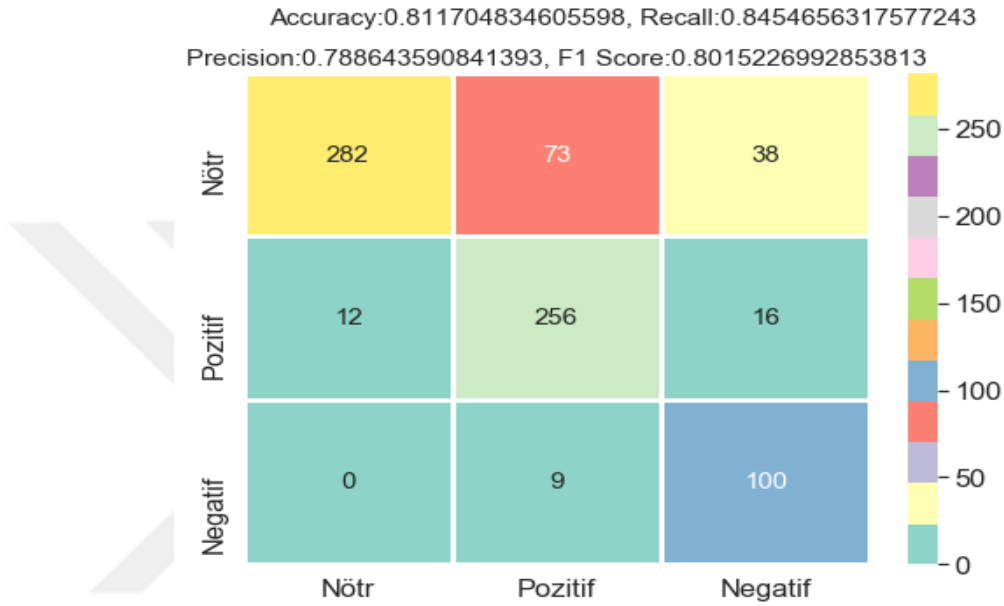
Şekil 3.122. Yapılan paylaşımlara ait kelime ağı analizi (50>Kelime Kombinasyonu)

Yapılan paylaşım metinleri BoW (Bag of words) metoduyla veri seti haline getirilmiştir. Söz konusu veri seti içerisinde de 50’den fazla kombinasyona sahip kelimelerin ağ (network) yapısı şekil 3.122’de gösterilmiştir.



Şekil 3.123. Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması

Textblob kütüphanesi aracılığıyla yapılan paylaşımların duygu durumları belirlenmiş olup, söz konusu tespitlerin kendi içinde tutarlı olup olmadığını incelemek adına makine öğrenimi modellerinden rastgele orman sınıflandırması (Random forest classification) ile veri seti şekil 3.123’de gösterildiği üzere test ve eğitim setine ayrılmıştır. Sınıflandırma için verinin %75’i eğitim, %25’i de test olarak ayrılmıştır.

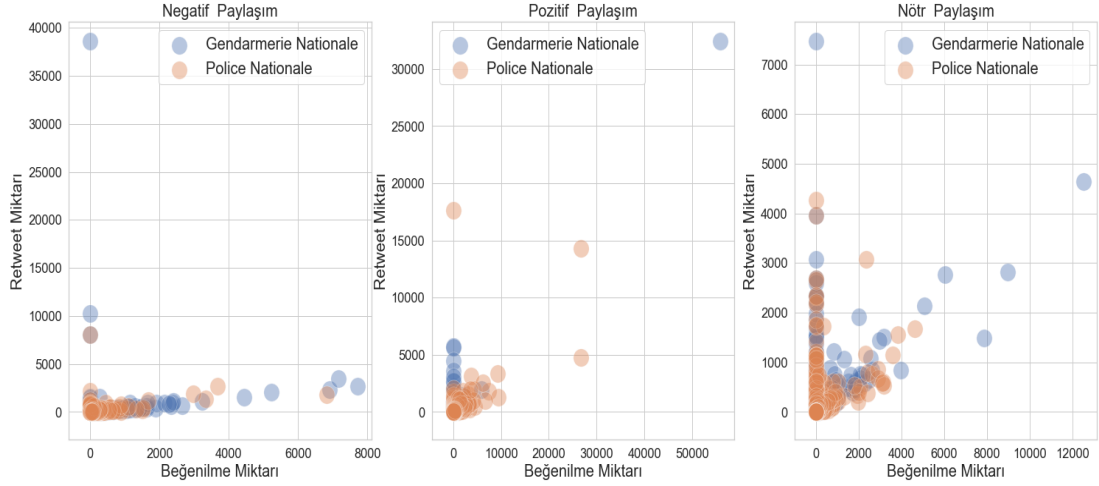


Şekil 3.124. Test verisinin duygusal matriksi

Makine öğrenimi sonrası test verileri üzerinde çıkan sonuca ilişkin matriks şekil 3.124’de görülmekte olup, yapılan öğrenme test verileri üzerinde %81.170 oranında doğruluk oranı sağlamıştır.

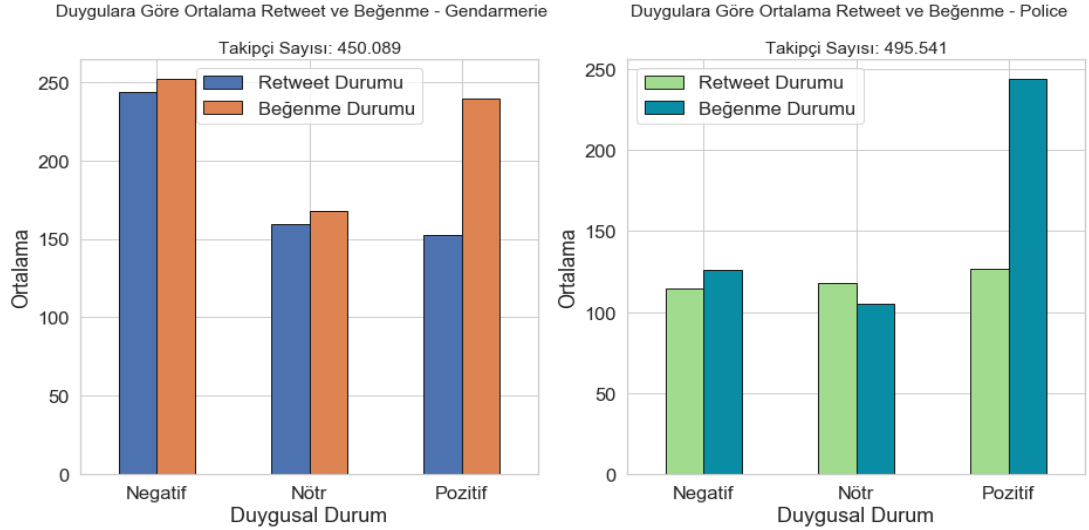
3.3.3. Gendarmerie Nationale ve Police Nationale Teşkilatlarının Mukayeseli Ortak Analizi

Fransa özelinde faaliyet gösteren kolluk teşkilatlarının topluma verdiği mesajları, mesajlarda hangi konu ve kelimeleri ağırlıkta kullandığı ile paylaştığı mesajların sosyal medyadaki etkisini mukayeseli olarak aşağıda incelenmektedir. Mukayese kapsamında korelasyon ve makine öğrenimi modellerinden karar destek vektörleri kullanılarak, tespitler görsel içerikle desteklenmektedir.



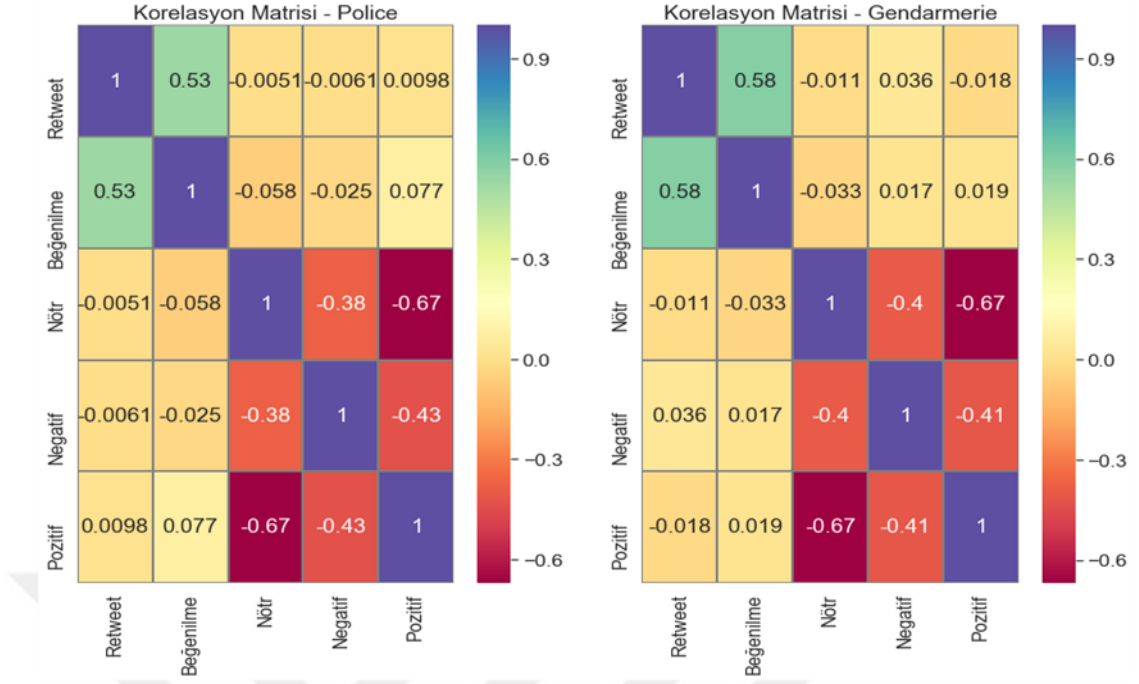
Şekil 3.125. Duygulara göre retweet ve beğenme durumu

Paylaşımlar içerisinde anomali tespiti yapmak maksadıyla şekil 3.125 incelendiğinde Gendarmerie özelinde bazı nötr ve negatif cümlelerde anomali gösterdiği, Police özelinde ise bazı pozitif cümlelerde özellikle retweet miktarının fazla olduğu gözlemlenmiştir. Kurumlar arası genel dağılımın birbirine yakın olduğu ve en fazla anomali gösteren dağılımın nötr paylaşımlarda yaşandığı belirlenmiştir.



Şekil 3.126. Duygulara göre ortalama retweet ve beğenme durumu

Paylaşımların genel ortalaması şekil 3.126'da görüldüğü üzere incelendiğinde, genel olarak tüm paylaşımlarda Gendarmerie'nin beğeni ve retweet ortalaması yüksek iken, özellikle pozitif paylaşımlarda Police'nin de ortalamasının Gendarmerie'ye yakın düzeyde olduğu gözlemlenmektedir.

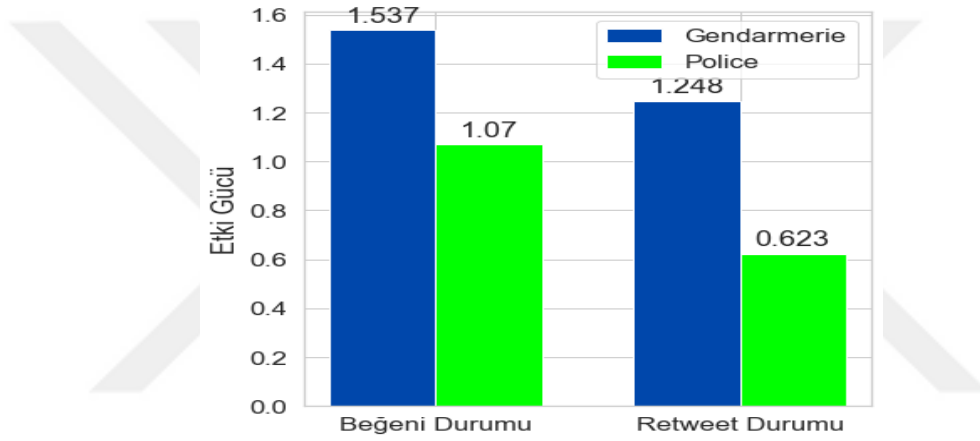


Şekil 3.127. Duygusal korelasyon matriksi

Duyguların, beğeni ve retweet üzerindeki gücünün korelasyon biçimindeki gösterimi matriks biçiminde şekil 3.127’de yapılmış olup, söz konusu verilerin analizinden;

- Beğeni ve Retweet arası ilişkinin Gendarmerie’de daha güçlü olduğu (Police Nationale: 0.53, Gendarmerie:0.58),
- Pozitif mesajlar kapsamında yapılan paylaşımları insanların beğenmesin Police’de daha yüksek ilişki düzeyinde olduğu (Police Nationale: 0.077, Gendarmerie:0.019),
- Pozitif mesajlar kapsamında yapılan paylaşımları insanların retweet etmesinin Police’de daha yüksek ilişki düzeyinde (Police Nationale: 0.0098, Gendarmerie: -0.018),
- Nötr mesajlar kapsamında yapılan paylaşımları insanların beğenmesin Gendarmerie’de daha yüksek ilişki düzeyinde olduğu (Police Nationale: -0.0058, Gendarmerie: -0.033),

- Nötr mesajlar kapsamında yapılan paylaşımları insanların retweet etmesinin Police’de daha yüksek ilişki düzeyinde olduğu (Police Nationale: -0.0051, Gendarmerie: -0.011),
- Negatif mesajlar kapsamında yapılan paylaşımları insanların beğenmesinin Gendarmerie’de daha yüksek ilişki düzeyinde olduğu (Police Nationale: -0.025, Gendarmerie:0.017),
- Negatif mesajlar kapsamında yapılan paylaşımları insanların retweet etmesinin Gendarmerie’de daha yüksek ilişki düzeyinde olduğu (Police Nationale: -0.0061, Gendarmerie:0.036) gözlemlenmiştir.



Şekil 3.128. Gendarmerie ve Police’nin etki gücü

Yapılan paylaşımların takipçileri oranında hangi derecede paylaşıldığının ve retweet edildiğinin ortaya konması maksadıyla etki gücü hesaplaması yapılmış, elde edilen sonuçlar da şekil 3.128’de gösterilmektedir. Etki gücü hesaplaması yapılırken, toplam beğeni sayısı/toplam takipçi sayısı ve toplam retweet sayısı/toplam takipçi sayısı formülasyonu kullanmıştır. Buradaki amaç her takipçinin ortalama ne kadar mesajı beğendiği ve retweet ettiği tespitiyle, mesajların ne kadar güçlü olduğunu karşılaştırmalı olarak ortaya koymaktır. Bu yapılan analiz sonucunda ise Gendarmerie’nin paylaşımlarının Police’ye kıyasla etki gücünün kısmi daha yüksek olduğu anlaşılmaktadır. Bu sebeple de aradaki farkın sebebini daha iyi anlayabilmek adına makine öğrenimine dayalı paylaşım yaptıkları konuların neler olduğunun ortaya konması gerekmektedir.

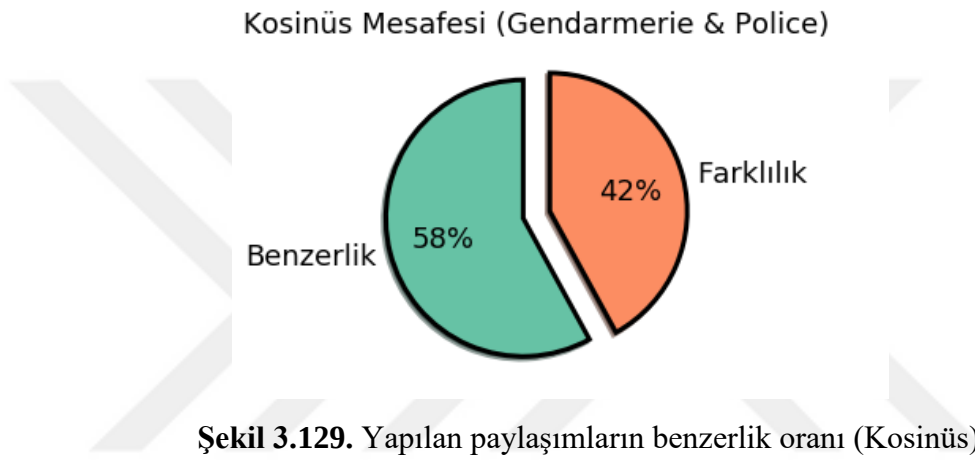
Çizelge 3.3. Gendarmerie ve Police'nin paylaşım yaptığı konuların LDA (Latent Dirichlet Allocation) analizi ve konular arası mesafe haritalandırması

Konu	Gendarmerie LDA Analizi	Police LDA Analizi
Mesafe Haritası		
1	('gendarmerie', 259.0277525888682) ('gendarme', 115.03920371000929) ('discover', 103.29101058687141) ('national', 64.41071389305971) ('new', 57.79551894492079) ('daily', 57.65445199292528) ('year', 56.188512368694575) ('school', 55.63006801802842) ('intervention', 52.08106019404698) ('mission', 52.067937235522706)	('police', 280.6183456937285) ('tip', 85.40297551432228) ('officer', 74.92972983472929) ('case', 66.23864776669524) ('prepare', 63.60347355752151) ('rescue', 59.00705964068796) ('scam', 58.982927210910674) ('academy', 58.510400834500786) ('new', 58.06290708571286) ('day', 53.54781809627977)
2	('arrest', 125.7941427677724) ('year', 87.90871525132498) ('individual', 61.1839705583431) ('theft', 58.10778063273236) ('group', 55.835221120305185) ('investigation', 54.881823632197445) ('gendarme', 49.27359948907761) ('criminal', 47.68839468986068) ('center', 44.70219965070011) ('state', 43.251853816244115)	('police', 386.92230412808516) ('arrest', 78.96501944035158) ('join', 77.1530020255498) ('network', 63.71281716871592) ('meet', 62.39933039126866) ('welcome', 61.51978531586831) ('follow', 61.11165234307347) ('officer', 52.351910415114304) ('news', 49.4101481106801) ('miss', 48.18965438980558)
3	('department', 92.94366006088342) ('general', 92.56783365404996) ('place', 78.77261792104088) ('lizurey', 61.98259696463784) ('national', 54.34080491599996) ('family', 53.10211297894315) ('thank', 51.12580823978127) ('come', 50.84334171457152) ('richard', 47.74976883655716) ('army', 47.73557884824715)	('report', 292.2527474603495) ('content', 186.73506638396768) ('internet', 162.73894564565958) ('illegal', 143.35889677612735) ('violence', 113.82474593886002) ('platform', 102.97675641568209) ('child', 102.32096573379448) ('sexual', 91.39251192937351) ('victim', 74.28349538231464) ('police', 64.58710953319712)
4	('day', 125.94607375136265) ('gendarme', 100.98680419756744) ('child', 72.15022262928852) ('today', 63.63261924528271) ('police', 61.73393041612808) ('night', 60.48191571060306) ('force', 57.95241351714881) ('student', 56.61854266176273) ('france', 54.11651886431158) ('control', 53.0368885002064)	('police', 130.63127270147115) ('day', 97.57963956978551) ('make', 84.15042154545795) ('emergency', 72.95179074188232) ('know', 64.474883468214) ('hello', 63.98583105419176) ('service', 61.82870892912812) ('open', 51.88541793669533) ('policeman', 41.896266404051154) ('specially', 41.189774713170785)

Çizelge 3.3. Devam

Konu	Jandarma LDA Analizi	Emniyet LDA Analizi
5	('traffic', 84.78655497702029) ('seize', 84.38711165089553) ('arrest', 82.3288036987594) ('end', 73.56231861529176) ('individual', 62.65842858902735) ('dismantle', 53.99803553223518) ('cannabis', 47.38524144803061) ('area', 41.467447822809014) ('operation', 40.13511469039132) ('gendarme', 39.83607162911481)	('video', 476.1811057115493) ('report', 472.29142714056525) ('thank', 291.4453151453306) ('investigator', 281.8486207170219) ('relay', 210.20144964765132) ('mobilize', 133.01073143272106) ('longer', 49.23800832494447) ('violence', 45.60884201016255) ('investigation', 42.887443653323515) ('identify', 39.436895760600706)
6	('advice', 71.04499360744957) ('make', 66.10289404473916) ('stay', 60.41931733497764) ('return', 56.84474233433957) ('leave', 53.39569389713648) ('protect', 49.44340906881808) ('start', 46.96433496887257) ('minister', 46.25180026976875) ('good', 41.90704292047283) ('information', 37.58529589571824)	('hello', 243.11447022183097) ('thank', 180.18514849601416) ('vigilance', 172.0434875001326) ('platform', 161.52030774798564) ('pharos', 137.11980877654247) ('good', 128.57327317934445) ('station', 116.26539674737668) ('police', 103.1720824679623) ('evening', 75.57866482426849) ('complaint', 72.95250676557694)
7	('follow', 123.35663328099116) ('road', 109.23513493139696) ('security', 104.52553551348856) ('amp', 75.65196418472466) ('vehicle', 63.08044259299111) ('car', 50.60586614684634) ('death', 47.72986758677078) ('drive', 42.47392429595646) ('driver', 41.06533461657607) ('comrade', 40.96006116292867)	('participate', 72.0119239127824) ('online', 61.18492862374128) ('use', 57.12259868742936) ('know', 56.89884291764002) ('dissemination', 53.72374972864158) ('end', 49.80558402870386) ('peace', 45.929591797491234) ('security', 45.608960394975774) ('advice', 43.57599546907767) ('avoid', 43.16365190384706)
8	('colonel', 69.02210208651543) ('morning', 64.30986037426376) ('population', 41.845352067752074) ('meet', 38.417098767961335) ('mountain', 38.30642924091985) ('use', 37.84981380533085) ('live', 36.522659949667215) ('summer', 35.899909571532746) ('come', 34.88145986247156) ('rescue', 34.51677008084602)	('police', 125.63054651760531) ('officer', 67.51318868987784) ('year', 66.55605540840533) ('discover', 59.74292305241493) ('crime', 58.67084950042754) ('help', 48.00005152581299) ('profession', 43.179583527300615) ('animal', 39.477922471679925) ('work', 38.42043831700504) ('action', 35.289653204694346)
9	('gendarme', 95.79747027704998) ('safety', 65.33008336184935) ('mobilize', 63.53542394302922) ('time', 60.849185041357366) ('respect', 59.69571095666065) ('allow', 57.346347238250104) ('operation', 48.7536992696439) ('little', 35.84029421805643) ('man', 32.634551475173815) ('action', 30.36694885228237)	('site', 54.84984548506399) ('sure', 51.171600099215325) ('brigade', 47.651796670283545) ('risk', 44.39333081982211) ('great', 44.06128761391639) ('dont', 36.18049859733037) ('forget', 36.15550356082883) ('make', 33.23414026101942) ('prevention', 29.999015153528987) ('weekend', 29.4873010342574)
10	('good', 73.59144872137011) ('team', 49.72316404043434) ('world', 49.41626670517922) ('near', 43.866388559715865) ('open', 40.36598036440977) ('military', 37.87656204100797) ('officer', 32.57411917045618) ('congratulation', 29.108238971783198) ('champion', 27.987127450842017) ('competition', 27.897020467429357)	('congratulation', 74.42508707563303) ('safety', 72.56494970867527) ('come', 70.68680432567629) ('road', 68.63071009425362) ('behavior', 56.90444396260361) ('night', 50.17297022260932) ('event', 44.97841286614543) ('question', 44.37015541148094) ('answer', 41.89960782043787) ('ensure', 35.913331008324604)

Çizelge 3.3’de konu modellemeye aktif olarak kullanılan LDA modellemesiyle, kurumların paylaşımları 10 konu başlığı altında incelenmiştir. 10 konu içinde de her konudan en yüksek ilişki düzeyine sahip 10 kelime örneklem olarak çizelgede gösterilmiştir. Her konunun vektörel düzeyde konular arası mesafe haritası çizelgenin başında verilmiş olup, kelimelerin yanında yer alan ifadeler de vektörel düzeyde konu içinde ilişkisini ifade etmektedir. Bu kapsamda konu bazında yapılan incelemelerde, paylaşım yapılan konuların ve vektörel düzlemdeki mesafelerinin ayrıştığı rahatlıkla görülmektedir.



Paylaşım yapılan konuların vektörel dağılımlarındaki farklılıkların, metin korpuslarının geneline yayılıp tespit edilmesi maksadıyla, öncelikle daha net bir sonucun çıkması için kelimelerin kökleri çözümlenerek (lemmatization) işlem yapılmıştır. Metin içindeki kelime (token) yapılarının her biri vektörel düzeye indirgenerek, şekil 3.129’da görüldüğü üzere kosinüs mesafesi yöntemi ile paylaşım yapılan metinlerin benzerlik oranları %58 olarak bulunmuştur.

3.4. İspanya Özelinde Kolluk Teşkilatlarının Paylaşımlarının İncelenmesi

Guardia Civil ve Policía Nacional’ın resmi paylaşımlarından alınan örneklem üzerinden keşfedici veri analizi yapılarak, sosyal medya kullanım alışkanlıkları üzerinden duyarlılık analizi ve söz konusu duyarlılığın makine öğrenimine dayalı tahminlemesi yapılmıştır.

3.4.1. Guardia Civil'in Paylaşımlarının İncelenmesi

İspanya'da faaliyet gösteren jandarma türü kolluk gücünün resmi hesabı olan “Guardia Civil” başlıklı ve “@guardiacivil” kullanıcı adlı Twitter hesabından 09 Mart 2020 tarihinde herkese açık olarak yapılan paylaşımlar Tweepy kütüphanesi aracılığıyla Python programlama dili kullanılarak, csv formatında çekilmiştir.



Şekil 3.130. Guardia Civil resmi Twitter sayfasının ekran resmi

	id	created_at	text	followers_count	retweet_count	favourite_count	source
0	1238773889288306889	2020-03-08 22:00:00	¿Necesitas ayuda para proteger a los peques en...	1438887	33	93	TweetDeck
1	1238758591828381697	2020-03-08 21:00:00	Hoy también queremos felicitar por su Patrón a...	1438887	289	1180	TweetDeck
2	1238751040373854208	2020-03-08 20:30:00	No podemos pasar sin felicitar por su #patrón,...	1438887	81	445	TweetDeck
3	1238743490710560789	2020-03-08 20:00:00	Los servicios humanitarios fueron un pilar bas...	1438887	103	478	TweetDeck
4	1238739714897432576	2020-03-08 19:45:00	Urgente!!! En esta es Diana y ha desaparecido ...	1438887	1209	149	TweetDeck

Şekil 3.131. @guardiacivil kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler – I

```
print(data.info())
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3243 entries, 0 to 3242
Data columns (total 7 columns):
id                3243 non-null int64
created_at        3243 non-null object
text              3243 non-null object
followers_count   3243 non-null int64
retweet_count     3243 non-null int64
favourite_count   3243 non-null int64
source            3243 non-null object
dtypes: int64(4), object(3)
memory usage: 177.5+ KB
None
```

Şekil 3.132. @guardiacivil kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler –II

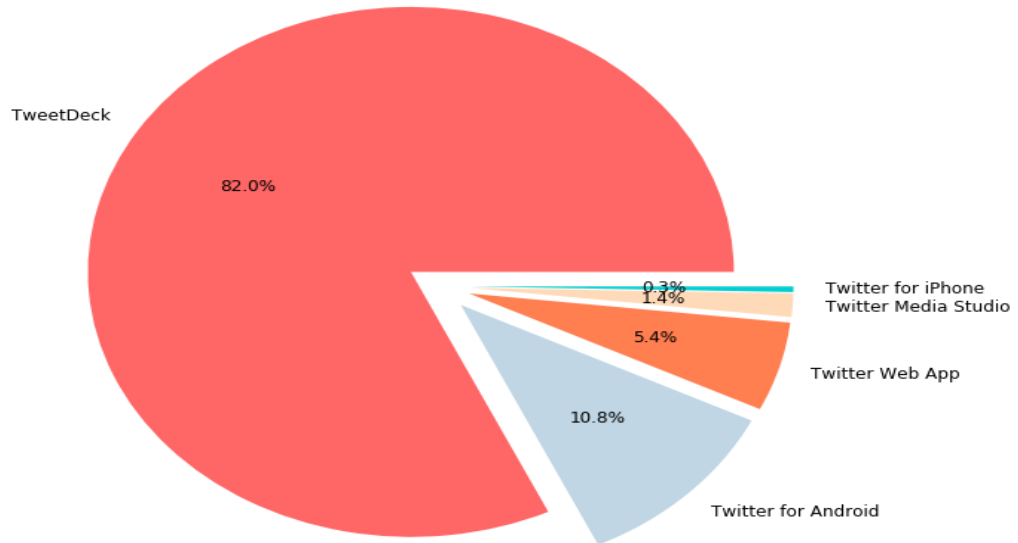
```
print(f" Data Available since {data.created_at.min()}")
print(f" Data Available upto {data.created_at.max()}")
```

Data Available since 2019-09-16 12:10:00
Data Available upto 2020-03-08 22:00:00

Şekil 3.133. @guardiacivil kullanıcı adlı hesaptan çekilen veriye ait ön bilgiler –III

Pandas kütüphanesi aracılığıyla @guardiacivil kullanıcı adlı hesaptan çekilen veriler üzerinde ön inceleme yapıldığında;

- Verinin 3243 satır ve 7 sütundan oluştuğu (Şekil 3.131, Şekil 3.132),
- Verinin sütun isimlerinin “id”, “created_at”, “text”, “followers_count”, “retweet_count”, “favourite_count” ve “source” olduğu (Şekil 3.132),
- Veri tipleri olarak “created_at”, “text” ve “source” sütunlarının object (nesne), diğer sütunların ise int64 (sayısal - integer) olarak tanımlandığı (Şekil 3.132),
- Veride 0' dan 3242'ye kadar numaralandırılmış olan 3243 adet girdi bulunduğu ve her bir sütunda boş olmayan (non-null) girdi sayısının tanımlandığı, bu değerlerden de değeri olmayan sütun değerinin bulunmadığının anlaşıldığı (Şekil 3.132),
- @guardiacivil adlı kullanıcıdan çekilen verilerin 16 Eylül 2019'dan 08 Mart 2020 tarihleri arasında yer aldığı (Şekil 3.133) anlaşılmaktadır.



Şekil 3.134. @guardiacivil adlı kullanıcının veri paylaşımında kullandığı kaynaklar

@guardiacivil adlı kullanıcının Twitter paylaşımı yaptığı cihazların kaynağı incelendiğinde; %82'sinin TweetDeck, %10,8'inin Android, %0,3'ünün Iphone, %5,4'ünün twitter web uygulaması ve %1,4'ünün de twitter medya stüdyosu aracılığıyla olduğu şekil Şekil 3.134'de gösterilmektedir.

```
print(f" Maximum number of retweets {data.retweet_count.max()}")
print(f" Maximum number of favorites {data.favourite_count.max()}")

Maximum number of retweets 16993
Maximum number of favorites 10617

data.loc[data['retweet_count']==16993,['text']].values

array([[ 'RT @policia: ► ¡MUCHO OJO, por favor, retuitea para que tus seguidores e
ción de billetes falsos en t...' ]],
      dtype=object)

data.loc[data['retweet_count']==16993, ['created_at']].values

array([[ '2019-10-03 14:07:58' ]], dtype=object)

data.loc[data['favourite_count']==10617,['text']].values

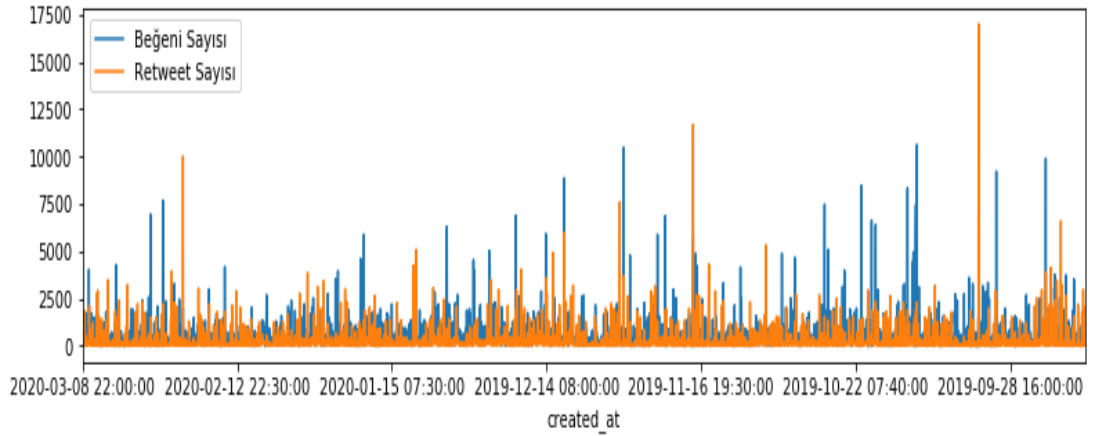
array([[ 'A ti, por estar ahí, sólo podemos ofrecerte una palabra de todo corazón.
vil\r\n\r\nhttps://t.co/anGR7xNm5y' ]],
      dtype=object)

data.loc[data['favourite_count']==10617,['created_at']].values

array([[ '2019-10-12 06:00:00' ]], dtype=object)
```

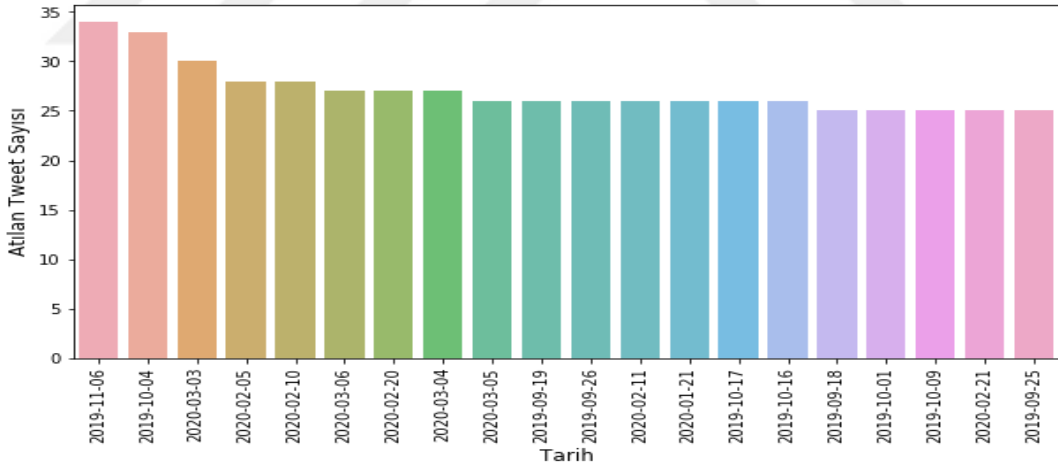
Şekil 3.135. Paylaşımların beğenilme ve tekrar paylaşılma durumunun incelenmesi

Yapılan paylaşımlar Şekil 3.135'de görüldüğü üzere incelendiğinde, 16993 tekrar paylaşım oranı ile en fazla retweet yapılan paylaşımın 03 Ekim 2019 tarihinde paylaşılan yine bir retweet mesajı (['RT @policia: ► ¡MUCHO OJO, por favor, retuitea para que tus seguidores estén al tanto! Hemos detectado la aparición de billetes falsos en t...']) olduğu, 10617 beğeni ile de en fazla beğeni alan paylaşımın ('A ti, por estar ahí, sólo podemos ofrecerte una palabra de todo corazón. GRACIAS. #120 #PatronaGuardiaCivil https://t.co/anGR7xNm5y') 12 Ekim 2019 tarihinde yapıldığı anlaşılmaktadır. Ayrıca Twitter paylaşımlarının 1.707.176 kez beğenildiği (favorites), 1.068.453 kez tekrar paylaşıldığı (retweet) ve toplam takipçi sayısının ise 1.438.667 olduğu belirlenmiştir.



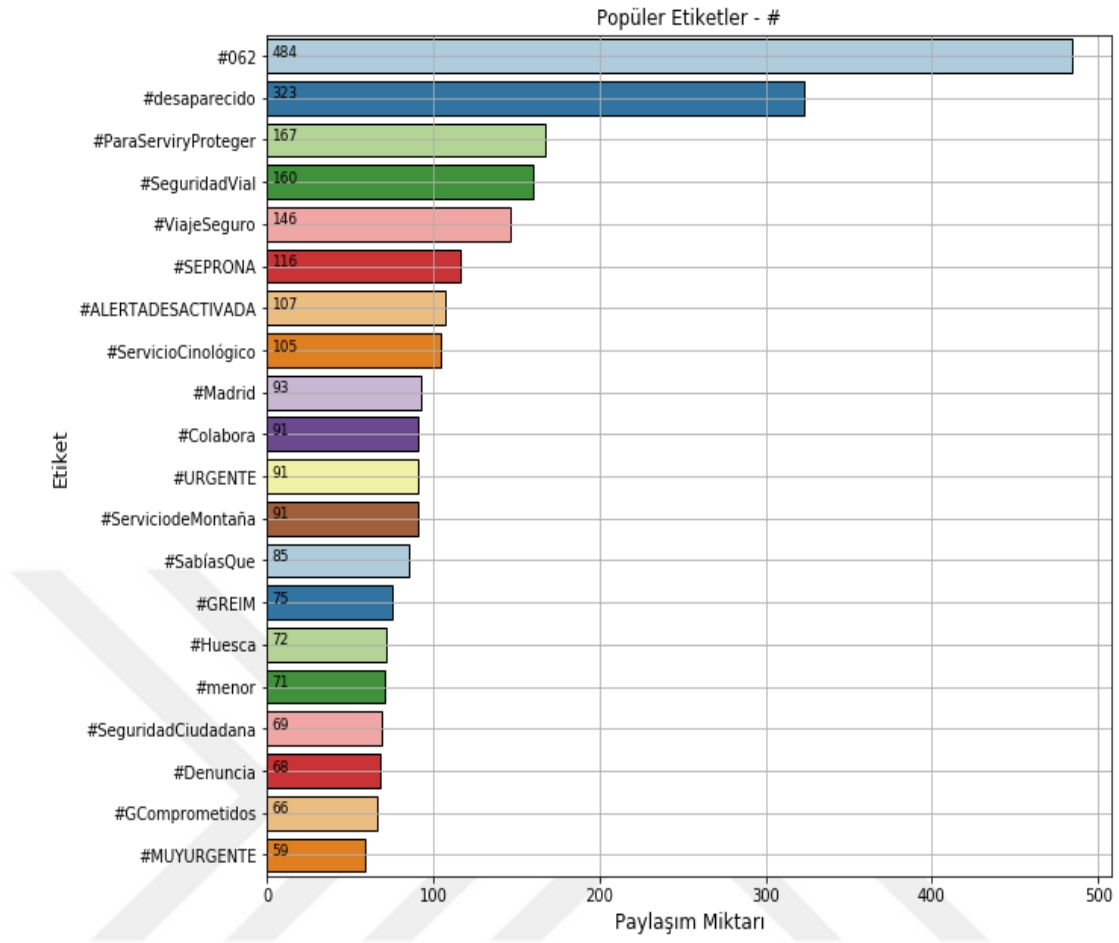
Şekil 3.136. @guardiacivil adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu

@guardiacivil adlı kullanıcının twitter paylaşımlarının beğenilme ve retweet durumlarının tarih bazlı ortak grafiği şekil 3.136’da gösterilmektedir (Takipçi sayısının tarih bazında sabit olmasından dolayı “followers_count” değeri ile tweet kimliğini içeren “id” değerinde ilişkisel analiz incelenmemiştir.).



Şekil 3.137. Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı

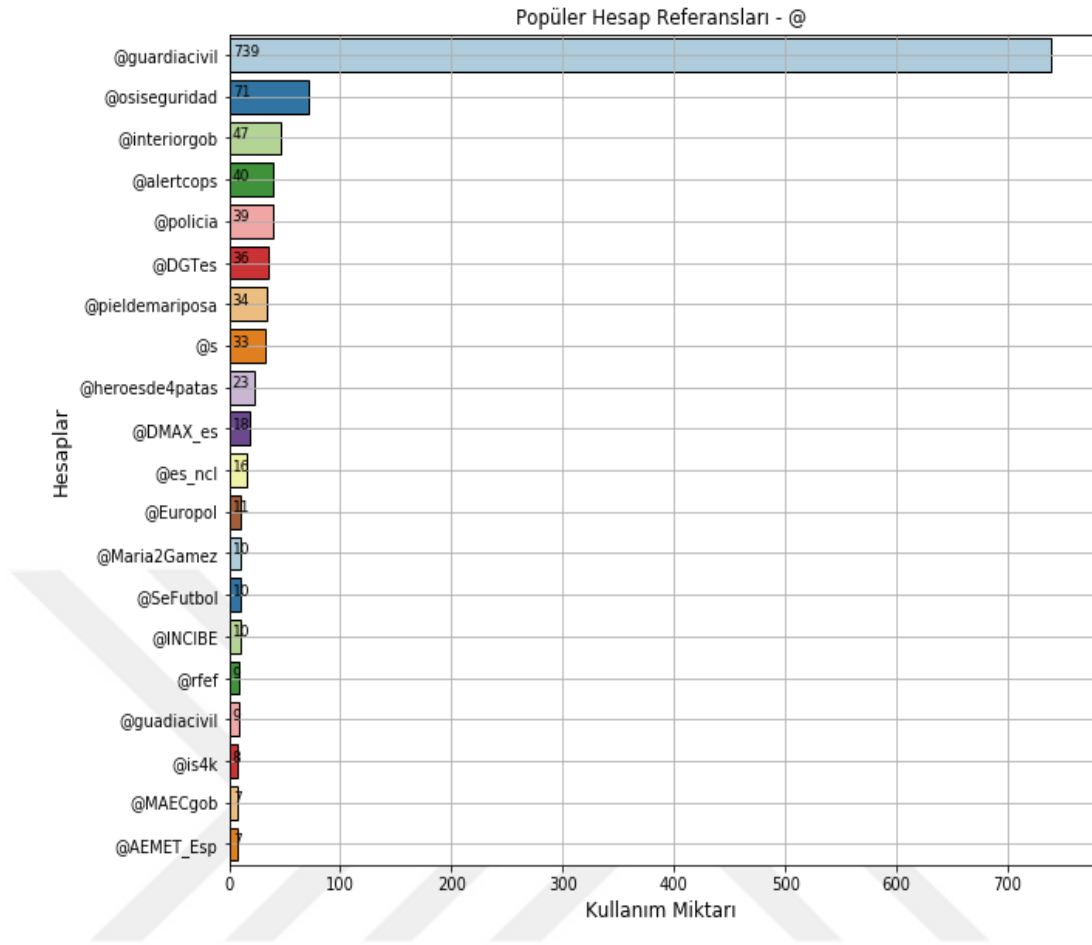
Şekil 3.137 incelendiğinde aynı gün içerisinde sık paylaşımın yapıldığı tarihler görülmekte olup, örnek olarak en sık paylaşımın yapıldığı 06 Kasım 2019 tarihindeki 34 adet yapılan paylaşım incelendiğinde, halka birbirinden farklı mesajlar veren çeşitli tweet’lerden oluştuğu görülmektedir.



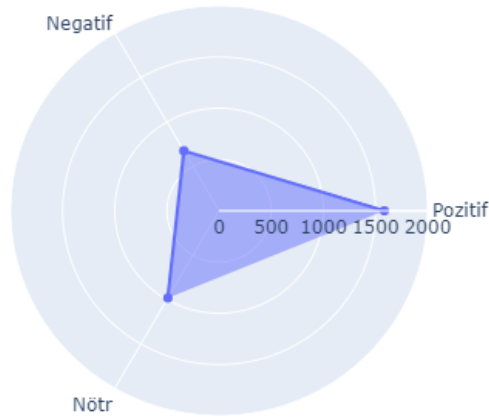
Şekil 3.140. En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)

Yapılan paylaşımlarda en fazla kullanılan sosyal medya etiketlerinin 20'sine ilişkin kullanım sıklığı ile sık kullanılan sosyal medya etiketlerine ilişkin hazırlanan kelime bulutu (Maksimum kelime girdisi 1000 olarak alınmıştır.) Şekil 3.139 ve 3.140'da gösterilmektedir.

Yapılan paylaşımlarda en fazla kullanılan sosyal medya hesaplarının 20'sine ilişkin kullanım sıklığı Şekil 3.141'de gösterilmektedir. Ayrıca @guardiacivil kullanıcı hesabından yapılan paylaşım metinleri, metin ön işlemesine (text preprocessing) tabi tutularak tüm emojiler kaldırılmış ve İspanyolca/İngilizce karakter farklılıklarından dolayı metindeki tüm harfler küçültülerek makine öğrenimine dayalı olarak İspanyolca'dan İngilizce'ye çeviri yapılmıştır.



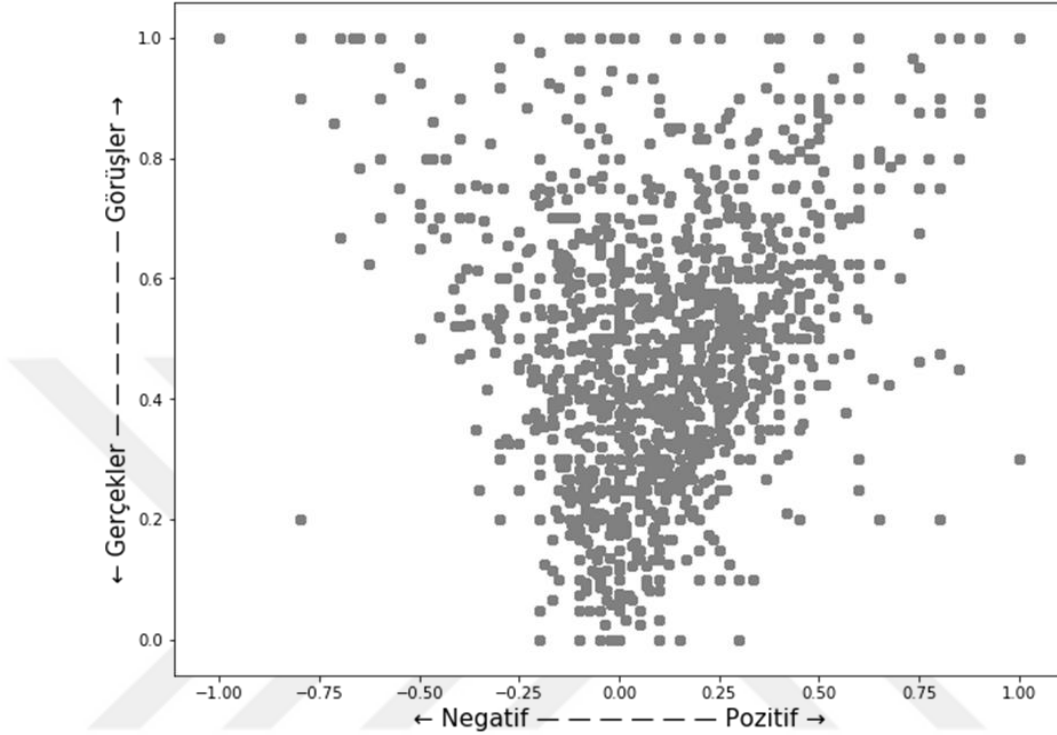
Şekil 3.141. En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)



Şekil 3.142. Metnin duygusal durumunun radar grafiği

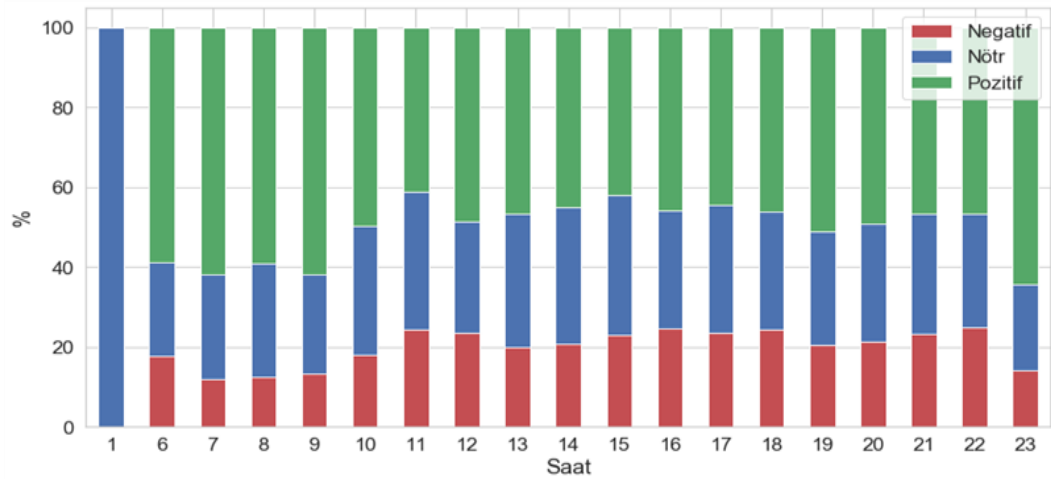
Çeviri metni üzerinde yapılan duygusal dağılım (0=Nötr, 0>Negatif, 0<Pozitif)
Şekil 3.142’da görüldüğü üzere 1580 tweet ile pozitif mesajların çoğunlukta olduğu,

984 tweet ile nötr mesajların pozitif mesajları izlediği ve son olarak da 679 tweet ile negatif skorlanan mesajların bulunduğu belirlenmiştir.



Şekil 3.143. Metnin duygusal durumunun öznellikle analizi

Elde edilen çeviri üzerinden kelime bazlı olarak, her kelimenin yarattığı duygusal durum ortaya konularak, her cümleye ait duygusal puan durumu, duygusal durumu ve ortaya konulan kanının öznellik ihtimali çıkarılmıştır. Metnin duygusal durumu ile öznelliğini beraber analiz ettiğimizde ise, cümlelerin içinde yer alan her kelimenin skorlanması ile cümle skoru elde edilmiş olup, ana korpusun çoğunluğunun 0.5 öznellik bandını geçmediği Şekil 3.143’de gösterilmektedir. Fakat az sayıda da olsa öznelliği 1 olarak yer alan 25 tweet için, negatif, nötr veya pozitif gibi değerlendirmelerde bulunulmuştur.



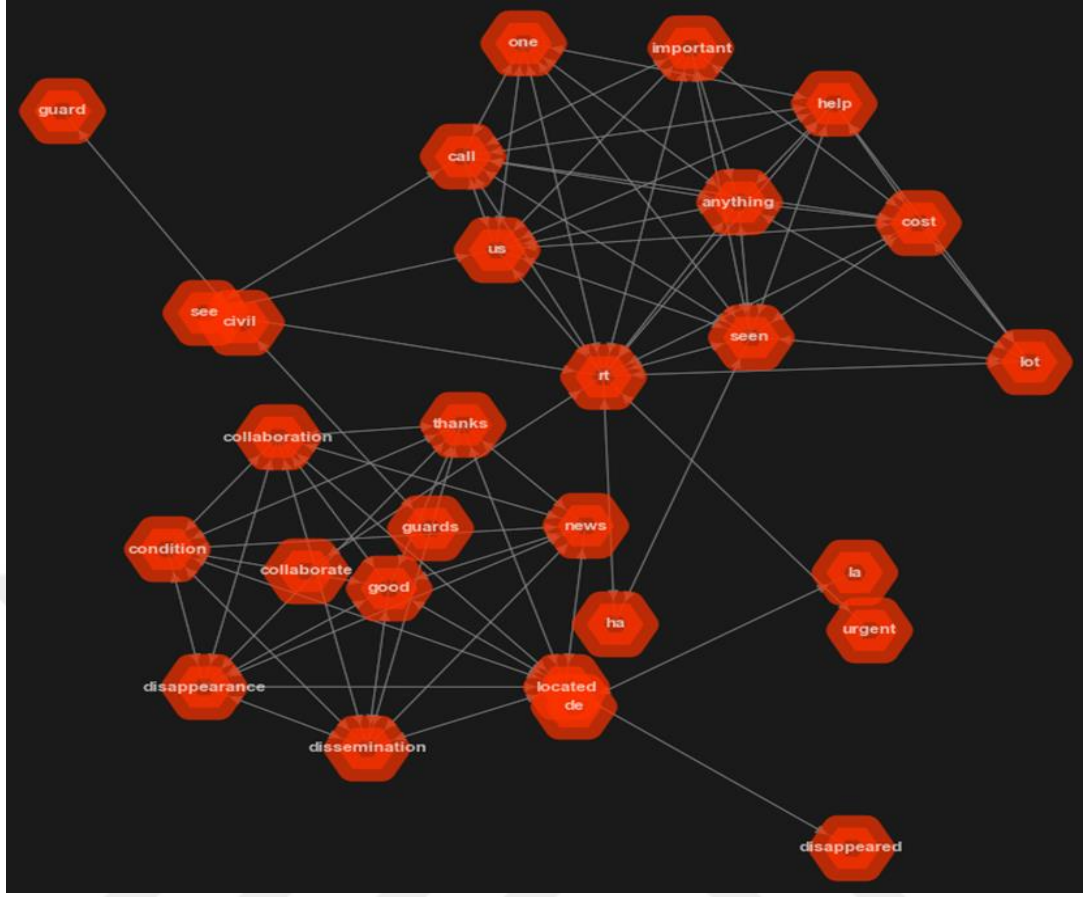
Şekil 3.144. Metnin duygusal durumunun saat bazlı incelenmesi

Paylaşımların yapıldığı zaman dilimi ile duygusal durumu arasındaki ilişki Şekil 3.144’de gösterilmekte olup, incelenen korpusun yaklaşık 6 aya yakın süreyi kapsamamasından dolayı zamanın da duygusal duruma etkisi olduğu söylenebilmektedir.



Şekil 3.145. Paylaşımlar arasında sık kullanılan ifadelere ait kelime bulutu

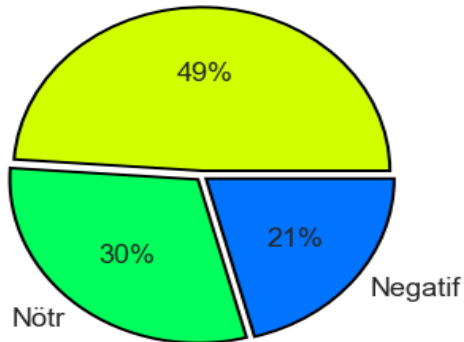
Yapılan paylaşımlarda en sık kullanılan ifadeler bazında (# ve @ gibi ifadeler olmaksızın) şekil 3.145’deki kelime bulutunun elde edildiği görülmektedir.



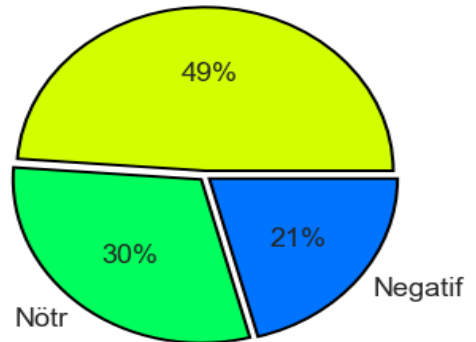
Şekil 3.146. Yapılan paylaşımlara ait kelime ağı analizi (50>Kelime Kombinasyonu)

Yapılan paylaşım metinleri BoW (Bag of words) metoduyla veri seti haline getirilmiştir. Söz konusu veri seti içerisinde de 50’den fazla kombinasyona sahip kelimelerin ağ (network) yapısı Şekil 3.146’da gösterilmiştir.

Eğitim Veri Seti - @guardiacivil %75
Pozitif

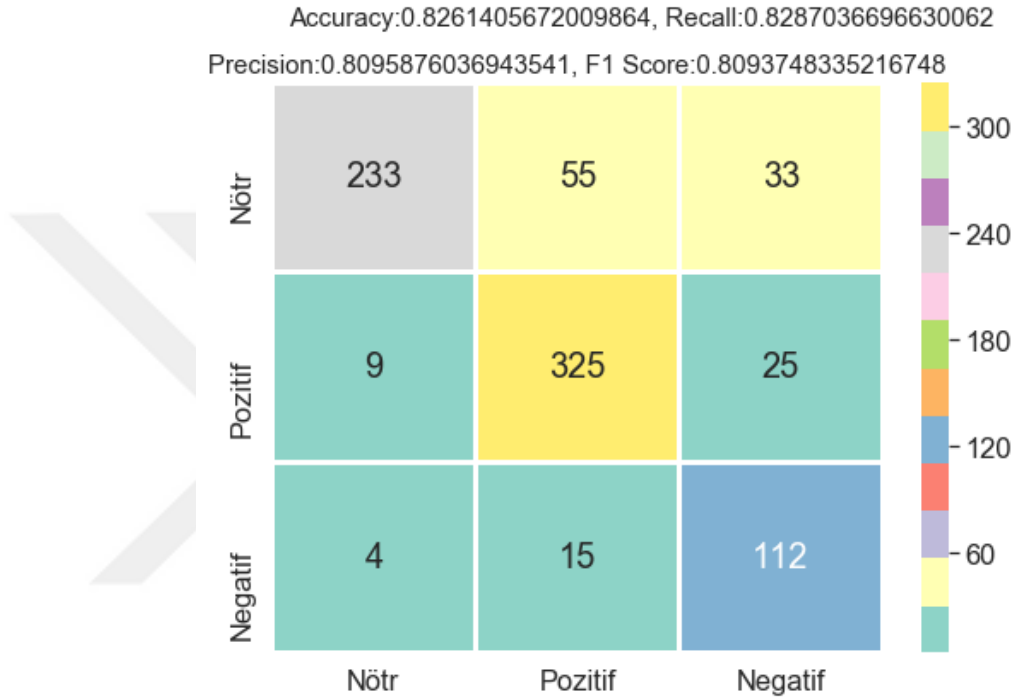


Test Veri Seti - @guardiacivil_ %25
Pozitif



Şekil 3.147. Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması

Textblob kütüphanesi aracılığıyla yapılan paylaşımların duygu durumları belirlenmiş olup, söz konusu tespitlerin kendi içinde tutarlı olup olmadığını incelemek adına makine öğrenimi modellerinden rastgele orman sınıflandırması (Random forest classification) ile veri seti Şekil 3.147’de gösterildiği üzere test ve eğitim setine ayrılmıştır. Sınıflandırma için verinin %75’i eğitim, %25’i de test olarak ayrılmıştır.



Şekil 3.148. Test verisinin duygusal matriksi

Makine öğrenimi sonrası test verileri üzerinde çıkan sonuca ilişkin matriks Şekil 3.148’de görülmekte olup, yapılan öğrenme test verileri üzerinde %82.614 oranında doğruluk oranı sağlamıştır.

3.4.2. Policía Nacional’in Paylaşımlarının İncelenmesi

İspanya’da faaliyet gösteren polis türü kolluk gücünün resmi hesabı olan “Policía Nacional” başlıklı ve @policia kullanıcı adlı Twitter hesabından 09 Mart 2020 tarihinde herkese açık olarak yapılan paylaşımlar Tweepy kütüphanesi aracılığıyla Python programlama dili kullanılarak, csv formatında çekilmiştir.



Şekil 3.149. Policía Nacional resmi Twitter sayfasının ekran resmi

	id	created_at	text	followers_count	retweet_count	favourite_count	source
0	1238619147180654593	2020-03-09 07:38:00	¡Buenos días y #FelizLunes! Comenzamos...	3311613	3	31	TweetDeck
1	1238768152345645056	2020-03-08 21:38:00	A pocas horas de que termine el #domingo, noso...	3311613	41	367	TweetDeck
2	1238733928909120514	2020-03-08 19:22:00	Hoy puede ser un buen día para hacer una copia...	3311613	23	109	TweetDeck
3	1238704988001575936	2020-03-08 17:27:00	Mientras conduces, olvídate de cualquier distr...	3311613	39	143	TweetDeck
4	1238675793847107584	2020-03-08 15:31:00	¿Conoces algún caso de tráfico de estupefacien...	3311613	37	143	TweetDeck

Şekil 3.150. @policia kullanıcı adlı hesaptan çekilen veriye ait bilgiler- I

```
print(data.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3227 entries, 0 to 3226
Data columns (total 7 columns):
id                3227 non-null int64
created_at        3227 non-null object
text              3227 non-null object
followers_count   3227 non-null int64
retweet_count     3227 non-null int64
favourite_count   3227 non-null int64
source            3227 non-null object
dtypes: int64(4), object(3)
memory usage: 176.6+ KB
None
```

Şekil 3.151. @policia kullanıcı adlı hesaptan çekilen veriye ait bilgiler-II

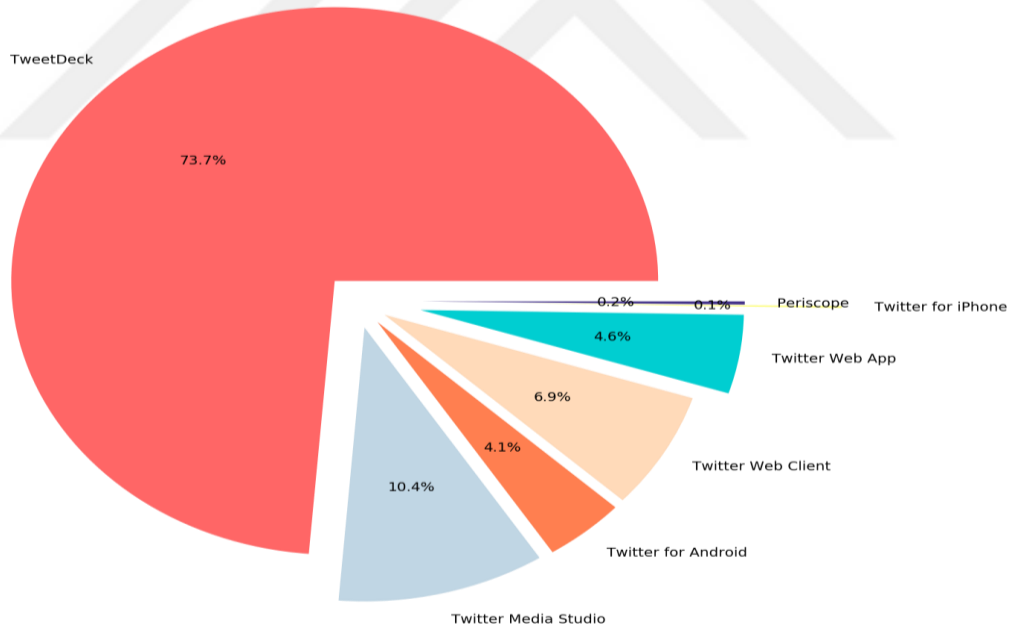
```
print(f" Data Available since {data.created_at.min()}")
print(f" Data Available upto {data.created_at.max()}")

Data Available since 2019-05-26 13:02:01
Data Available upto 2020-03-09 07:38:00
```

Şekil 3.152. @policia kullanıcı adlı hesaptan çekilen veriye ait bilgiler- III

Pandas kütüphanesi aracılığıyla @policia kullanıcı adlı hesaptan çekilen veriler üzerinde ön inceleme yapıldığında;

- Verinin 3227 satır ve 7 sütundan oluştuğu (Şekil 3.150, Şekil 3.151),
- Verinin sütun isimlerinin “id”, “created_at”, “text”, “followers_count”, “retweet_count”, “favourite_count” ve “source” olduğu (Şekil 3.151),
- Veri tipleri olarak “created_at”, “text” ve “source” sütunlarının object (nesne), diğer sütunların ise int64 (sayısal - integer) olarak tanımlandığı (Şekil 3.151),
- Veride 0' dan 3226'ya kadar numaralandırılmış olan 3227 adet girdi bulunduğu ve her bir sütunda boş olmayan (non-null) girdi sayısının tanımlandığı, bu değerlerden de değeri olmayan sütun değerinin bulunmadığının anlaşıldığı (Şekil 3.151),
- @policia adlı kullanıcıdan çekilen verilerin 26 Mayıs 2019'dan 09 Mart 2020 tarihleri arasında yer aldığı (Şekil 3.152) anlaşılmaktadır.



Şekil 3.153. @policia adlı kullanıcının veri paylaşımında kullandığı kaynaklar

@policia adlı kullanıcının Twitter paylaşımı yaptığı cihazların kaynağı incelendiğinde; %73.7'sinin de TweetDeck, %4,1'inin Android, %6.9'unun Twitter web istemcisi (web client), %4,6'sının Twitter web uygulaması, %0.2'sinin Periscope,

%10.4'ünün Twitter medya stüdyosu ve %0.1'inin de Iphone aracılığıyla olduğu şekil 3.153'de gösterilmektedir.

```
print(f" Maximum number of retweets {data.retweet_count.max()}")
print(f" Maximum number of favorites {data.favourite_count.max()}")

Maximum number of retweets 17627
Maximum number of favorites 48730

data.loc[data['retweet_count']==17627,['text']].values

array([[ 'ÚLTIMA HORA\r\n\r\nLa @policia Nacional ha detenido a la persona que agredió ayer
ortar una bandera española\r\n\r\nHa sido arrestado por delito contra el ejercicio de los Der
úblicas y delito leve de lesiones\r\n#EstamosPorTi https://t.co/DnDQAYkTyd']],
      dtype=object)

data.loc[data['retweet_count']==17627, ['created_at']].values

array(['2019-10-15 12:30:37'], dtype=object)

data.loc[data['favourite_count']==48730,['text']].values

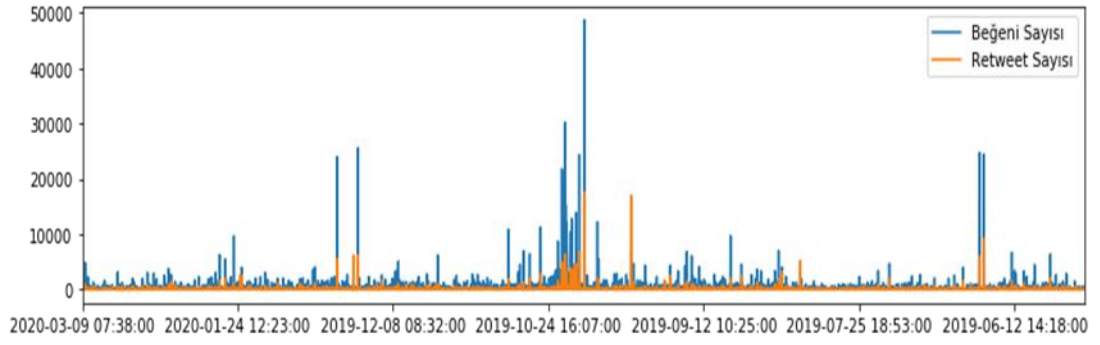
array([[ 'ÚLTIMA HORA\r\n\r\nLa @policia Nacional ha detenido a la persona que agredió ayer
ortar una bandera española\r\n\r\nHa sido arrestado por delito contra el ejercicio de los Der
úblicas y delito leve de lesiones\r\n#EstamosPorTi https://t.co/DnDQAYkTyd']],
      dtype=object)

data.loc[data['favourite_count']==48730, ['created_at']].values

array(['2019-10-15 12:30:37'], dtype=object)
```

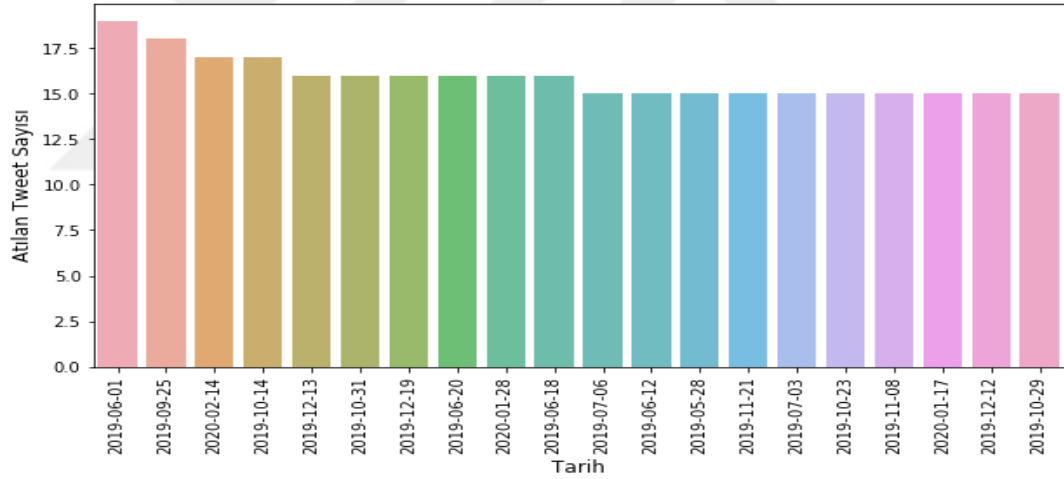
Şekil 3.154. Paylaşımların beğenilme ve retweet durumunun incelenmesi

Yapılan paylaşımlar şekil 3.154'de görüldüğü üzere incelendiğinde, 17627 tekrar paylaşım oranı ile en fazla retweet yapılan paylaşımın 08 Kasım 2018 tarihinde paylaşılan bir retweet mesajı ['ÚLTIMA HORA\r\n\r\nLa @policia Nacional ha detenido a la persona que agredió ayer a una ciudadana de Tarragona por portar una bandera española\r\n\r\nHa sido arrestado por delito contra el ejercicio de los Derechos Fundamentales y Libertades Públicas y delito leve de lesiones\r\n#EstamosPorTi https://t.co/DnDQAYkTyd'] olduğu, 48730 beğeni ile de en fazla beğeni alan paylaşımın retweet paylaşımıyla aynı olduğu anlaşılmaktadır. Ayrıca Twitter paylaşımlarının 1.871.762 kez beğenildiği (favorites), 550.948 kez tekrar paylaşıldığı (retweet) ve toplam takipçi sayısının ise 3.311.613 olduğu belirlenmiştir.



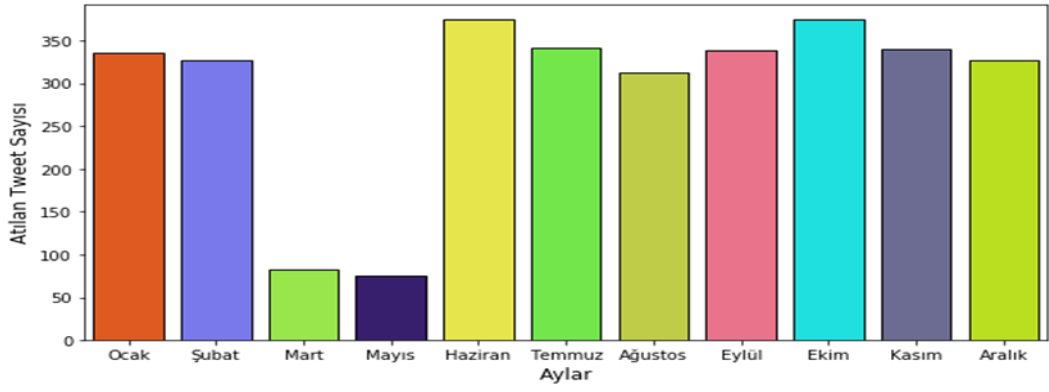
Şekil 3.155. @policia adlı kullanıcının paylaşımlarının beğenilme ve retweet durumu

@policia adlı kullanıcının twitter paylaşımlarının beğenilme ve retweet durumlarının tarih bazlı ortak grafiği şekil 3.155’de gösterilmektedir (Takipçi sayısının tarih bazında sabit olmasından dolayı “followers_count” değeri ile tweet kimliğini içeren “id” değerinde ilişkisel analiz incelenmemiştir.).



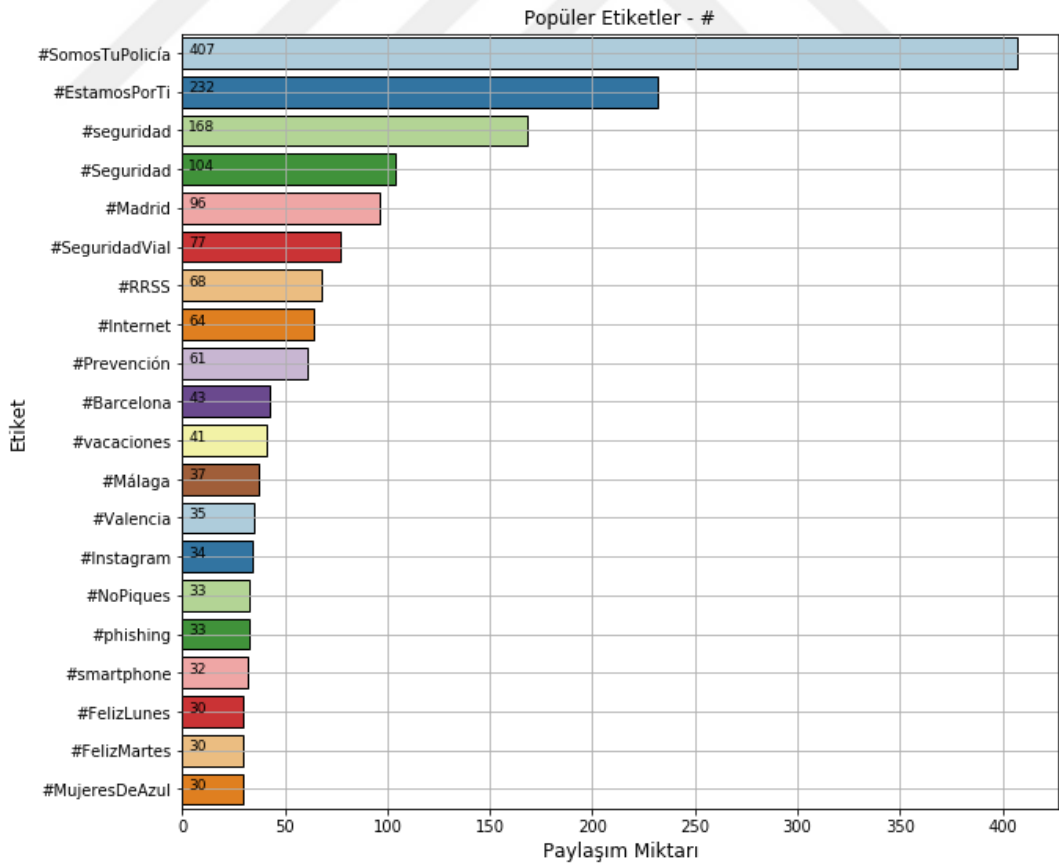
Şekil 3.156. Verilerin çekildiği zaman aralığı içinde en fazla paylaşımın yapıldığı 20 tarih ve yapılan paylaşım sıklığı

Şekil 3.156 incelendiğinde aynı gün içerisinde sık paylaşımın yapıldığı tarihler görülmekte olup, örnek olarak en sık paylaşımın yapıldığı 01 Haziran 2019 tarihindeki 19 adet yapılan paylaşım incelendiğinde, paylaşımların çoğunun 2019 yılında İspanya'nın Madrid şehrindeki Wanda Metropolitano Stadyumu'nda oynanan şampiyonlar ligi finali kapsamında alına güvenlik tedbirleri ile ilgili olduğu görülmektedir.



Şekil 3.157. Verilerin çekildiği zaman aralığı içinde yapılan paylaşımların ay bazında dağılımı

Twitter aracılığıyla sosyal medyadan yapılan paylaşımlara şekil 3.157’de olduğu üzere aylık dağılımına bakıldığında 375 Tweet ile ekim ayı en fazla sosyal medya paylaşımının yapıldığı ay iken, genel seyrinde paylaşımların düzenli aralıklarla yapıldığı gözlemlenmektedir.

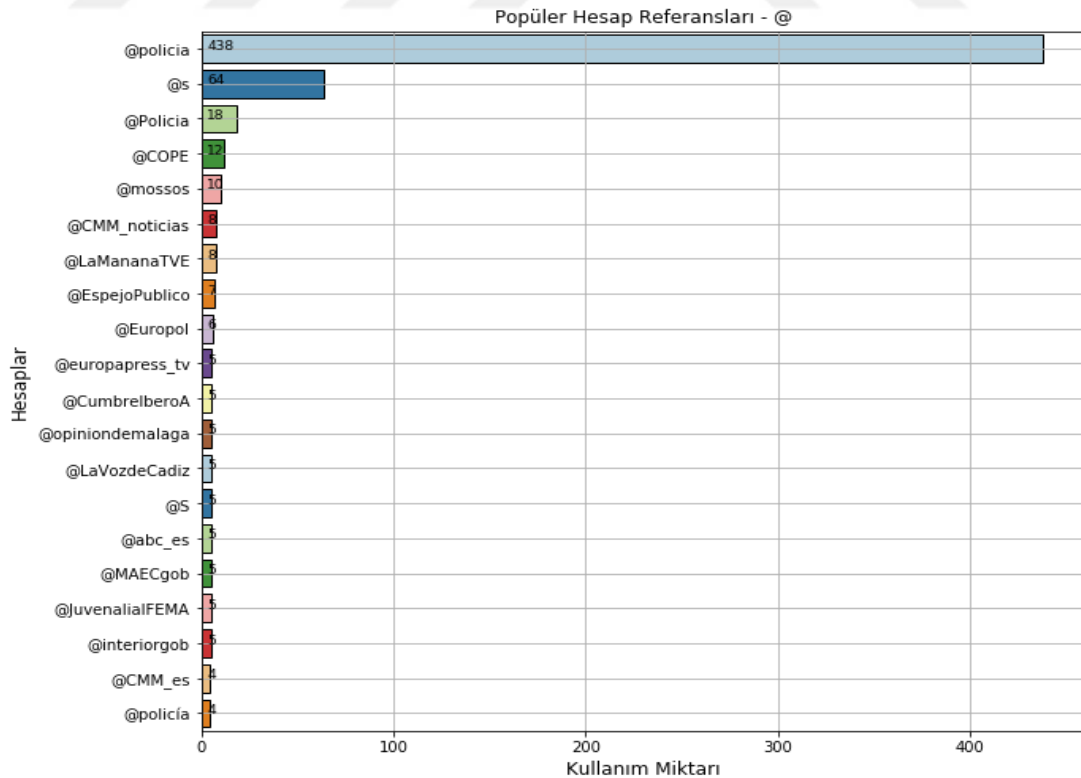


Şekil 3.158. En fazla kullanılan 20 sosyal medya etiketi (Hashtag - #)



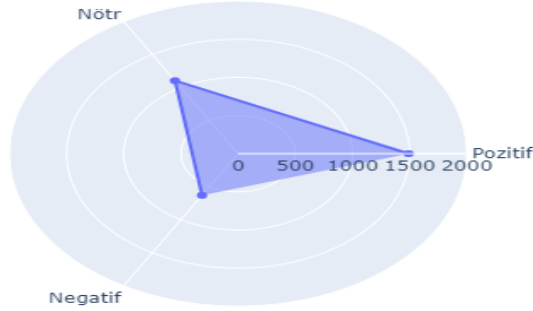
Şekil 3.159. En fazla kullanılan sosyal medya etiketlerinin (Hashtag -#) kelime bulutu

Yapılan paylaşımlarda en fazla kullanılan sosyal medya etiketlerinin 20'sine ilişkin kullanım sıklığı ile sık kullanılan sosyal medya etiketlerine ilişkin hazırlanan kelime bulutu (Maksimum kelime girdisi 1000 olarak alınmıştır.) şekil 3.158 ve 3.159'da gösterilmektedir.



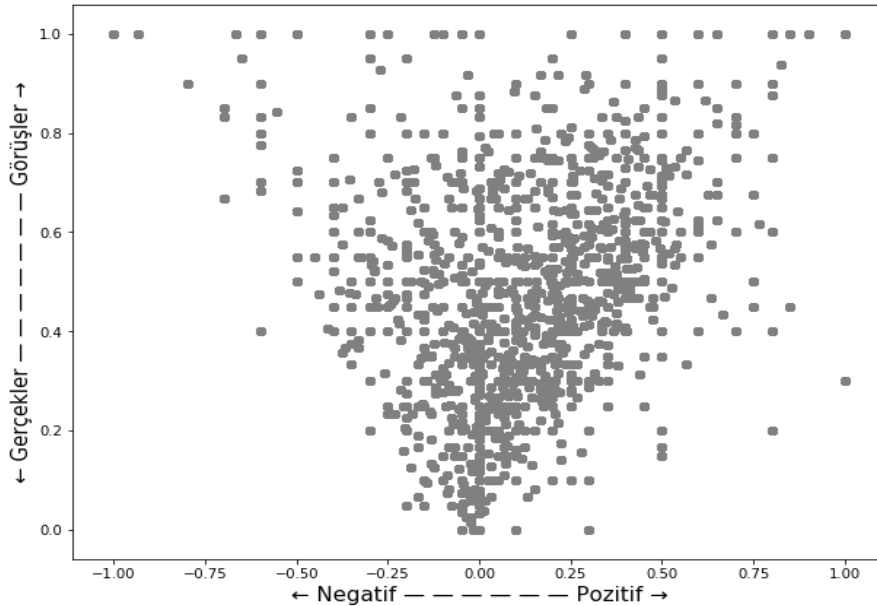
Şekil 3.160. En fazla kullanılan 20 sosyal medya hesabı (Hesap - @)

Yapılan paylaşımlarda en fazla kullanılan sosyal medya hesaplarının 20'sine ilişkin kullanım sıklığı şekil 3.160'da gösterilmektedir.



Şekil 3.161. Metnin duygusal durumunun radar grafiği

@policia kullanıcı hesabından yapılan paylaşım metinleri, metin ön işlemesine (text preprocessing) tabi tutularak tüm emojiler kaldırılmış ve İspanyolca/İngilizce karakter farklılıklarından dolayı metindeki tüm harfler küçültülerek makine öğrenimine dayalı olarak İspanyolca'dan İngilizce'ye çeviri yapılmıştır. Çeviri metni üzerinde yapılan duygusal dağılım (0=Nötr, 0>Negatif, 0<Pozitif) şekil 3.161'de görüldüğü üzere 1492 tweet ile pozitif mesajların çoğunlukta olduğu, 1104 tweet ile nötr mesajların pozitif mesajlarını izlediği ve son olarak da 631 tweet ile negatif skorlanan mesajların bulunduğu belirlenmiştir.

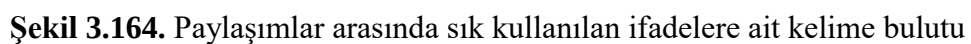


Şekil 3.162. Metnin duygusal durumunun öznellik analizi

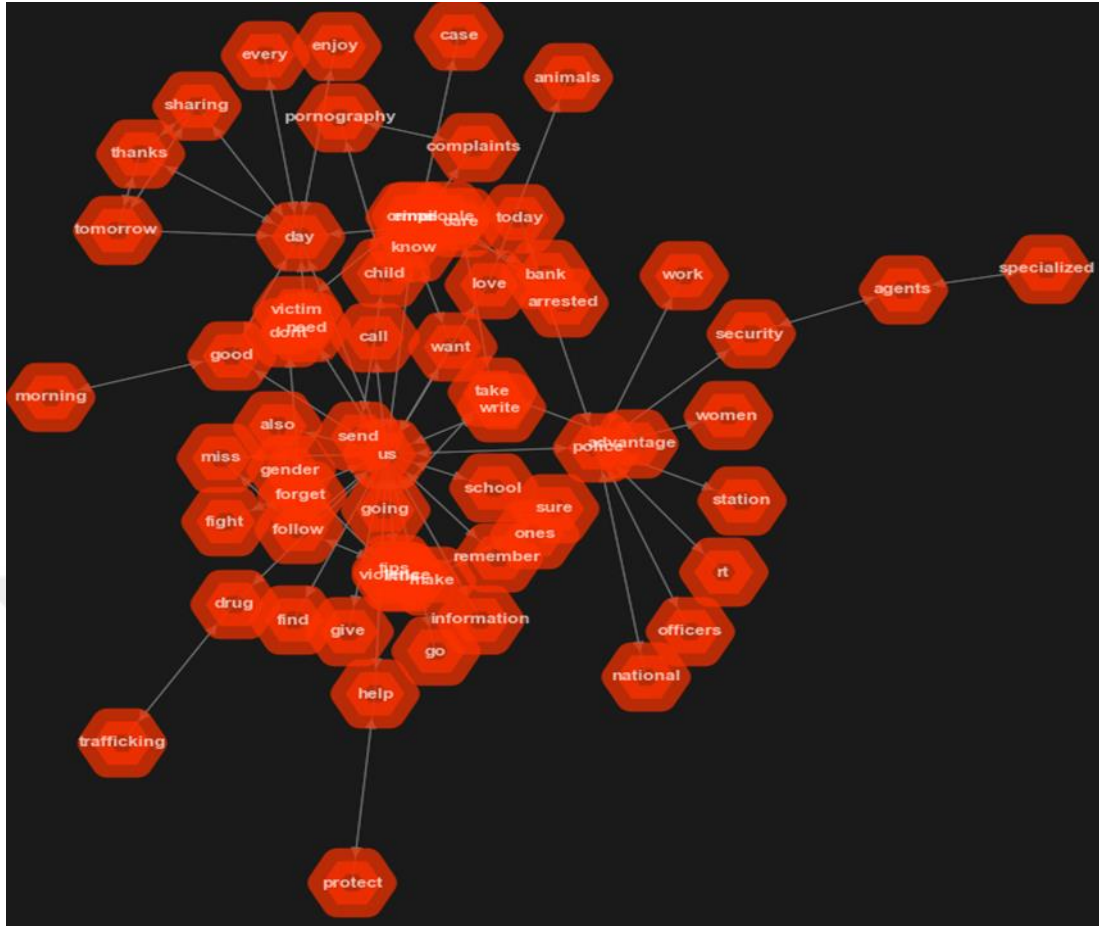


Şekil 3.163. Metnin duygusal durumunun saat bazlı incelenme

Paylaşımların yapıldığı zaman dilimi ile duygusal durumu arasındaki

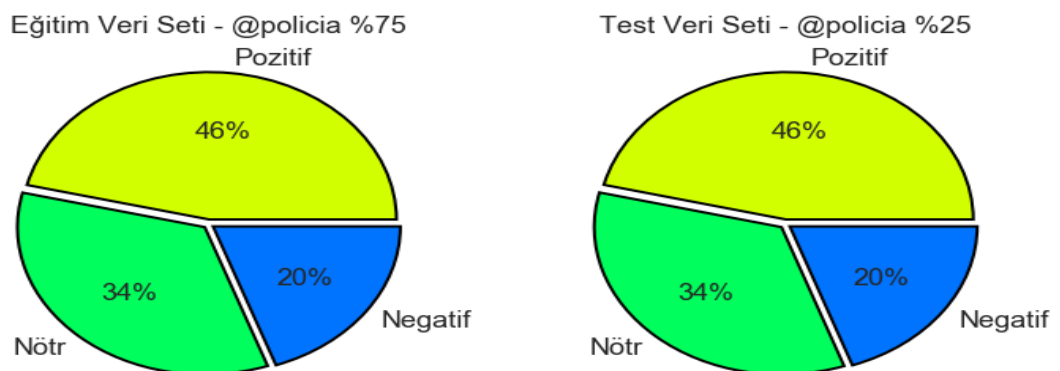


Yapılan paylaşımlarda en sık kullanılan ifadeler bazında (# ve @ gibi ifadeler olmaksızın) şekil 3.121'deki kelime bulutunun elde edildiği görülmektedir.



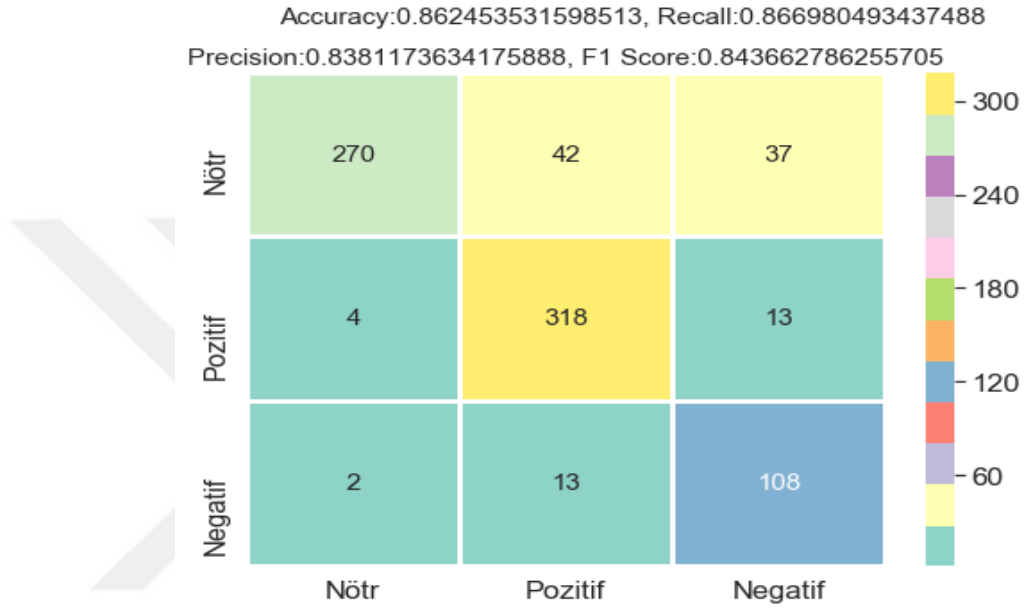
Şekil 3.165. Yapılan paylaşımlara ait kelime ağı analizi (25>Kelime Kombinasyonu)

Yapılan paylaşım metinleri BoW (Bag of words) metoduyla veri seti haline getirilmiştir. Söz konusu veri seti içerisinde de 25’den fazla kombinasyona sahip kelimelerin ağ (network) yapısı şekil 3.165’de gösterilmiştir.



Şekil 3.166. Makine öğrenimi için veri setinin eğitim ve test verilerine ayrılması

Textblob kütüphanesi aracılığıyla yapılan paylaşımların duygu durumları belirlenmiş olup, söz konusu tespitlerin kendi içinde tutarlı olup olmadığını incelemek adına makine öğrenimi modellerinden rastgele orman sınıflandırması (Random forest classification) ile veri seti şekil 3.166’da gösterildiği üzere test ve eğitim setine ayrılmıştır. Sınıflandırma için verinin %75’i eğitim, %25’i de test olarak ayrılmıştır.

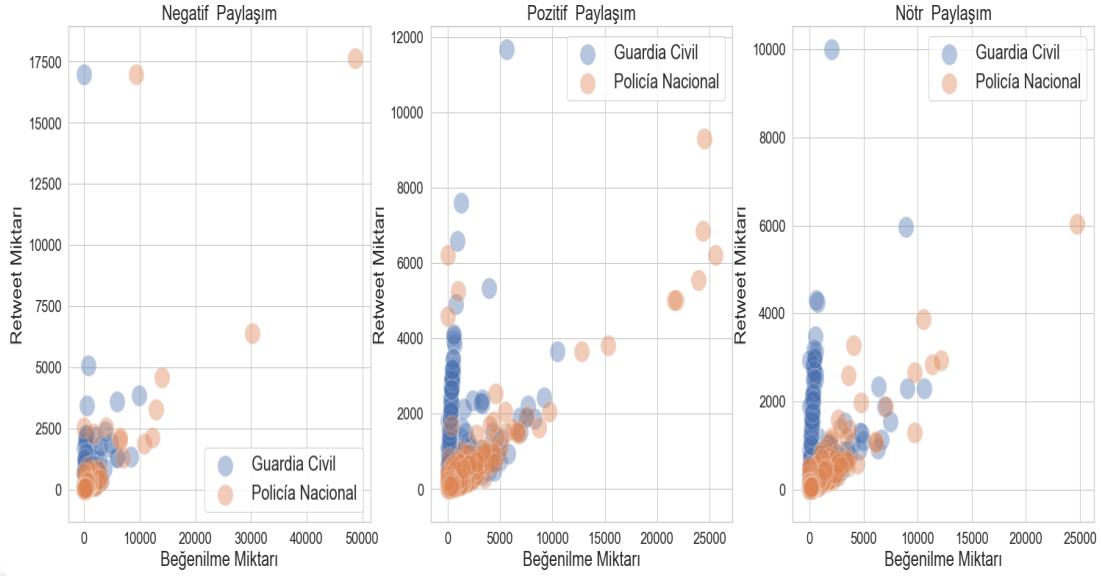


Şekil 3.167. Test verisinin duygusal matriksi

Makine öğrenimi sonrası test verileri üzerinde çıkan sonuca ilişkin matriks şekil 3.167’de görülmekte olup, yapılan öğrenme test verileri üzerinde %86.245 oranında doğruluk oranı sağlamıştır.

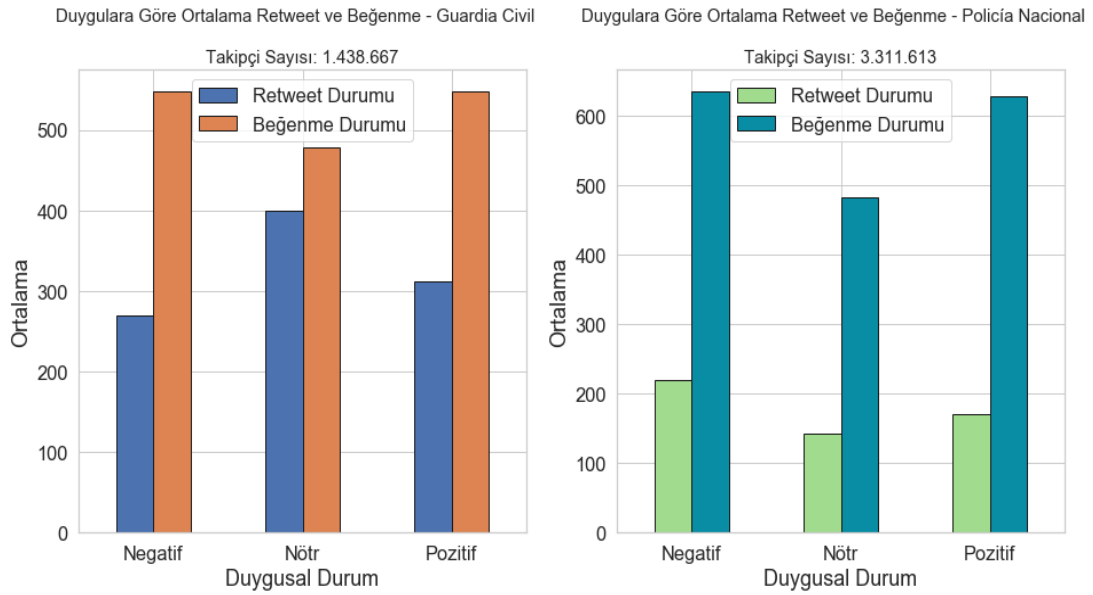
3.4.3. Guardia Civil ve Policía Nacional Teşkilatlarının Mukayeseli Ortak Analizi

İspanya özelinde faaliyet gösteren kolluk teşkilatlarının topluma verdiği mesajları, mesajlarda hangi konu ve kelimeleri ağırlıkta kullandığı ile paylaştığı mesajların sosyal medyadaki etkisini mukayeseli olarak aşağıda incelenmektedir. Mukayese kapsamında korelasyon ve makine öğrenimi modellerinden karar destek vektörleri kullanılarak, tespitler görsel içerikle desteklenmektedir.



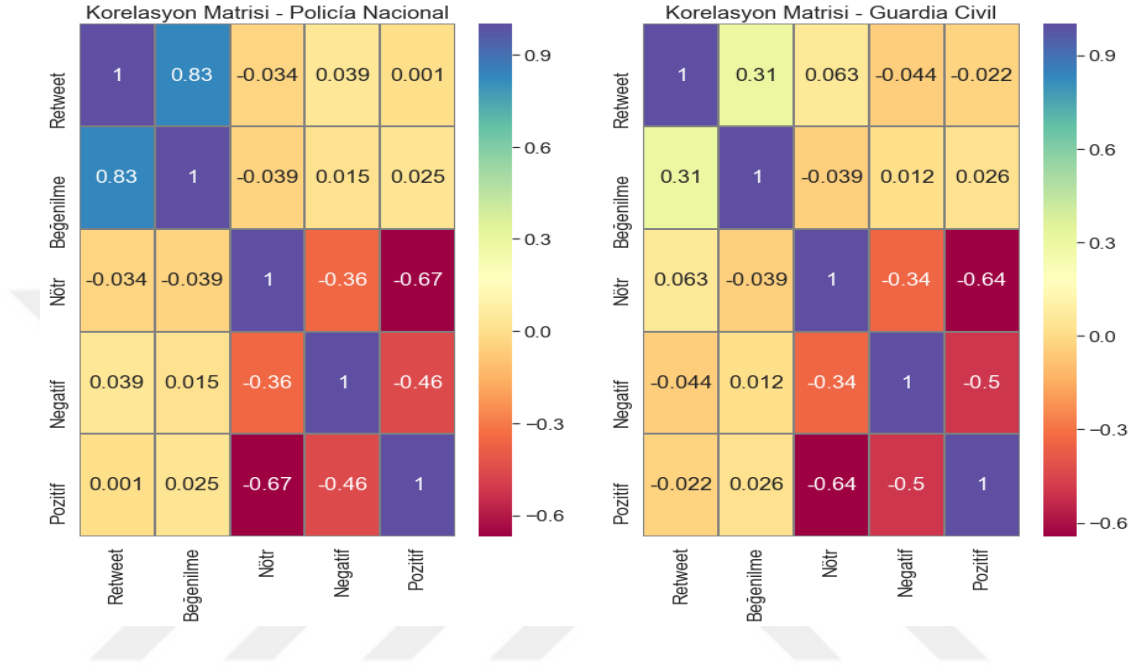
Şekil 3.168. Duygulara göre retweet ve beğenme durumu

Paylaşımlar içerisinde anomali tespiti yapmak amacıyla şekil 3.168 incelendiğinde, Guardia Civil özelinde bazı pozitif ve nötr cümlelerde anomalinin fazla olduğu, Policía Nacional özelinde ise yine bazı pozitif ve nötr cümlelerde anomalinin fazla olduğu gözlemlenmiştir. Kurumlar arası genel dağılımın birbirine yakın olduğu ve en fazla anomali gösteren dağılımın pozitif paylaşımlarda yaşandığı belirlenmiştir.



Şekil 3.169. Duygulara göre ortalama retweet ve beğenme durumu

Paylaşımların genel ortalaması şekil 3.169’da görüldüğü üzere incelendiğinde, genel olarak pozitif paylaşımlarda Guardia Civil’nin negatifte beğeni ve nötr retweet ortalaması yüksek iken, Polizia di Stato’nun ise negatif beğeni ve retweet ortalamasının daha yüksek olduğu gözlemlenmektedir.

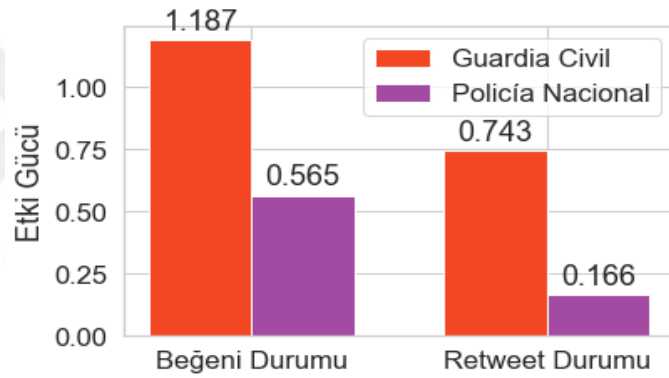


Şekil 3.170. Duygusal korelasyon matrisi

Duyguların, beğeni ve retweet üzerindeki gücünün korelasyon biçimindeki gösterimi matris biçiminde şekil 3.84’de yapılmış olup, söz konusu verilerin analizinden;

- Beğeni ve Retweet arası ilişkinin Policia Nacional’da daha güçlü olduğu (Policia Nacional: 0.83, Guardia Civil:0.31),
- Pozitif mesajlar kapsamında yapılan paylaşımları insanların beğenmesin Policia Nacional’da daha yüksek ilişki düzeyinde olduğu (Policia Nacional: 0.025, Guardia Civil:0.026),
- Pozitif mesajlar kapsamında yapılan paylaşımları insanların retweet etmesinin Policia Nacional’da daha yüksek ilişki düzeyinde (Policia Nacional: 0.001, Guardia Civil:-0.022),

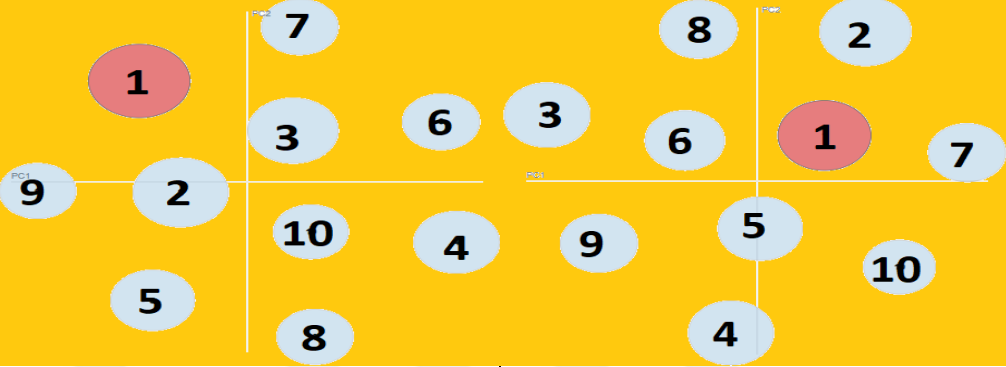
- Nötr mesajlar kapsamında yapılan paylaşımları insanların beğenmesinin aynı ilişki düzeyinde olduğu (Policía Nacional: -0.039, Guardia Civil:-0.039),
- Nötr mesajlar kapsamında yapılan paylaşımları insanların retweet etmesinin Policía Nacional’da daha yüksek ilişki düzeyinde olduğu (Policía Nacional: -0.034, Guardia Civil:-0.044),
- Negatif mesajlar kapsamında yapılan paylaşımları insanların beğenmesinin Policía Nacional’da daha yüksek ilişki düzeyinde olduğu (Policía Nacional: 0.015, Guardia Civil:0.012),
- Negatif mesajlar kapsamında yapılan paylaşımları insanların retweet etmesinin Policía Nacional’da daha yüksek ilişki düzeyinde olduğu (Policía Nacional: 0.039, Guardia Civil:-0.044) gözlemlenmiştir.



Şekil 3.171. Guardia Civil ve Policía Nacional’un etki gücü

Yapılan paylaşımların takipçileri oranında hangi derecede paylaşıldığının ve retweet edildiğinin ortaya konması amacıyla etki gücü hesaplaması yapılmış, elde edilen sonuçlar da şekil 3.171’de gösterilmektedir. Etki gücü hesaplaması yapılırken, toplam beğeni sayısı/toplam takipçi sayısı ve toplam retweet sayısı/toplam takipçi sayısı formülasyonu kullanılmıştır. Buradaki amaç her takipçinin ortalama ne kadar mesajı beğendiği ve retweet ettiği tespitiyle, mesajların ne kadar güçlü olduğunu karşılaştırmalı olarak ortaya koymaktır. Bu yapılan analiz sonucunda ise Guardia Civil’in paylaşımlarının Policía Nacional’a kıyasla etki gücünün kısmi daha yüksek olduğu anlaşılmaktadır. Bu sebeple de aradaki farkın sebebinin daha iyi anlayabilmek adına makine öğrenimine dayalı paylaşım yaptıkları konuların neler olduğunun ortaya konması gerekmektedir.

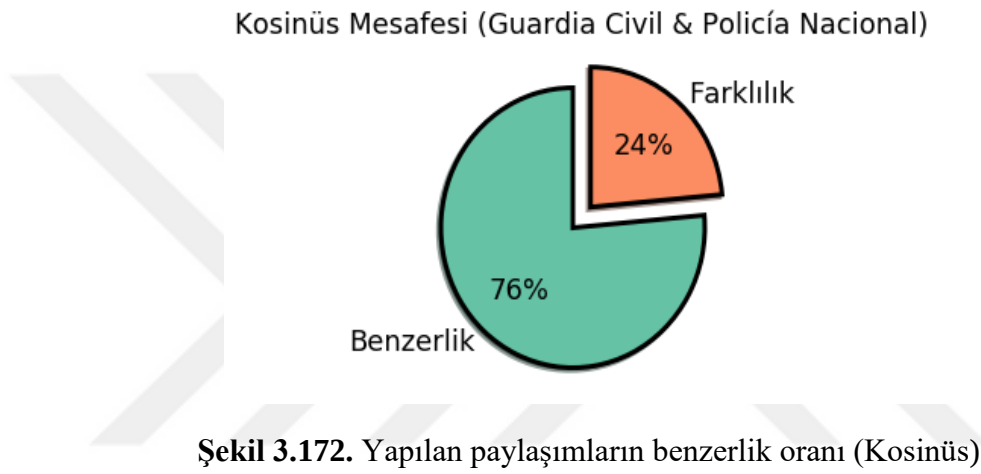
Çizelge 3.4. Guardia Civil ve Policía Nacional’un paylaşım yaptığı konuların LDA (Latent Dirichlet Allocation) analizi ve konular arası mesafe haritalandırması

Konu	Guardia Civil LDA Analizi	Policía Nacional LDA Analizi
Mesafe Haritası		
1	('dog', 140.12360727521875) ('carry', 109.25004136328626) ('know', 94.2456067973439) ('rescue', 80.91266700699393) ('mountain', 80.88968158080019) ('information', 73.07070196038802) ('work', 57.53230813505805) ('guide', 56.006084110273854) ('mobile', 50.19170836656235) ('detect', 49.983675549482925)	('need', 123.57261609246687) ('remember', 118.53535766361226) ('good', 90.53600946898162) ('protect', 88.50228367611831) ('time', 82.02959007819506) ('help', 80.5471024778533) ('leave', 78.6011551159099) ('start', 72.53257081378932) ('dont', 60.06253106738799) ('lose', 59.96964462050105)
2	('guard', 171.6234876351302) ('civil', 156.89451598771547) ('today', 107.13782534345776) ('challenge', 88.01455962907478) ('life', 75.61300207934545) ('national', 73.8354403053037) ('receive', 54.44797332815768) ('league', 52.66445892010823) ('friend', 49.56646188791655) ('cyberspace', 49.306273663804994)	('police', 204.57313467087343) ('agent', 145.71443934616238) ('security', 129.03810736619076) ('national', 111.12505792920368) ('year', 82.61156935021918) ('today', 74.43173856851307) ('work', 67.93515092912436) ('ready', 55.863227822014025) ('woman', 51.638958866265) ('officer', 47.647019394015885)
3	('good', 240.98516052735965) ('locate', 192.61661487290357) ('thanks', 161.05792790965796) ('important', 151.76053819711876) ('condition', 126.3839150173466) ('collaboration', 104.05057073120321) ('dissemination', 103.20174603513865) ('disappearance', 94.22542274699718) ('news', 88.81589921002727) ('security', 88.41515969226212)	('follow', 123.73838920333394) ('police', 71.0776658557637) ('criminal', 70.15618470634676) ('avoid', 70.03505172703053) ('tip', 69.73958568307481) ('operation', 67.10616093382696) ('station', 62.56877184134121) ('want', 61.412532626957386) ('miss', 58.44766420477939) ('network', 46.181004642876026)
4	('help', 306.0360402075204) ('cost', 129.5860339035234) ('lot', 105.31834385254652) ('know', 85.49261749264821) ('way', 64.4186882851402) ('urgent', 58.11124087382037) ('advice', 54.94934376868562) ('act', 54.144620838745276) ('road', 51.76102447888996) ('activity', 51.67886645075702)	('arrest', 135.21942176973744) ('come', 88.38564375020745) ('man', 70.99916236820377) ('people', 60.799671211395484) ('info', 49.7610294810635) ('year', 49.04167308810061) ('car', 48.58598900859447) ('door', 44.163956267947775) ('home', 42.60960470122065) ('stop', 41.361786592706444)

Çizelge 3.4. Devam

Konu	Jandarma LDA Analizi	Emniyet LDA Analizi
5	('day', 213.95426761221924) ('need', 99.49795836931102) ('avoid', 64.34981030418653) ('phase', 62.17810409505514) ('sign', 56.76204173192272) ('say', 56.07701547775157) ('enjoy', 52.97035079342097) ('start', 42.16089004293906) ('road', 40.21207415556871) ('allow', 39.92107644213804)	('crime', 128.1936211443837) ('send', 90.24457864651222) ('information', 83.7281430968176) ('victim', 83.15648906517691) ('ask', 82.18998689479909) ('bank', 73.83816880583059) ('email', 68.50592054835256) ('child', 58.817232293267665) ('money', 52.87460310248417) ('fight', 51.497398975608256)
6	('use', 78.88092070284462) ('car', 74.51704769991736) ('road', 65.89747777346996) ('control', 62.0720118793563) ('dont', 52.918054896225506) ('vehicle', 51.95355181586284) ('right', 49.81159327418055) ('come', 42.97022571136982) ('let', 40.198177113477) ('think', 32.367502952014625)	('life', 85.16075548140357) ('work', 85.1186075723716) ('enjoy', 77.77588573365627) ('respect', 65.70642964335366) ('dont', 64.01092723215973) ('day', 61.79215907458359) ('case', 60.75377773486911) ('forget', 56.684214286884206) ('write', 54.095618180904914) ('good', 51.7006054608994)
7	('traffic', 107.17008181868412) ('year', 69.55932589223868) ('light', 62.572030341818234) ('vehicle', 53.81768675787) ('continue', 48.429045077202005) ('follow', 39.33986071298433) ('drive', 35.936518288462246) ('crime', 33.9309389454114) ('driver', 33.49574148408183) ('video', 29.54077850779849)	('check', 126.12133862798596) ('look', 122.86193109221095) ('make', 102.2378690597005) ('leave', 62.241268089611694) ('link', 45.291693964786376) ('sure', 37.16418199886552) ('ignore', 34.30495886311034) ('bank', 33.94387982335918) ('real', 33.04662401885159) ('appointment', 32.22377808870452)
8	('work', 97.4617728200705) ('service', 68.53503485045162) ('old', 49.75253606518471) ('year', 47.83025278725235) ('false', 45.14845871280926) ('message', 42.18785489791469) ('environment', 41.17285135538745) ('prevent', 40.603327403900856) ('water', 37.5991005578023) ('play', 37.18047558928227)	('want', 114.24065118545477) ('animal', 99.23815368826787) ('know', 98.74254775065425) ('like', 79.64586075630291) ('love', 66.47119452062958) ('school', 63.88130080552201) ('family', 57.49097304261369) ('need', 54.97683481144524) ('participate', 52.781945503564145) ('care', 52.74709402974297)
9	('place', 90.38333601035352) ('child', 59.34226768969066) ('use', 55.91534356763945) ('cause', 48.44045621106232) ('wait', 44.01480209759281) ('family', 43.2144915154734) ('rest', 42.0932638913492) ('people', 39.49075645732905) ('new', 39.02159716475898) ('minor', 38.10352401667828)	('detain', 92.5504822754041) ('victim', 59.338011598336465) ('detainee', 54.84774693722547) ('woman', 54.37809162460692) ('tell', 45.337305251200455) ('organization', 44.6627408917468) ('agent', 43.47666010771452) ('dismantle', 38.748151259103295) ('gender', 37.64028036974687) ('sexual', 35.48514900365341)
10	('like', 98.73560736549894) ('look', 98.15210294485394) ('animal', 92.77173654946898) ('know', 72.41602098299377) ('end', 55.429430358517905) ('stop', 50.6133884872589) ('leave', 46.631361361475925) ('think', 39.13362659261753) ('step', 35.15363046707888) ('dog', 33.9664182600579)	('day', 279.14041552320674) ('thanks', 250.08221916691716) ('share', 241.18033347712057) ('tomorrow', 204.8579657426856) ('drug', 134.8109134140004) ('help', 56.61974200628514) ('traffic', 38.188370025461126) ('safety', 36.79003568058359) ('return', 34.49840667951601) ('face', 33.761507841123866)

Çizelge 3.4’de konu modellemeye aktif olarak kullanılan LDA modellemesiyle, kurumların paylaşımları 10 konu başlığı altında incelenmiştir. 10 konu içinde de her konudan en yüksek ilişki düzeyine sahip 10 kelime örneklem olarak çizelgede gösterilmiştir. Her konunun vektörel düzeyde konular arası mesafe haritası çizelgenin başında verilmiş olup, kelimelerin yanında yer alan ifadeler de vektörel düzeyde konu içinde ilişkisini ifade etmektedir. Bu kapsamda konu bazında yapılan incelemelerde, paylaşım yapılan konuların ve vektörel düzlemdeki mesafelerinin ayrıştığı rahatlıkla görülmektedir.

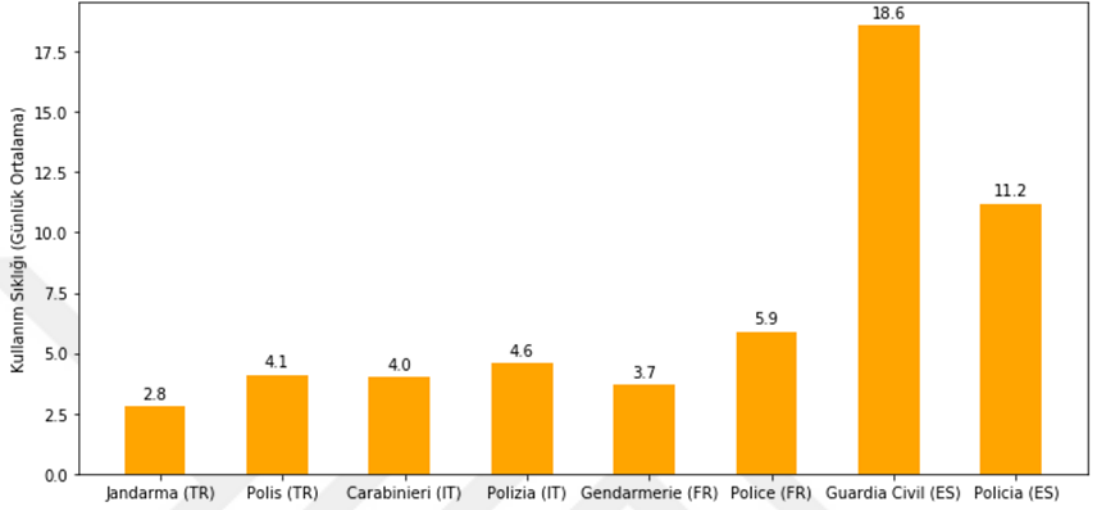


Paylaşım yapılan konuların vektörel dağılımlarındaki farklılıkların, metin korpuslarının geneline yayılıp tespit edilmesi maksadıyla, öncelikle daha net bir sonucun çıkması için kelimelerin kökleri çözümlenerek (lemmatization) işlem yapılmıştır. Metin içindeki kelime (token) yapılarının her biri vektörel düzeye indirgenerek, şekil 3.172’de görüldüğü üzere kosinüs mesafesi yöntemi ile paylaşım yapılan metinlerin benzerlik oranları %76 olarak bulunmuştur.

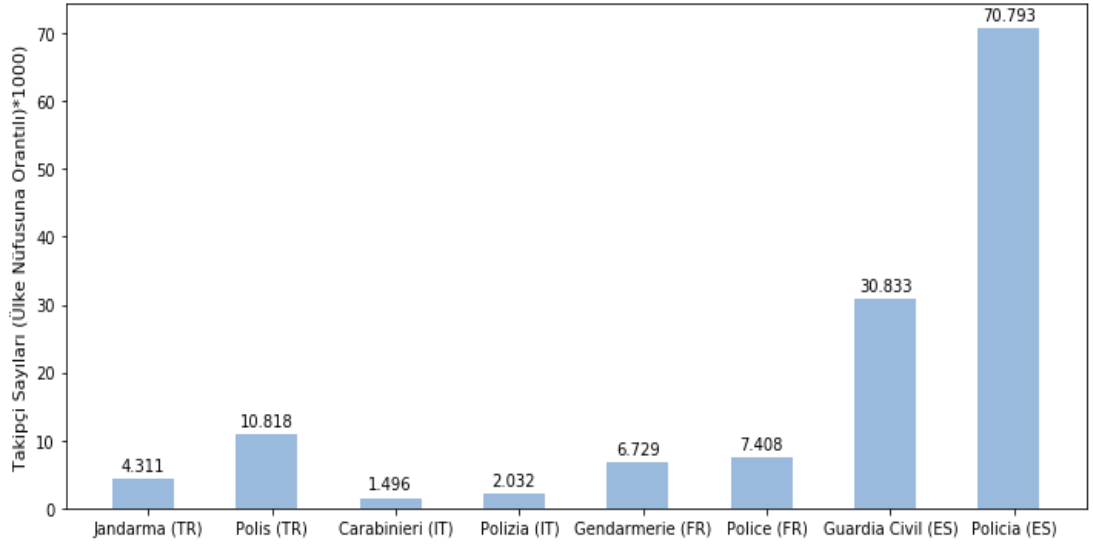
3.5. Kolluk Paylaşımlarının Uluslararası Düzeyde Mukayeseli İncelenmesi

Toplum destekli güvenlik hizmeti kapsamında, kolluk kuvvetlerinin halkın güvenini ve sempatisini kazanarak, işlem yapması esastır. Halkın kolluğa duyduğu ilgi ve sempati, adalet düzeninin sağlanması adına işbirliği teşkil eder. Bu sebeple de daha önceki başlıklarda kurum ve ülke bazında incelenen kolluk kuvvetlerinin, Türkiye,

İtalya, Fransa ve İspanya özelinde tümleşik olarak analizi yapılarak, rastgele orman sınıflandırıcısıyla (Random Forest Classification) test edilen etki duygusal analizle birlikte etki gücü yüksek olan kolluk kuvvetinin bulunarak, paylaşım konularının vektörel düzeyde TF-IDF düzeyde benzeşim oranları bulunmuştur.



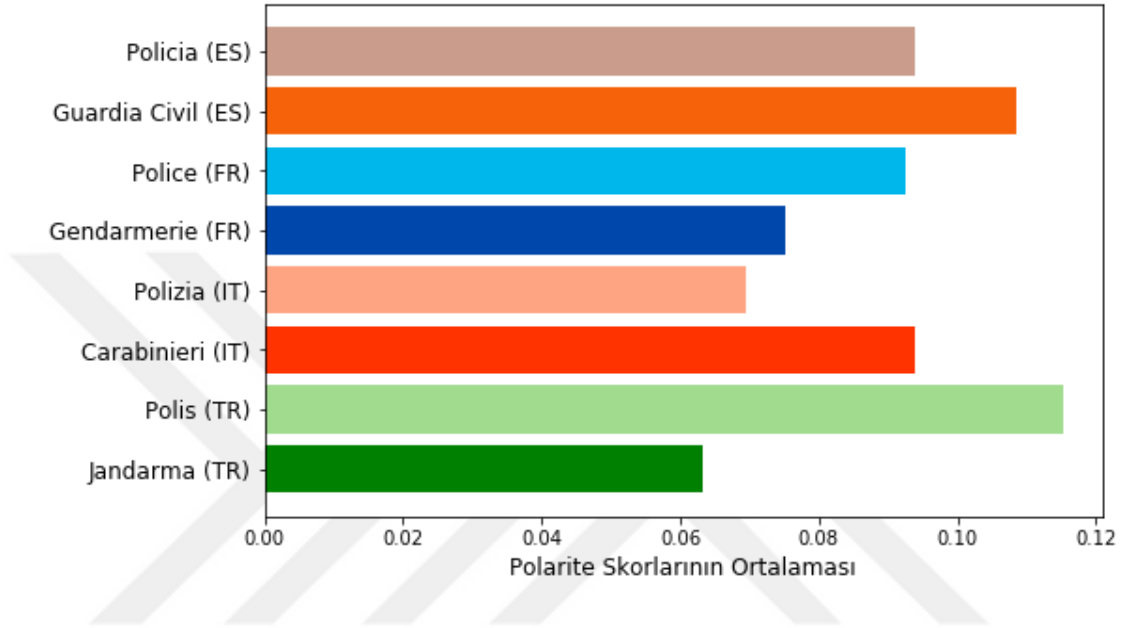
Şekil 3.173. Kolluk teşkilatlarının Twitter kullanım sıklığı



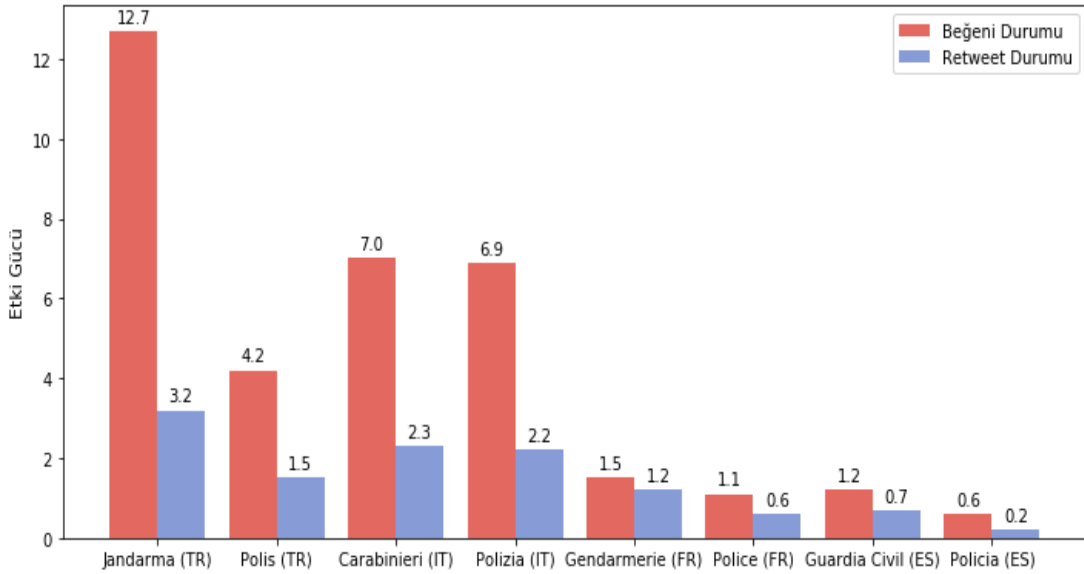
Şekil 3.174. Nüfus ve takipçi sayılarının orantısı (*1000)

Şekil 3.173 ve 5.174’de yer alan veriler incelendiğinde, İspanya hariç Jandarma teşkilatlarının, Polis Teşkilatlarına nazaran kullanım sıklığının daha az olduğu ve söz

konusu ülke nüfusuna orantılı takipçi sayılarında da benzeri bir tabloyu ortaya çıkardığı görülmektedir İspanya’da ise ülkenin Jandarma görev gücü olan Guardia Civil’in diğer Jandarma güçlerine kıyasla 5.3 kat, kendi ülkesindeki polis gücüne kıyasla 1.66 kat daha fazla paylaşım yapmasına karşın, takipçi sayısı diğer Jandarma kuvvetlerinde olduğu gibi ülkesindeki Polis gücüne karşı düşüktür.



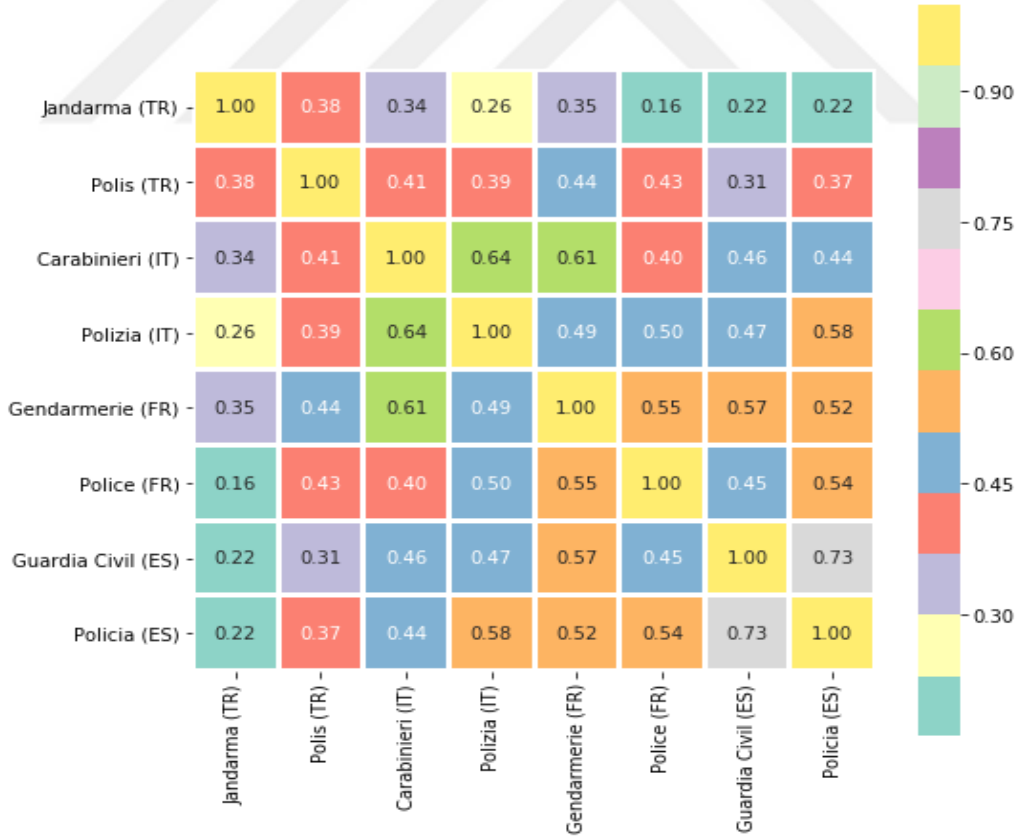
Şekil 3.175. Ortalama duygu durum puanları (Kelime bazlı cümle skoru üzerinden)



Şekil 3.176. Kolluk teşkilatlarının etki gücü (Takipçi sayısı orantılı)

Şekil 3.175 ve 5.176 incelendiğinde;

- Kolluk teşkilatlarının paylaşımlarının ortalamasının hepsi olumlu olup, bunların içinde de en olumlu mesajlar kullanan kolluğun Türk Polis teşkilatı olduğu ve hemen müteakinde Guardia Civil (ES) yer aldığı,
- Türkiye ve Fransa’da Polis teşkilatları, İtalya ve İspanya’da ise Jandarma teşkilatlarının halka daha çok olumlu mesajlar vermeye çalıştığı,
- Jandarma teşkilatlarının, Polis teşkilatlarına göre beğeni ve retweet gücünün oldukça yüksek olduğu,
- Göreli olarak en az olumlu mesajlar içeren paylaşımları yapan Türk Jandarması’nın en yüksek etki gücüne sahip olduğu,
- Paylaşım sıklığı, takipçi sayısı ve olumlu mesajlar vermede üst sıralarda olan İspanya kolluk kuvvetlerinin en düşük etki düzeyine sahip olduğu, yine benzeri şekilde Fransa özelinde de İspanya’ya yakın sonuçların olduğu gözlemlenmiştir.



Şekil 3.177. Kolluk kuvvetleri paylaşım benzerlik matrisi (TF-IDF)

Paylaşım sıklığının etki gücü üzerinde negatif etkisi bulunmakla birlikte, her ülke ve kolluğun içinde bulunduğu şartlardan mütevellik farklı konularda paylaşımlar yaptığı belirlenerek, ülke analizleri içerisinde paylaşım konularının (LDA) analiz yapılmıştır. (Çizelge 3.1 - 3.4) Söz konusu analize ilave olarak, metin içinde kullandıkları ifade köklerinin (lemmatization işlemi yapılarak) birbirlerine ne derece benzerlik taşıdığı TF-IDF yöntemiyle vektörel düzlemde bulunmuş olup, şekil 3.177’de görülmektedir.

Şekil 3.177’de yer alan içerik benzerlik analizi incelendiğinde;

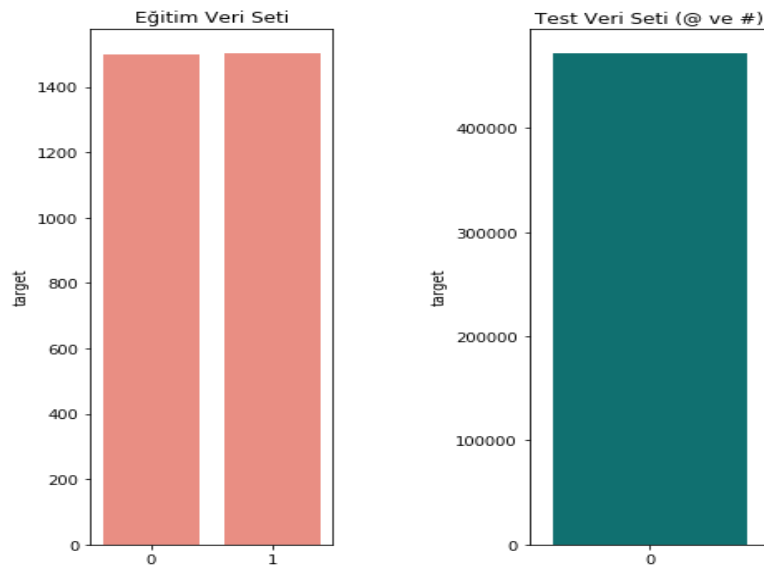
- Jandarma’nın (TR) içeriklerinin %38 oranında Polis (TR) ile benzerlik gösterdiği,
- Polis’in (TR) içeriklerinin %44 oranında Gendarmerie (FR) ile benzerlik gösterdiği,
- Carabinieri’nin (IT) içeriklerinin %64 oranında Polizia (IT) ile benzerlik gösterdiği,
- Gendarmerie’nin (FR) içeriklerinin %61 oranında Carabinieri (IT) ile benzerlik gösterdiği,
- Police’nin (FR) içeriklerinin %55 oranında Gendarmerie (FR) ile benzerlik gösterdiği,
- Guardia Civil’in (ES) içeriklerinin %74 oranında Policia (ES) ile benzerlik gösterdiği,
- İçerik benzeşme oranına göre en özgün konularda paylaşım yapan kolluğun Jandarma (TR) olduğu,
- Benzeşme oranlarını kolluk kuvveti bazında toplamı alındığında, benzerlik oranı en düşük ülkenin Türkiye olduğu ve bununla Fransa, İspanya ve İtalya örnekleminde halka verilen mesajların farklılaşmasından kaynaklı olduğu,
- Türkiye dışındaki diğer ülkelerde halka verilen mesajlardaki içeriklerin diğer uluslararası örneklerle daha çok benzerlik taşıdığı belirlenmiştir.

3.6. Yarı Denetimli Makine Öğrenimine Dayalı Siber Zorbalık Tespiti

Twitter örnekleminde sosyal ağlar üzerinde yapılan paylaşımlardan yola çıkarak makine öğrenmesi temelinde otomasyonel olarak paylaşımın siber zorbalık türü bir mesaj olup olmasının tespiti, çalışmanın amacıdır. Çalışmanın amacı kapsamında 19 (on dokuz) farklı makine öğrenimi kullanılmış olup, çalışma kapsamında elde edilen bulgular metotlarıyla birlikte alt maddelerde belirtilmektedir.

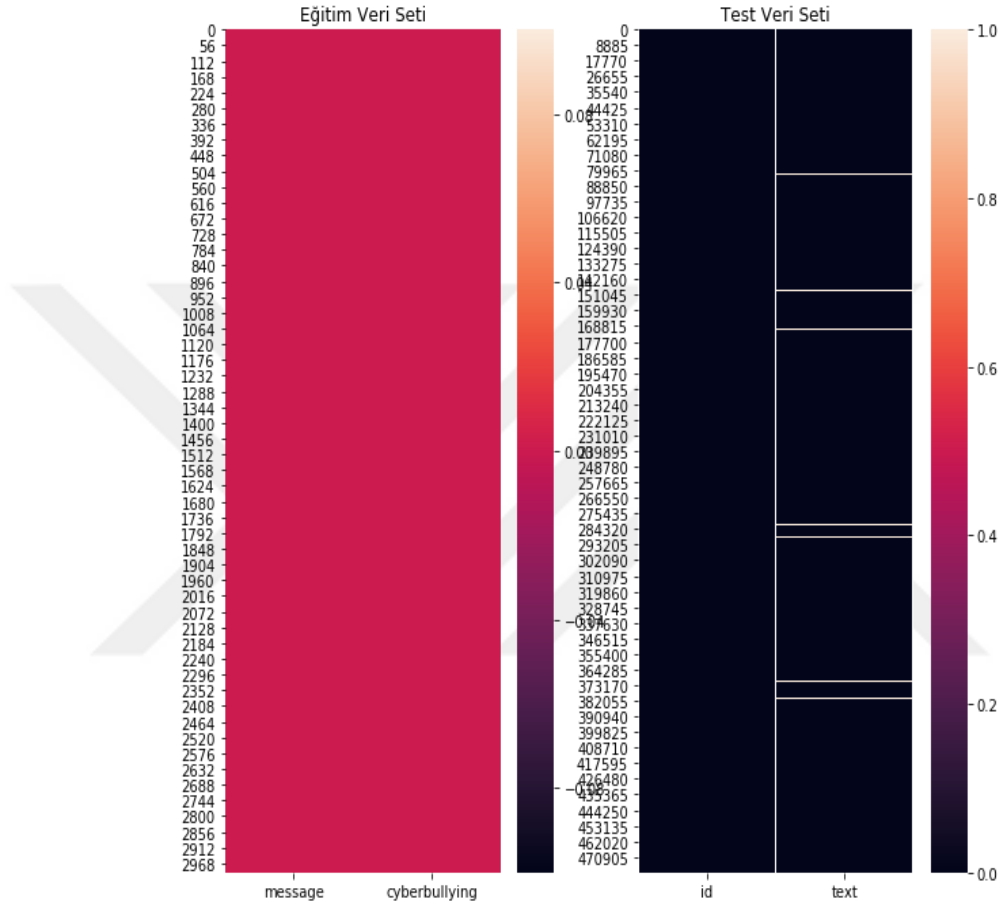
3.6.1. Eğitim ve Test Veri Seti Üzerinde Keşifsel Veri Analizi

Türkçe veri seti geliştirmeye yönelik Bozyiğit (2018) tarafından yayınlanan çalışma, eğitim verisi olarak kullanılmıştır. Analiz kapsamında test veri seti ise 479779 adet Tweet'den oluşmaktadır (Twitter API üzerinden 10 Eylül 2019 - 10 Aralık 2019 tarihleri arasında Türkçe dil süzgeci içererek, içinde etiketleme imgesi olarak kullanılan “@” simgesi içeren 353.874 adet Tweet ile 04 Ekim 2019 - 19 Kasım 2019 tarihleri arasında Türkçe dil süzgeci içererek, içinde hashtag imgesi olarak kullanılan “#” simgesi içeren 125.905 adet Tweet mevcuttur).



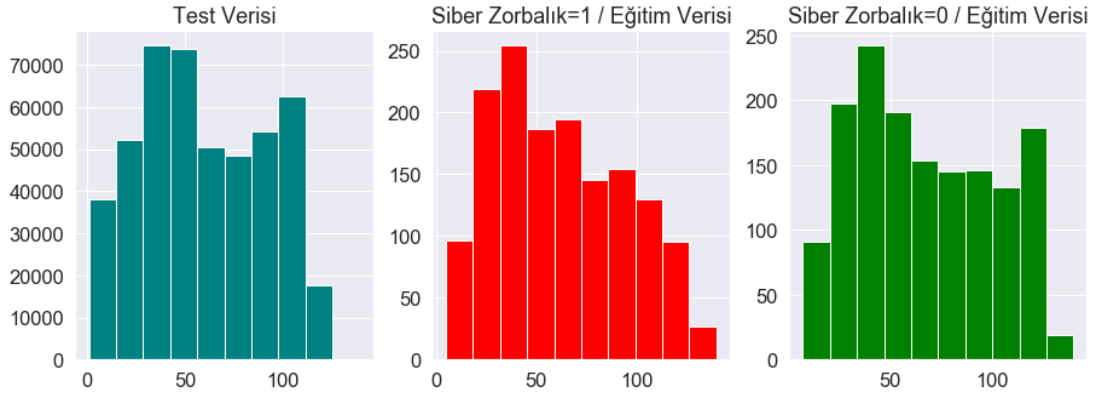
Şekil 3.178. Tespitte kullanılan veri setlerinin etiket bazında dağılımı

Şekil 3.178’de görüldüğü üzere, kullanılan veri setinde olan 3000 paylaşımın 1500’ünün siber zorbalık içerdiği (Etiket:1), kalan 1500’ün ise siber zorbalık içermediği (Etiket:0) belirlenmiş olup, test veri setinde ise siber zorbalık tahminlemesi yapılmadığı için etiket değeri 0 olarak yer almaktadır.

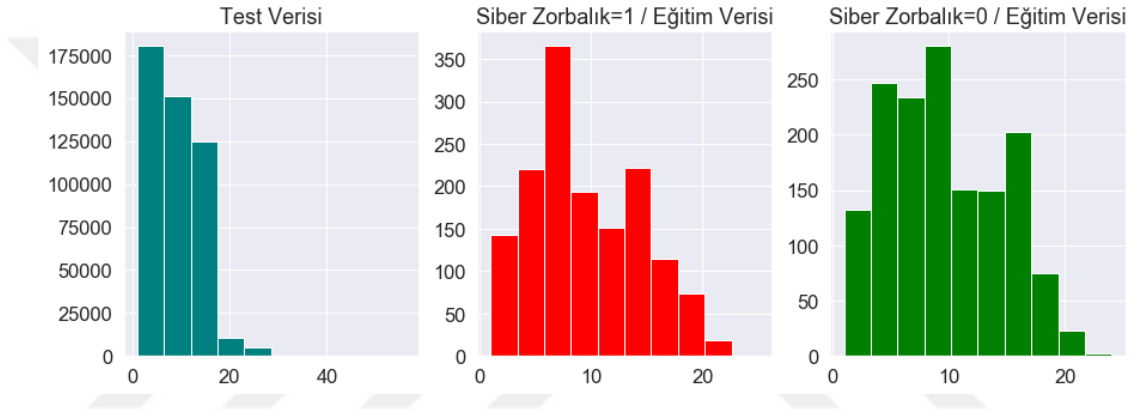


Şekil 3.179. Veri setinde yer alan boş değerler

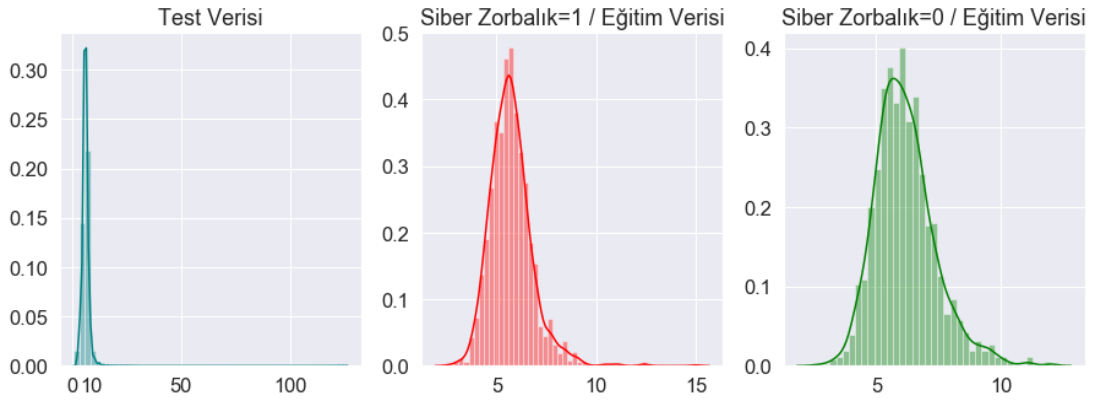
Test ve eğitim veri setleri gereç ve yöntem kısmında belirtilen ön işleme tabi tutulmuş olup, ön işlem sonucu elde edilen değerlerde boş olarak (NaN value) görünen değerlerin tespiti şekil 3.179’da yer alan diyagramda gösterilmiştir. Bu kapsamda söz konusu şekil incelendiğinde, eğitim veri setinde kayıp olmadığı, test eğitim veri setinde ise “id” sütununda eksik veri olmadığı, fakat “text” sütununda beyaz renkle gösterilen 7779 metin (Test veri setinin %0.01’ine tekâmül etmektedir.), boş olduğu için yani siber zorbalık da içermediği için analizden çıkarılmış olup, yapılan incelemeler kalan 472.000 Tweet üzerinden yapılmıştır.



Şekil 3.180. Tweetler içindeki karakter sayıları

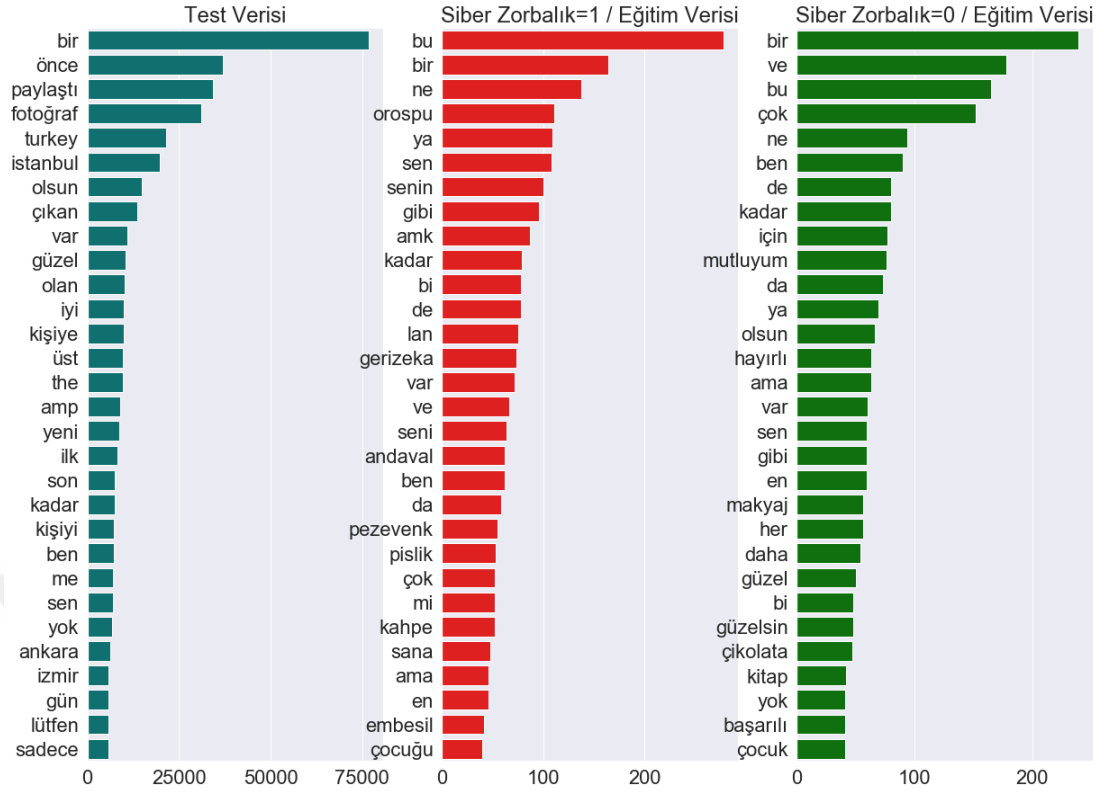


Şekil 3.181. Tweetler içindeki kelime sayıları

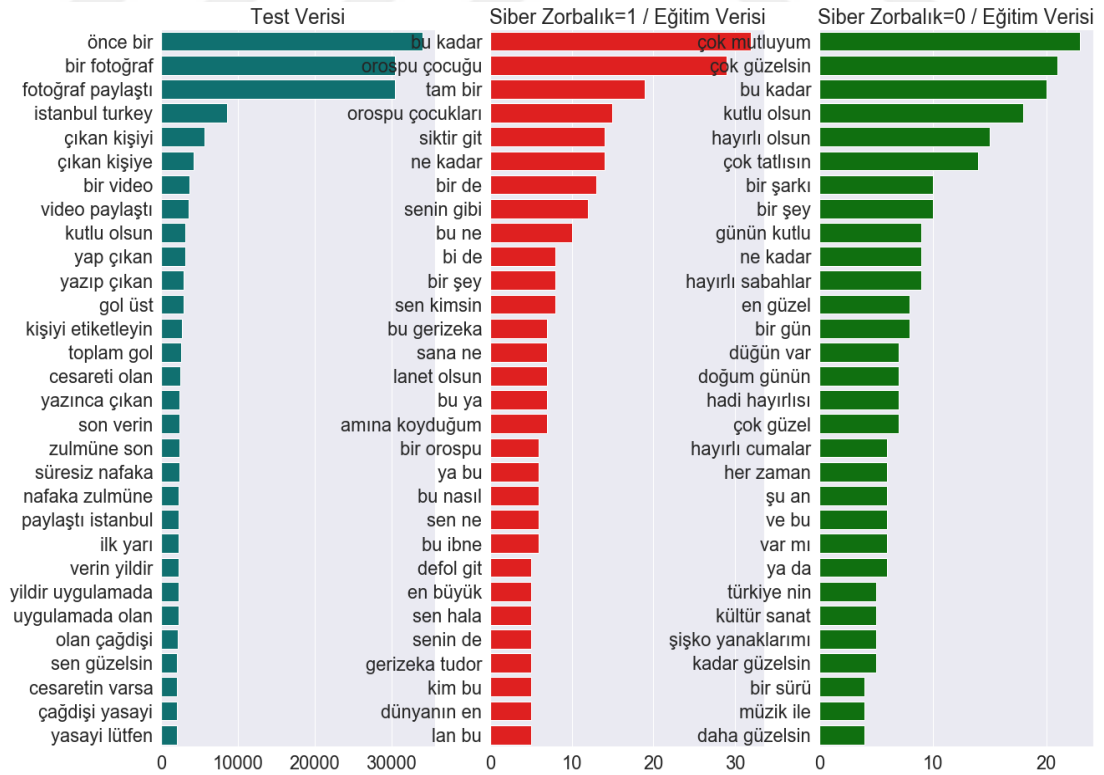


Şekil 3.182. Her bir tweet içindeki kelime uzunluğu

Eğitim ve test verisi üzerinde şekil 3.180, 3.181 ve 3.182 incelendiğinde, tweetler içindeki karakter sayısının 0-100 arasında, tweetler içindeki kelime sayılarının 0-20 arasında ve her bir tweet içindeki kelime uzunluğunun da 0-10 arasında ağırlık gösterdiği görülmektedir.



Şekil 3.183. Tweetler içinde en sık kullanılan kelimeler (n-gram, n:1)



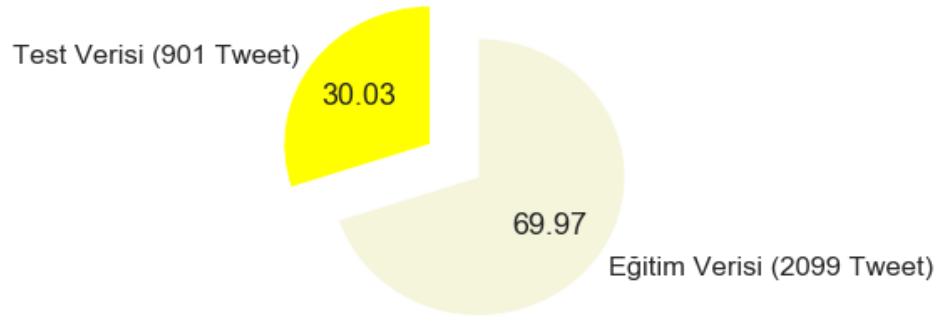
Şekil 3.184. Tweetler içinde en sık kullanılan iki kelimeler (n-gram, n:2)

Şekil 3.183’de tek gramlık modelleme üzerinden, şekil 3.184’de ise iki gramlık (bigram) modelleme üzerinden en çok kullanılan 30 kelime/kelimeler (ikili kelime) görülmekte olup, özellikle eğitim verisi içinde siber zorbalık içeren mesajlar ve siber zorbalık içermeyen mesajlar arasında fark gözlemlenmektedir.

3.6.2. Eğitim Veri Seti Üzerinde Makine Öğrenimi Modelleri

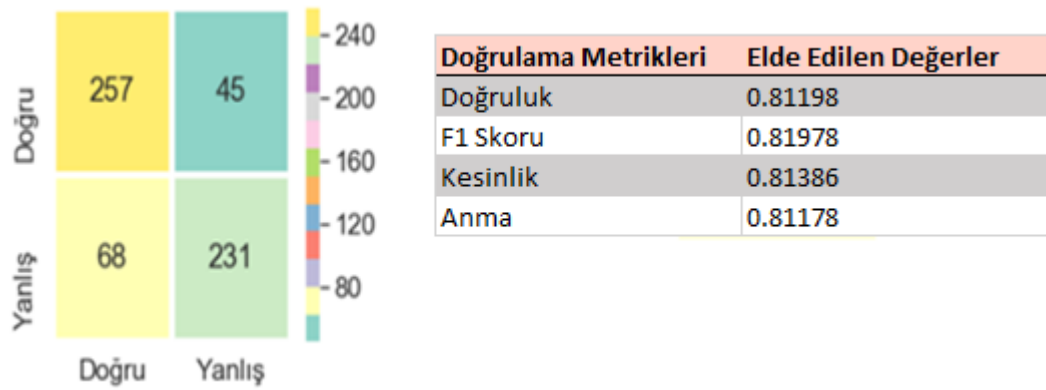
Test verisi üzerinden yarı denetimli öğrenmenin en iyi sınıflamayla gerçekleşmesi için eğitim verisi üzerinde öğrenmenin farklı algoritmalar temelinde optimum düzeyde olması sistemin ideal düzenidir. Bu sebeple eğitim veri seti üzerinde çeşitli makine öğrenimi modelleri sırayla uygulanmış olup, kullanılan modeller;

- Gaussian Naive Bayes,
- Multinomial Naive Bayes,
- Bernoulli Naive Bayes,
- Karar Ağacı Modeli (Decision Tree Model),
- Ekstra Ağaçlar Sınıflandırması (Extra Trees Classifier),
- Lineer Diskriminant Analizi (Linear Discriminant Analysis),
- Karesel Diskriminant Analizi (Quadratic Discriminant Analysis),
- AdaBoost (Adaptive Boosting),
- Gradyan Artırma Modeli (Gradient Boosting Model),
- Rastgele Orman (Random Forest),
- Lojistik Regresyon Modeli (Logistic Regression Model),
- Perceptron Sinir Ağı (Perceptron Neural Network),
- Lineer Karar Destek Sınıflandırması (Linear Support Vector Classification),
- XGBoost,
- K-En Yakın Komşu Modeli (K - Nearest Neighbors Model (KNN)),
- Destek Vektör Makineleri (Support Vector Machine (SVM)),
- Stokastik Gradyan Azaltma (Stochastic Gradient Descent (SGD)),
- Light Gradyan Artırma Modeli (Light Gradient Boosting Model (LGBM)),
- Votin Sınıflandırma Modeli (Voting Classifier Model) olarak yer almaktadır.



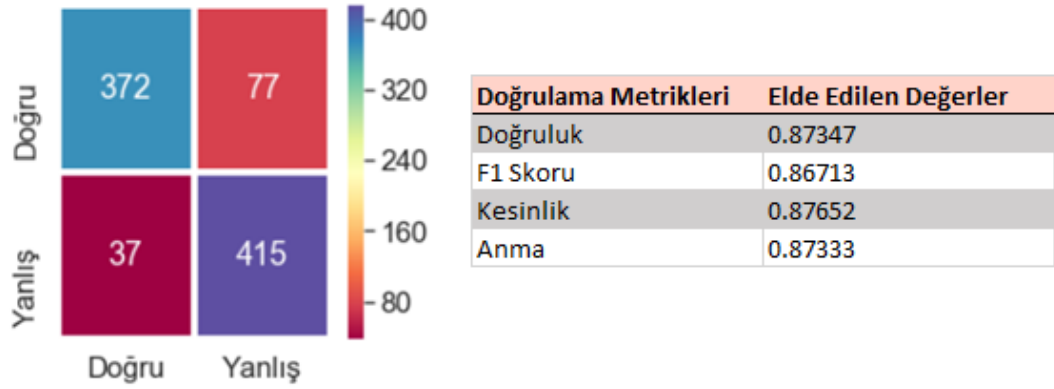
Şekil 3.185. Eğitim setinin test ve eğitim olarak bölünmesi

Veri üzerinde ön temizleme işlemine müteakip, içinde siber zorbalık etiketlerini barındıran veri seti, öğrenmenin denetlenebilmesi adına şekil 3.185’de yer aldığı şekliyle %30’u test olmak üzere bölünmüş, bölünme sonrası 2099 tweet öğrenme için ayrılmış, geri kalan 901 tweet de öğrenme sonrası öğrenmenin başarısını sınamak için ayrılmıştır. Test veri seti üzerinden elde edilen değerler, kullanılan parametre değerleriyle birlikte karışıklık matrisi ile doğruluk, kesinlik, anma ve F1 skoru şeklinde aşağıda belirtilmektedir.



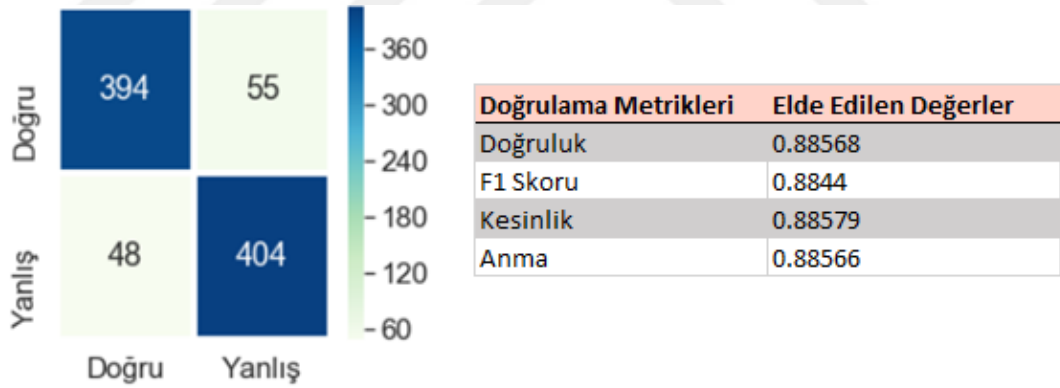
Şekil 3.186. Gaussian naive bayes ile elde edilen değerler

Şekil 3.186’da görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (*, priors=None, var_smoothing=1e-09). Öğrenme sonucunda %81.198 doğruluk ve harmonik ortalama olan %81.978 F1 skoru elde edilmiştir.



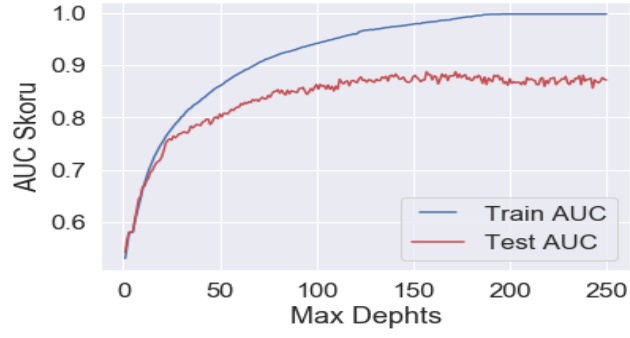
Şekil 3.187. MultinomialNB ile elde edilen değerler

Şekil 3.187’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (*, alpha=1.0, fit_prior=True, class_prior=None). Öğrenme sonucunda %87.347 doğruluk ve harmonik ortalama olan %86.713 F1 skoru elde edilmiştir.

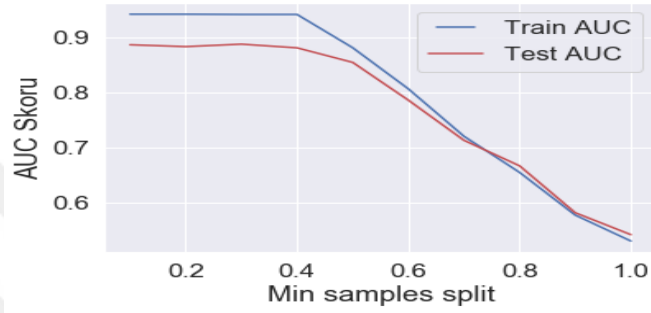


Şekil 3.188. BernoulliNB ile elde edilen değerler

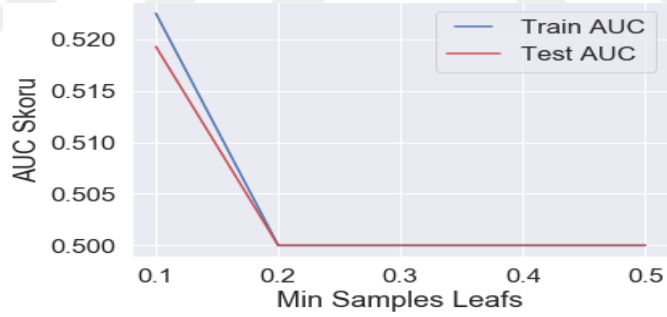
Şekil 3.188’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (*, alpha=1.0, binarize=0.0, fit_prior=True, class_prior=None). Öğrenme sonucunda %88.568 doğruluk ve harmonik ortalama olan %88.44 F1 skoru elde edilmiştir.



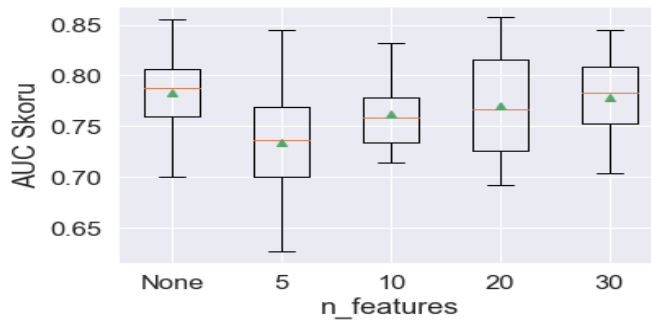
Şekil 3.189. Karar ağacında max dephts ile AUC skoru



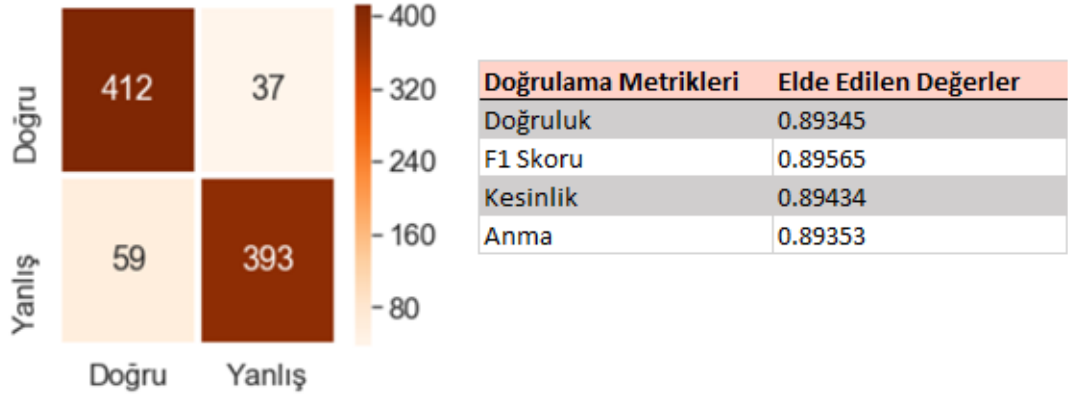
Şekil 3.190. Karar ağacında min samples split ile AUC skoru



Şekil 3.191. Karar ağacında min samples leafs ile AUC skoru

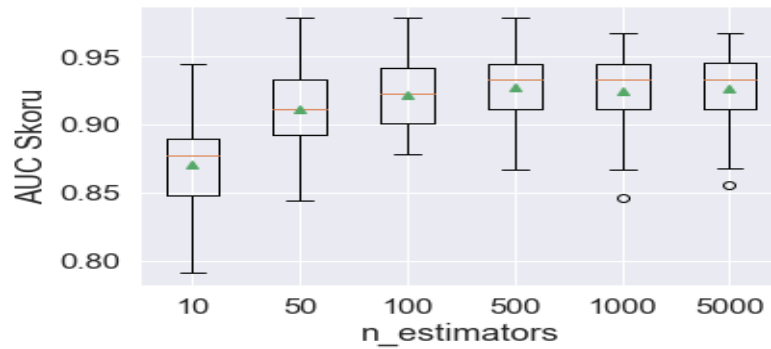


Şekil 3.192. Karar ağacında n_features ile AUC skoru

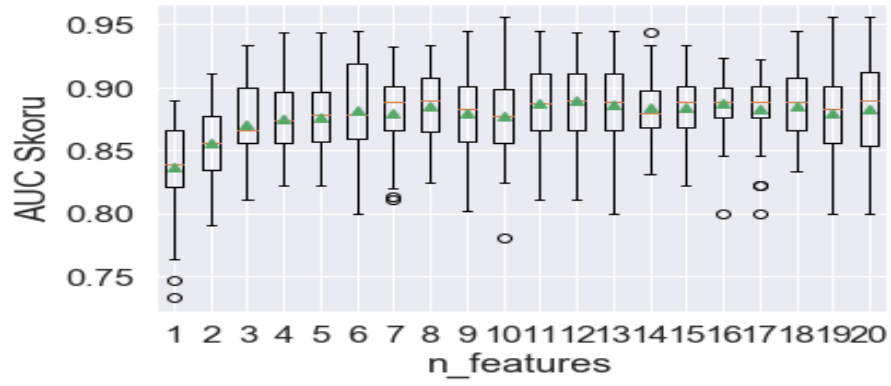


Şekil 3.193. Karar ağacı ile elde edilen değerler

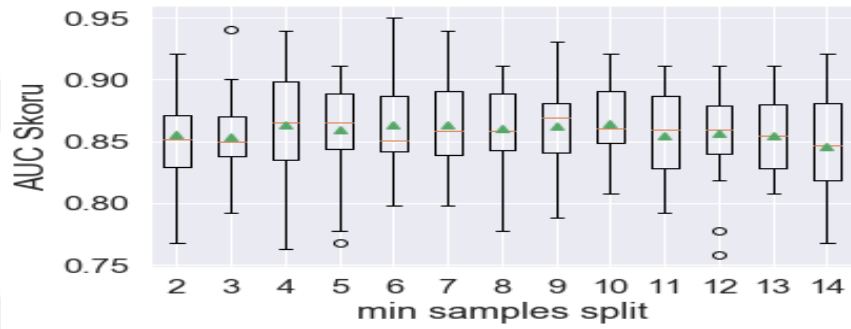
Karar ağacı modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.189’den max depths 180, şekil 3.190’dan min samples split 0.2, şekil 3.191’den min samples leafs için 0.1 dışında değerler stabil olduğu için varsayılan 1, şekil 3.192’den n features hiç (none) olarak alınmıştır. Bu kapsamda şekil 3.193’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (*, criterion='entropy', splitter='best', max_depth=180, min_samples_split=0.2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=42, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, presort='deprecated', ccp_alpha=0.0). Öğrenme sonucunda %89.345 doğruluk ve harmonik ortalama olan %89.565 F1 skoru elde edilmiştir.



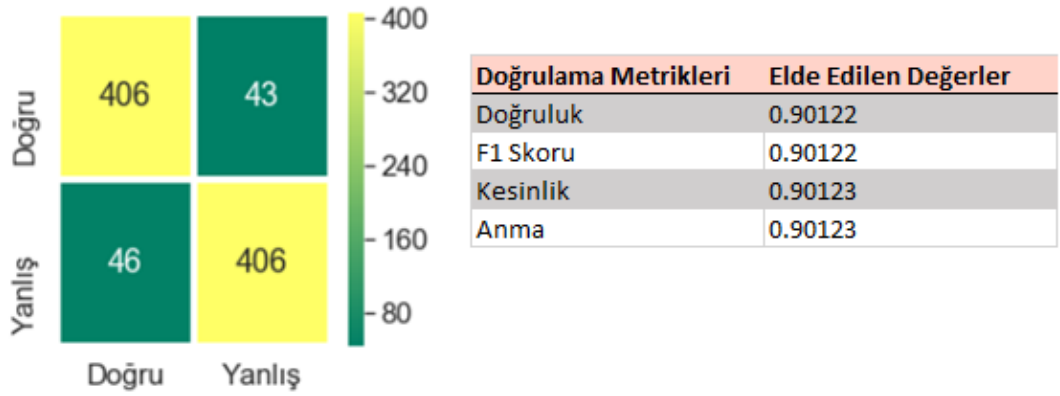
Şekil 3.194. Ekstra ağaçlarda n_estimators ile AUC skoru



Şekil 3.195. Ekstra ağaçlarda n_features ile AUC skoru



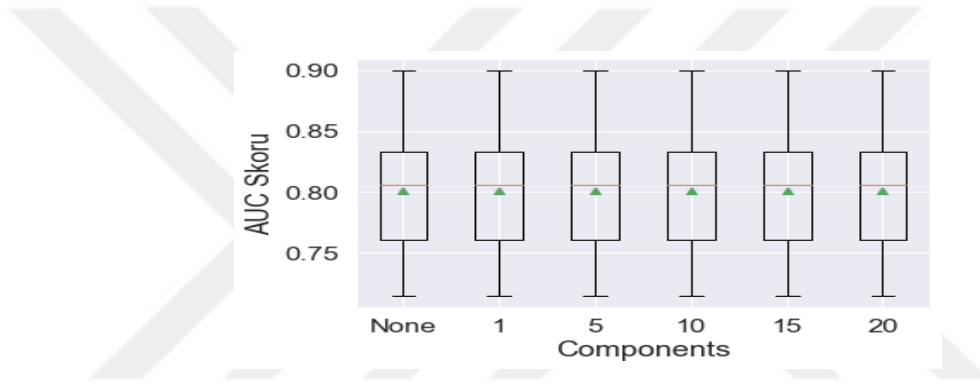
Şekil 3.196. Ekstra ağaçlarda min samples split ile AUC skoru



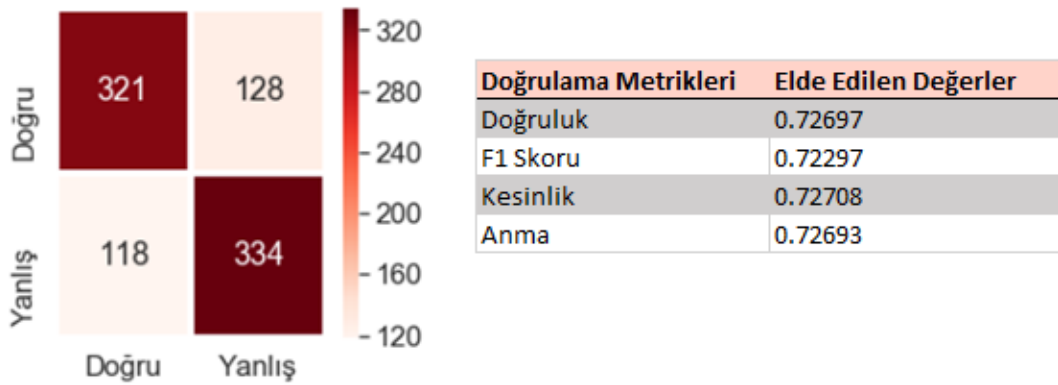
Şekil 3.197. Ekstra ağaçlar ile elde edilen değerler

Eksra ağaçlar modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.194’den n estimators 1000, şekil 3.195’den n features 18, şekil 3.196’dan

min samples split 9 olarak alınmıştır. Bu kapsamda şekil 3.197’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (n_estimators=1000, *, criterion='entropy', max_depth=None, min_samples_split=9, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=18, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=False, oob_score=False, n_jobs=None, random_state=42, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None). Öğrenme sonucunda %90.122 doğruluk ve harmonik ortalama olan %90.122 F1 skoru elde edilmiştir.



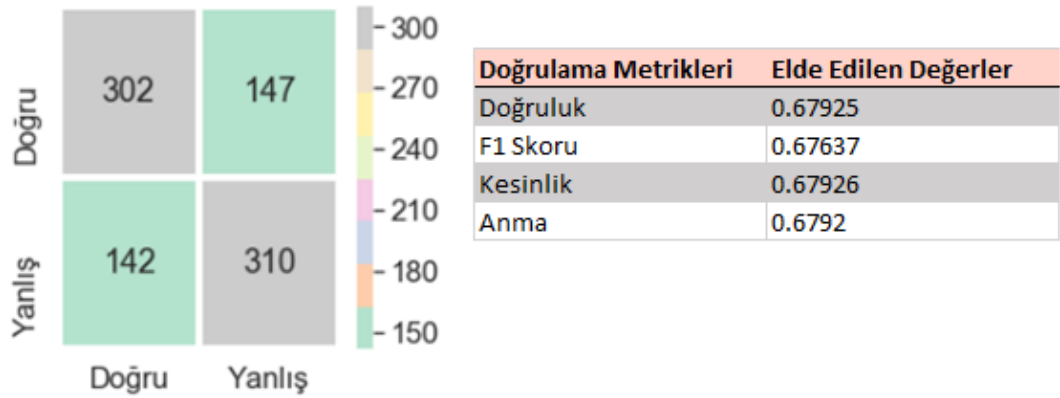
Şekil 3.198. Lineer diskriminant analizde components ile AUC skoru



Şekil 3.199. Lineer diskriminant analiz ile elde edilen değerler

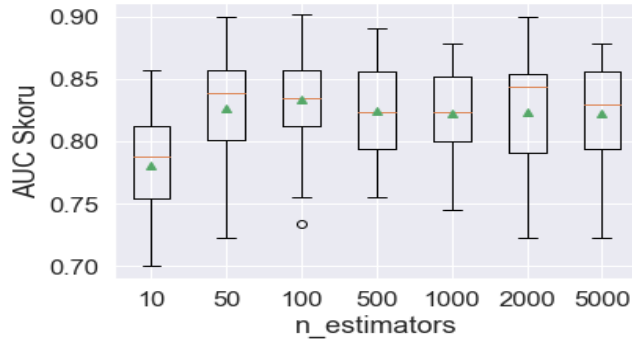
Lineer diskriminant modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.198’de component değerleri incelenmiş, fakat optimizasyonda anlamlı fark

ortaya çıkmadığı için varsayılan olan none değeri alınmıştır. Bu kapsamda şekil 3.199’da görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (*, solver='svd', shrinkage=None, priors=None, n_components=None, store_covariance=False, tol=0.0001). Öğrenme sonucunda %72.697 doğruluk ve harmonik ortalama olan %72.297 F1 skoru elde edilmiştir.

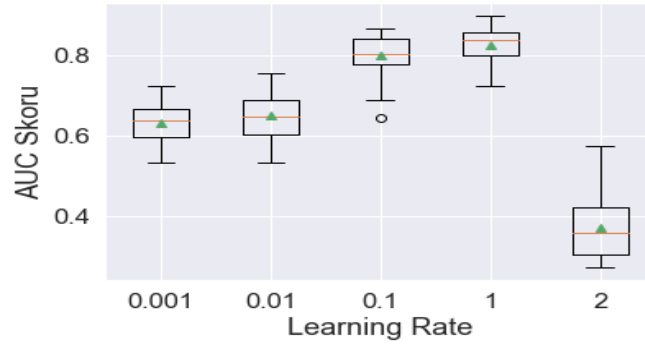


Şekil 3.200. Karesel diskriminant analiz ile elde edilen değerler

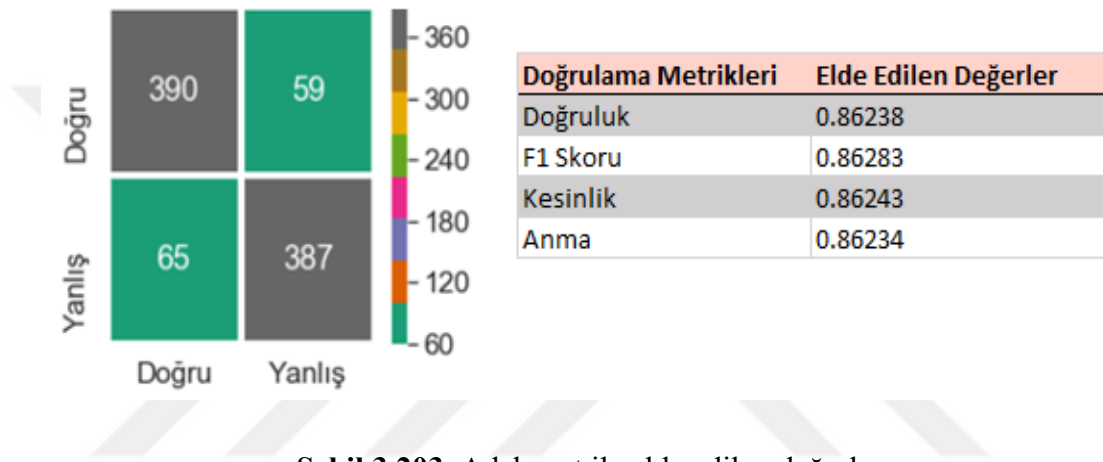
Karesel diskriminant analizi ile şekil 3.200’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (*, priors=None, reg_param=0.0, store_covariance=False, tol=0.0001). Öğrenme sonucunda %67.925 doğruluk ve harmonik ortalama olan %67.637 F1 skoru elde edilmiştir.



Şekil 3.201. Adaboost sınıflandırıcıda n estimators ile AUC skoru

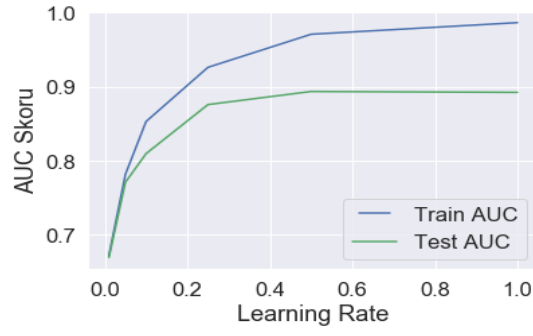


Şekil 3.202. Adaboost sınıflandırıcıda öğrenme oranı (learning rate) ile AUC skoru

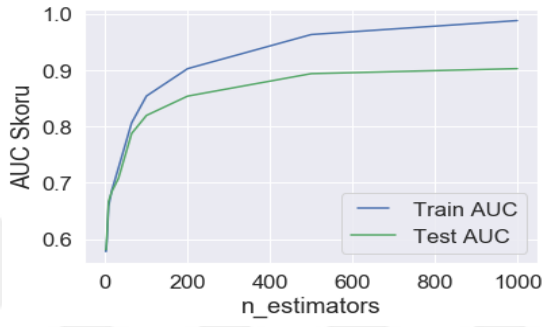


Şekil 3.203. Adaboost ile elde edilen değerler

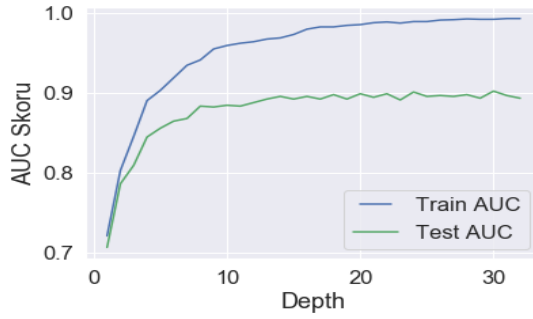
Adaboost modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.201’de n estimators değeri 2000, şekil 3.202’de öğrenme oranı (learning rate) 1 olarak alınmıştır. Bu kapsamda şekil 3.203’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (base_estimator=None, *, n_estimators=2000, learning_rate=1.0, algorithm='SAMME.R', random_state=None). Öğrenme sonucunda %86.238 doğruluk ve harmonik ortalama olan %86.283 F1 skoru elde edilmiştir.



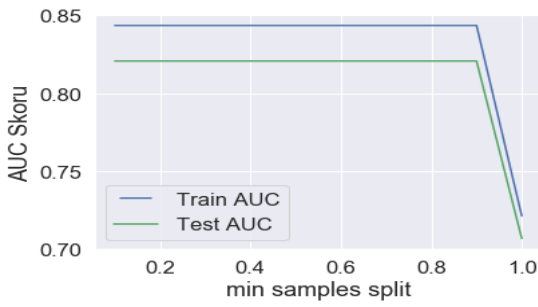
Şekil 3.204. Gradyan artırmada öğrenme oranı (learning rate) ile AUC skoru



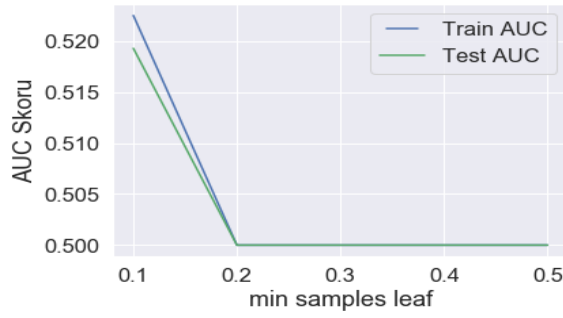
Şekil 3.205. Gradyan artırmada n estimators ile AUC skoru



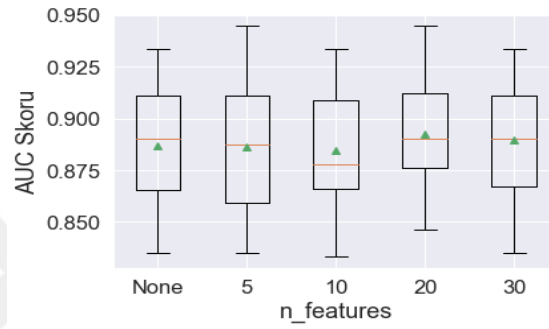
Şekil 3.206. Gradyan artırmada derinlik (depth) ile AUC skoru



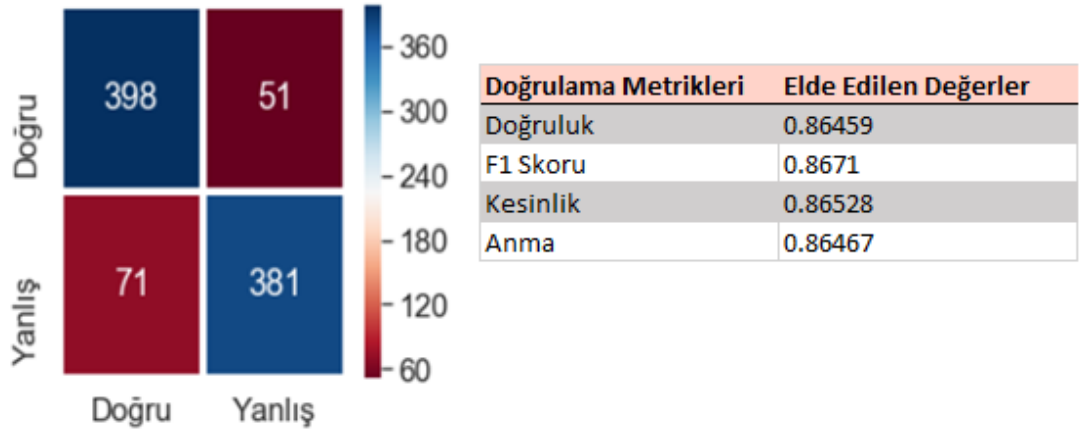
Şekil 3.207. Gradyan artırmada min samples split ile AUC skoru



Şekil 3.208. Gradyan artırmada min samples leaf ile AUC skoru



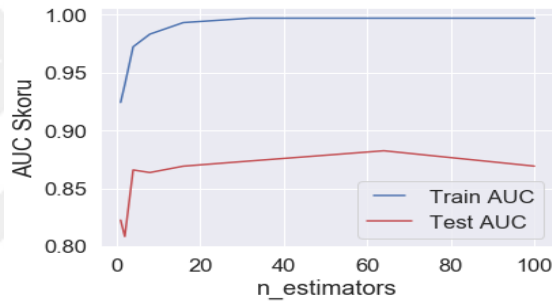
Şekil 3.209. Gradyan artırmada n features ile AUC skoru



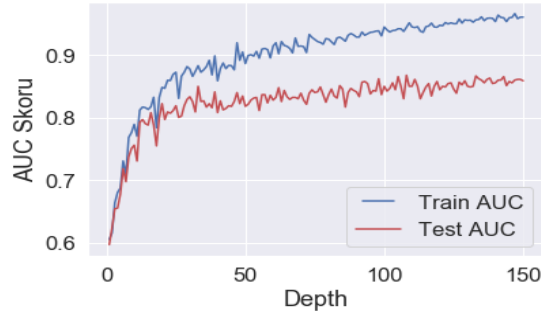
Şekil 3.210. Gradyan artırma ile elde edilen değerler

Gradyan artırma modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.204’de öğrenme oranı (learning rate) 0.5, şekil 3.205’de n estimators değeri 500, şekil 3.206’da derinlik (depth) 30, şekil 3.207’de min samples split 0.2 (Temel değer 2 alınmıştır.), şekil 3.208’de min samples leaf 0.1 (Temel değer 1 alınmıştır.),

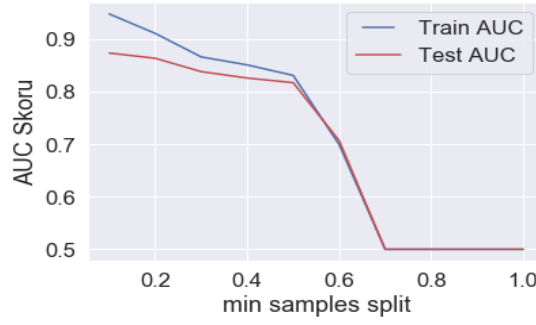
şekil 3.209’da n features 20 olarak alınmıştır. Bu kapsamda şekil 3.210’da görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (*, loss='deviance', learning_rate=0.5, n_estimators=500, subsample=1.0, criterion='friedman_mse', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_depth=30, min_impurity_decrease=0.0, min_impurity_split=None, init=None, random_state=42, max_features=20, verbose=0, max_leaf_nodes=None, warm_start=False, presort='deprecated', validation_fraction=0.1, n_iter_no_change=None, tol=0.0001, ccp_alpha=0.0). Öğrenme sonucunda %86.459 doğruluk ve harmonik ortalama olan %86.71 F1 skoru elde edilmiştir.



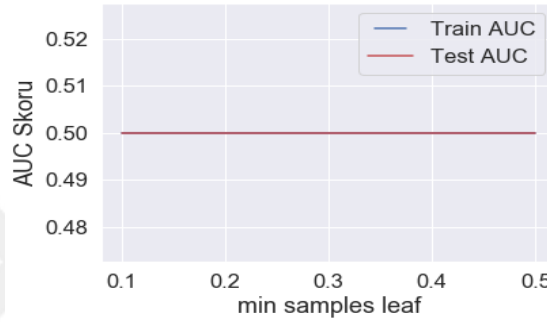
Şekil 3.211. Rastgele ormanda n_estimators ile AUC skoru



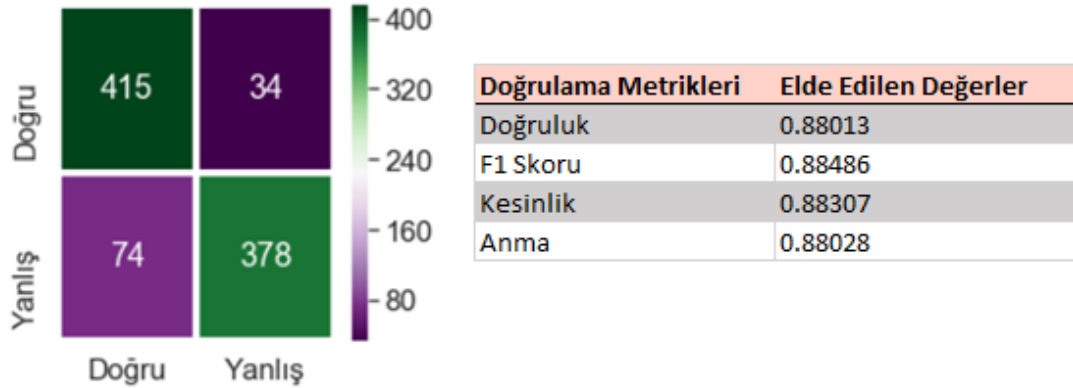
Şekil 3.212. Rastgele ormanda derinlik (depth) ile AUC skoru



Şekil 3.213. Rastgele ormanda min samples split ile AUC skoru



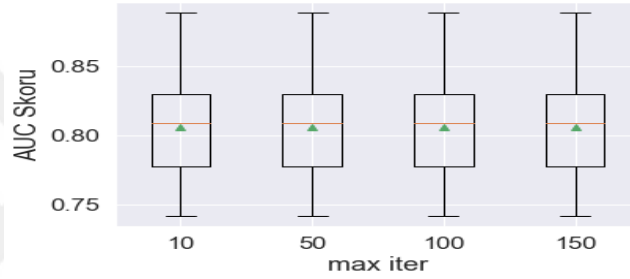
Şekil 3.214. Rastgele ormanda min samples leaf ile AUC skoru



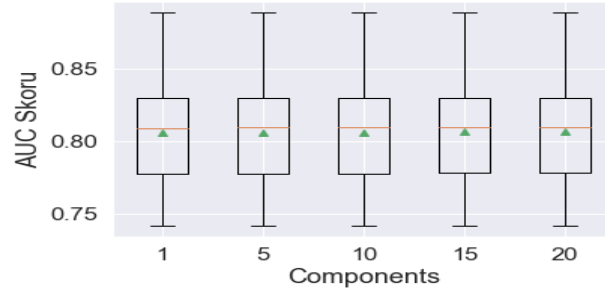
Şekil 3.215. Rastgele orman ile elde edilen değerler

Rastgele orman modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.211’de n estimators 65, şekil 3.212’de derinlik (depht) 110, şekil 3.213’de min samples split 0.2 (Temel değer 2 alınmıştır.), şekil 3.214’de min samples leaf stabil olduğundan 0.1 (Temel değer 1 alınmıştır.) olarak alınmıştır. Bu kapsamda şekil

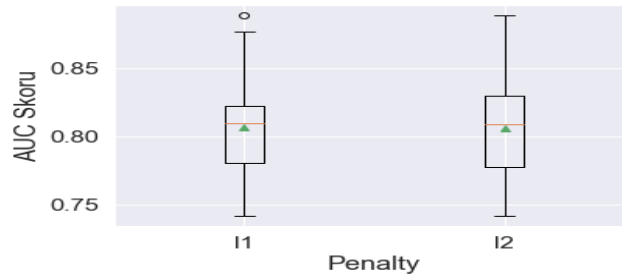
3.215’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (n_estimators=65, *, criterion=' entropy', max_depth=110, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='log2', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=42, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None). Öğrenme sonucunda %88.013 doğruluk ve harmonik ortalama olan %88.486 F1 skoru elde edilmiştir.



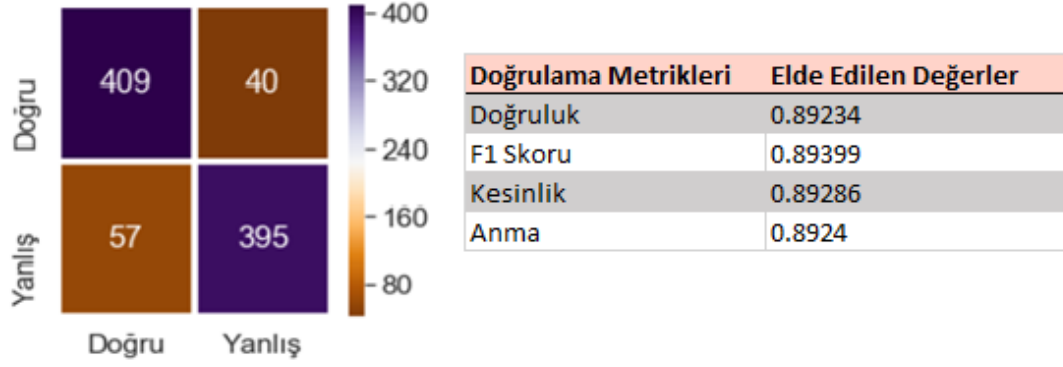
Şekil 3.216. Lojistik resgresyonda max iter ile AUC skoru



Şekil 3.217. Lojistik resgresyonda components ile AUC skoru

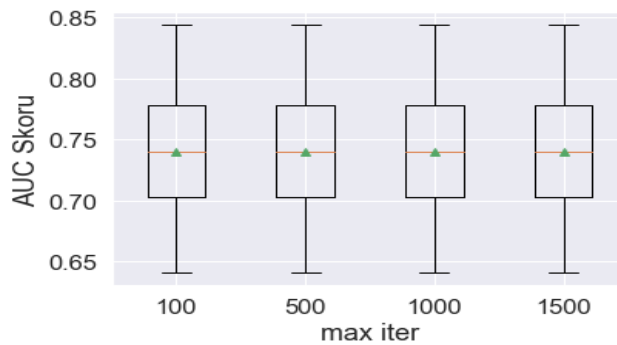


Şekil 3.218. Lojistik resgresyonda penalty ile AUC skoru

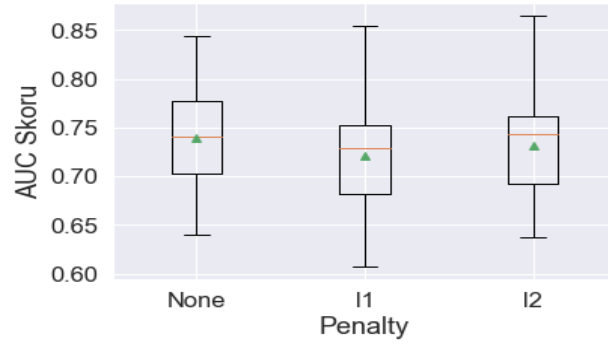


Şekil 3.219. Lojistik regresyon ile elde edilen değerler

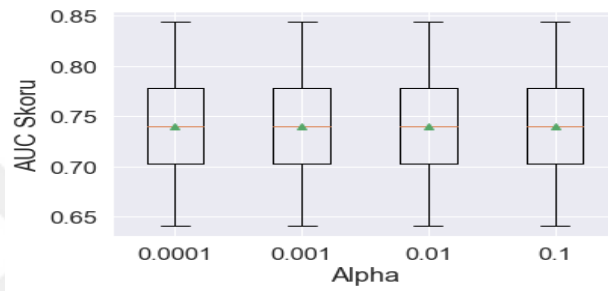
Lojistik regresyon modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.216’da max iter stabil olduğundan varsayılan değer olan 100, şekil 3.217’de components 110 stabil olduğundan varsayılan değer olan 1.0, şekil 3.218’de penalty l2 olarak alınmıştır. Bu kapsamda şekil 3.219’da görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (penalty='l2', *, dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=42, solver='lbfgs', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None). Öğrenme sonucunda %89.234 doğruluk ve harmonik ortalama olan %89.399 F1 skoru elde edilmiştir.



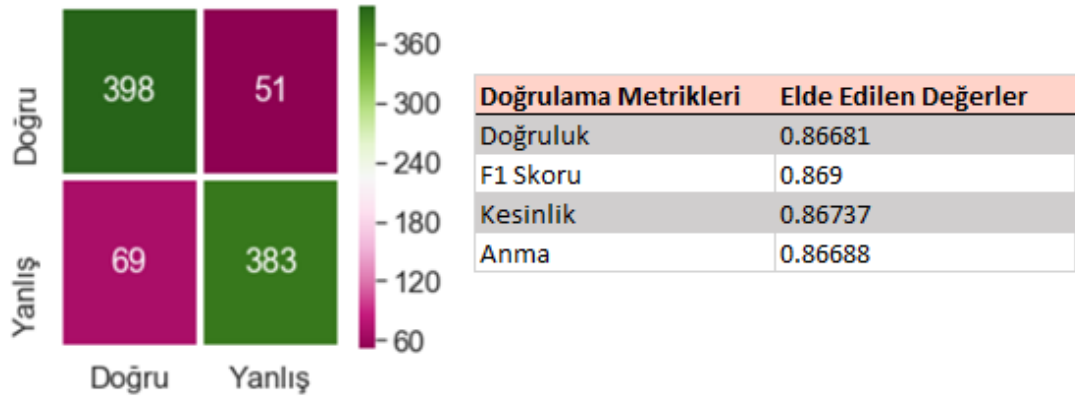
Şekil 3.220. Perceptron modelinde max iter ile AUC skoru



Şekil 3.221. Perceptron modelinde penalty ile AUC skoru



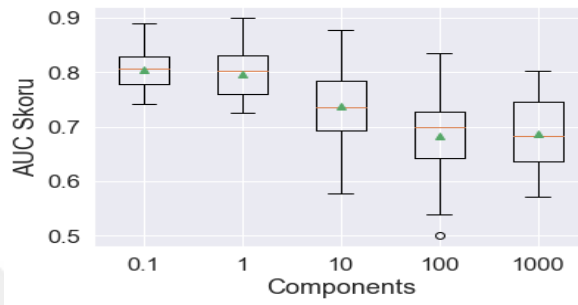
Şekil 3.222. Perceptron modelinde alpha ile AUC skoru



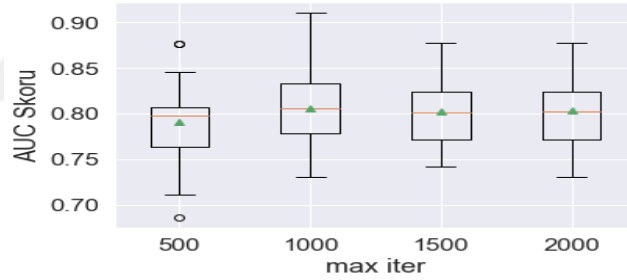
Şekil 3.223. Perceptron regresyon ile elde edilen değerler

Perceptron modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.220’de max iter stabil olduğundan varsayılan değer olan 1000, şekil 3.221’de penalty none, şekil 3.222’de alpha stabil olduğundan varsayılan değer 0.0001 olarak alınmıştır. Bu kapsamda şekil 3.223’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler

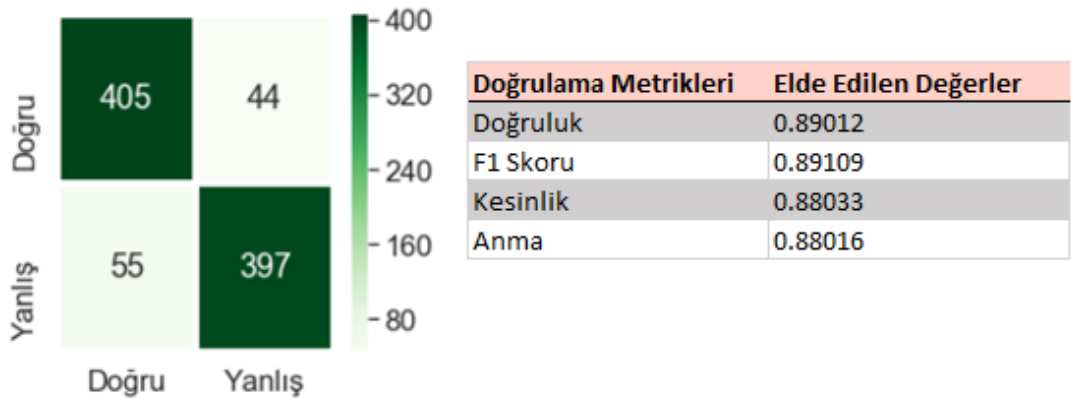
çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (*, penalty=None, alpha=0.0001, fit_intercept=True, max_iter=1000, tol=0.001, shuffle=True, verbose=0, eta0=1.0, n_jobs=None, random_state=42, early_stopping=False, validation_fraction=0.1, n_iter_no_change=5, class_weight=None, warm_start=False). Öğrenme sonucunda %86.681 doğruluk ve harmonik ortalama olan %86.9 F1 skoru elde edilmiştir.



Şekil 3.224. Lineer SVC modelinde components ile AUC skoru

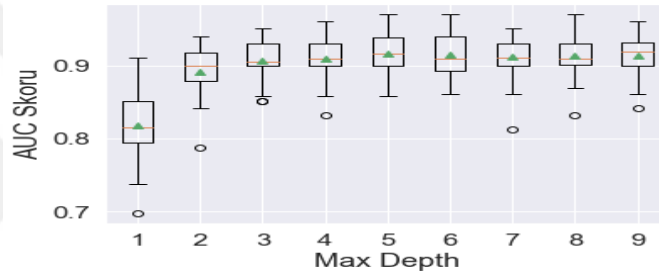


Şekil 3.225. Lineer SVC modelinde max iter ile AUC skoru

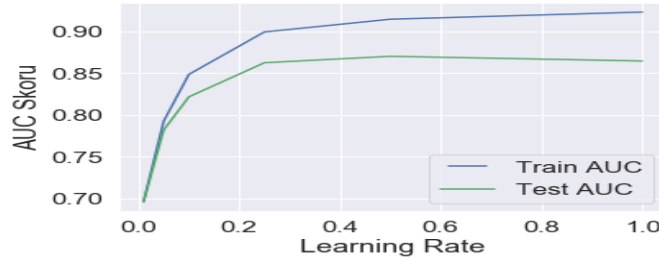


Şekil 3.226. Lineer SVC regresyon ile elde edilen değerler

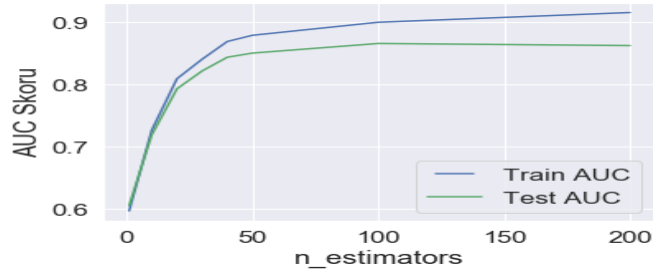
Lineer karar destek sınıflandırması modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.224’de components 1, şekil 3.225’de max iter 1000 olarak alınmıştır. Bu kapsamda şekil 3.226’da görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (penalty='l2', loss='squared_hinge', *, dual=True, tol=0.0001, C=1.0, multi_class='ovr', fit_intercept=True, intercept_scaling=1, class_weight=None, verbose=0, random_state=42, max_iter=1000). Öğrenme sonucunda %89.012 doğruluk ve harmonik ortalama olan %89.109 F1 skoru elde edilmiştir.



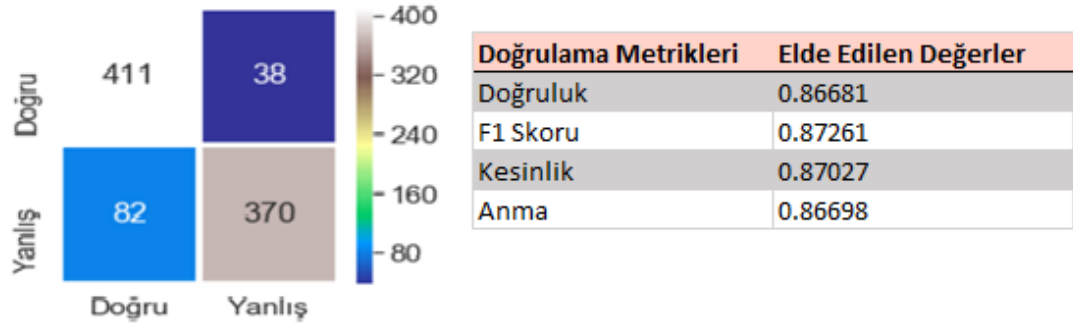
Şekil 3.227. XGBoost modelinde derinlik (max depth) ile AUC skoru



Şekil 3.228. XGBoost modelinde öğrenme oranı (learning rate) ile AUC skoru

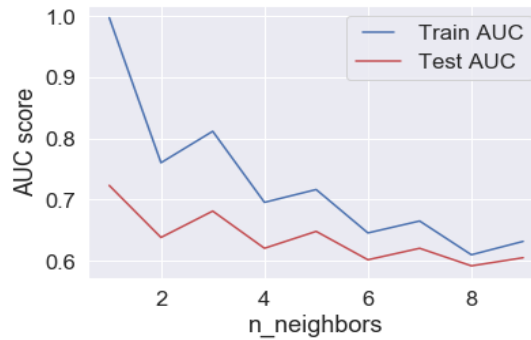


Şekil 3.229. XGBoost modelinde n estimators ile AUC skoru

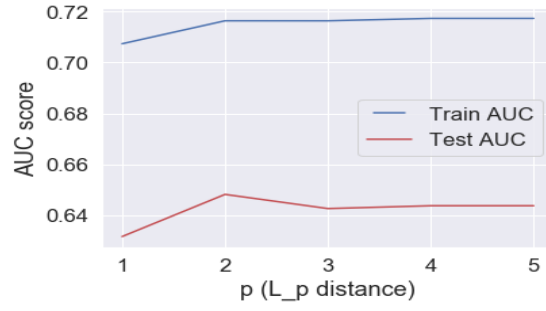


Şekil 3.230. XGBoost ile elde edilen değerler

XGBoost modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.227’de maksimum derinlik 5, şekil 3.228’de öğrenme hızı 0.5, şekil 3.229’da n estimators 100 olarak alınmıştır. Bu kapsamda şekil 3.230’da görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (base_score=0.5, colsample_bylevel=1, colsample_bytree=1, gamma=0, learning_rate=0.5, max_delta_step=0, max_depth=5, min_child_weight=1, missing=None, n_estimators=100, nthread=-1, objective='binary:logistic', reg_alpha=0, random_state=42, reg_lambda=1, scale_pos_weight=1, seed=0, silent=True, subsample=1). Öğrenme sonucunda %86.681 doğruluk ve harmonik ortalama olan %87.261 F1 skoru elde edilmiştir.



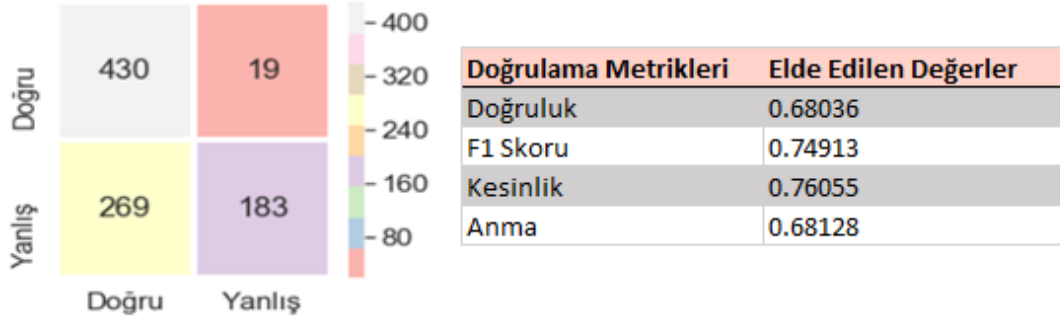
Şekil 3.231. K-en yakın komşu modelinde n komşu (n neighbors) ile AUC skoru



Şekil 3.232. K-en yakın komşu modelinde p değeri ile AUC skoru



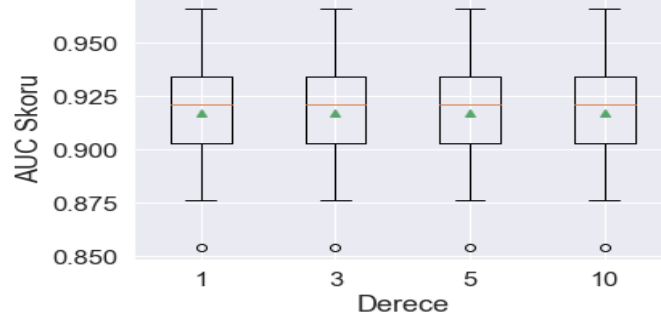
Şekil 3.233. K-en yakın komşu modelinde leaf size ile AUC skoru



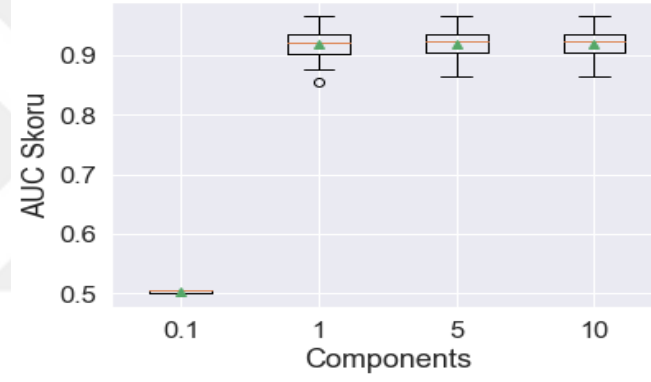
Şekil 3.234. K-en yakın komşu ile elde edilen değerler

K-en yakın komşu (KNN) modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.231’de n komşu 3, şekil 3.232’de p değeri 2 ve şekil 3.233’de leaf size 20 olarak alınmıştır. Bu kapsamda şekil 3.234’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (n_neighbors=3, *, weights='uniform', algorithm='auto', leaf_size=20, p=2, metric='minkowski',

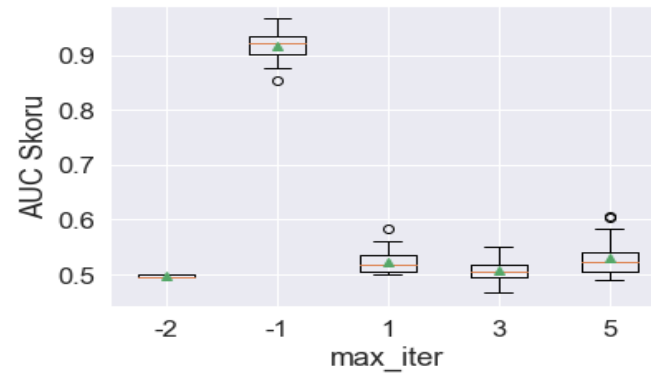
metric_params=None, n_jobs=None, **kwargs). Öğrenme sonucunda %68.036 doğruluk ve harmonik ortalama olan %74.913 F1 skoru elde edilmiştir.



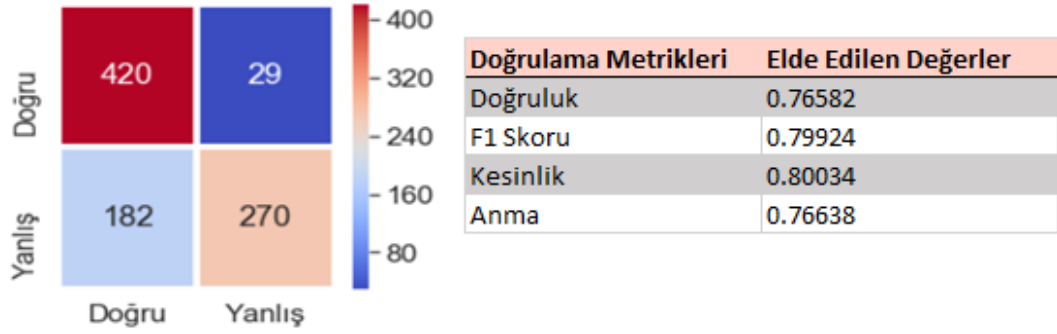
Şekil 3.235. SVM modelinde derece ile AUC skoru



Şekil 3.236. SVM modelinde components ile AUC skoru

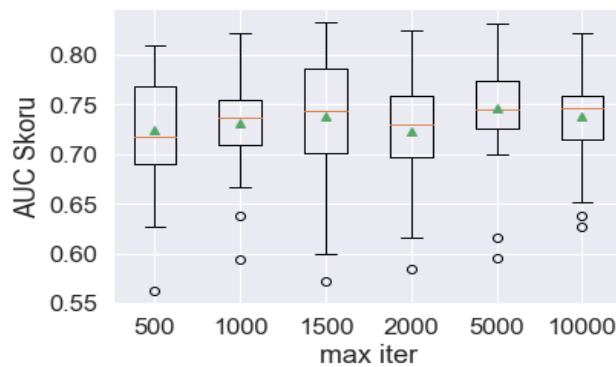


Şekil 3.237. SVM modelinde max iter ile AUC skoru

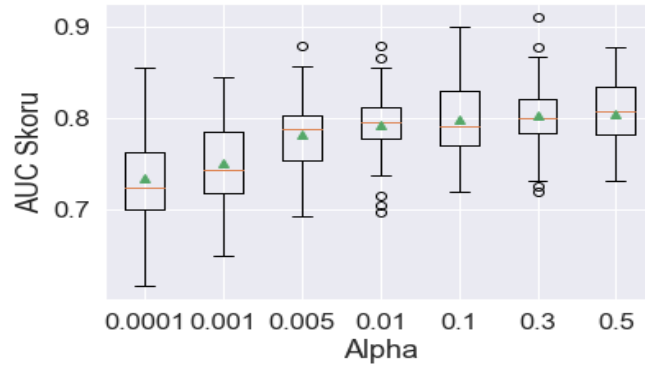


Şekil 3.238. SVM ile elde edilen değerler

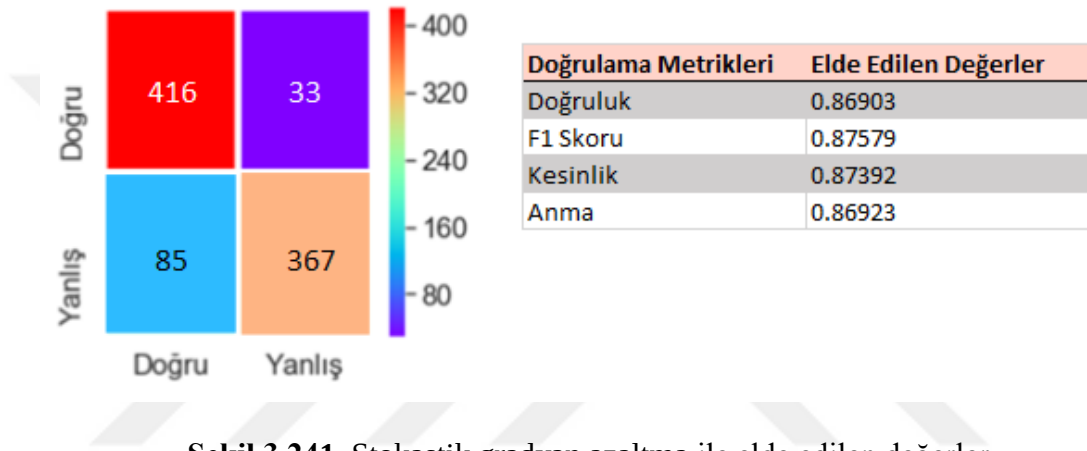
Destek vektör makineleri (SVM) modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.235’de derece 3, şekil 3.236’da components 5 ve şekil 3.237’de max iter -1 olarak alınmıştır. Bu kapsamda şekil 3.238’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (*, C=5.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=True, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', break_ties=False, random_state=42). Öğrenme sonucunda %76.582 doğruluk ve harmonik ortalama olan %79.924 F1 skoru elde edilmiştir.



Şekil 3.239. Stokastik gradyan azaltma modelinde max iter ile AUC skoru

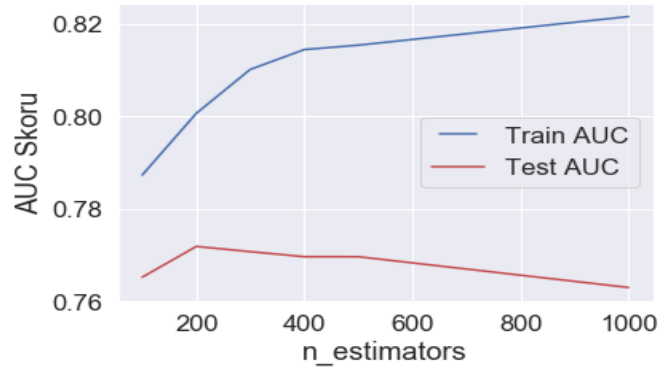


Şekil 3.240. Stokastik gradyan azaltma modelinde alpha ile AUC skoru

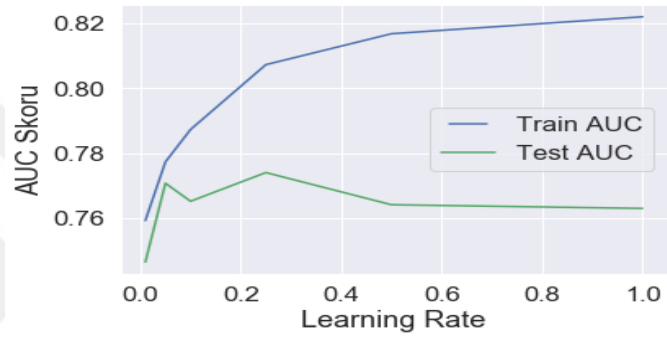


Şekil 3.241. Stokastik gradyan azaltma ile elde edilen değerler

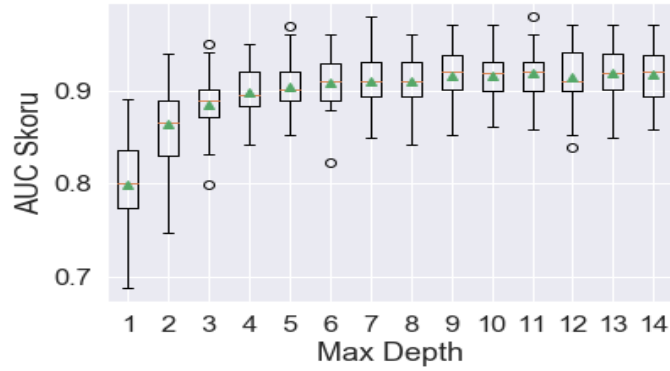
Stokastik gradyan azaltma (SGD) modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.239’da max iter 1500 ve şekil 3.240’da alpha 0.01 (0.01 ile 0.5 arası değerler, birbirine oldukça yakın olup, ufak farklılığın veri gürültüsünden kaynaklandığı değerlendirilmekte olup, en düşük optimize değer alınmıştır.) olarak alınmıştır. Bu kapsamda şekil 3.241’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (loss='hinge', *, penalty='l2', alpha=0.01, l1_ratio=0.15, fit_intercept=True, max_iter=1500, tol=0.001, shuffle=True, verbose=0, epsilon=0.1, n_jobs=None, random_state=42, learning_rate='optimal', eta0=0.0, power_t=0.5, early_stopping=False, validation_fraction=0.1, n_iter_no_change=5, class_weight=None, warm_start=False, average=False). Öğrenme sonucunda %86.903 doğruluk ve harmonik ortalama olan %87.579 F1 skoru elde edilmiştir.



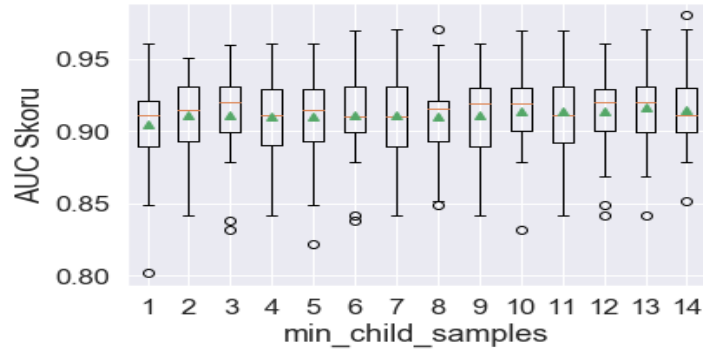
Şekil 3.242. LGBM modelinde n estimators ile AUC skoru



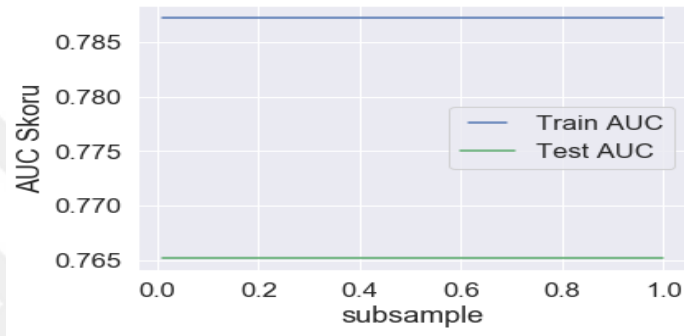
Şekil 3.243. LGBM modelinde öğrenme oranı (learning rate) ile AUC skoru



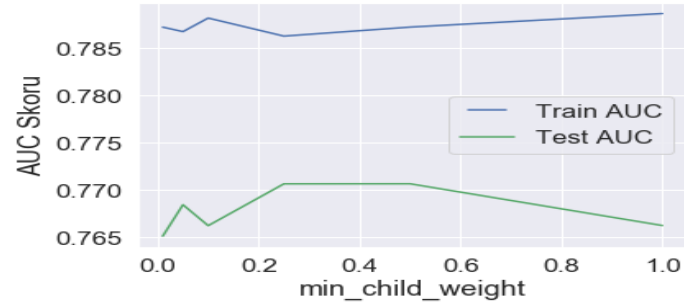
Şekil 3.244. LGBM modelinde max depth ile AUC skoru



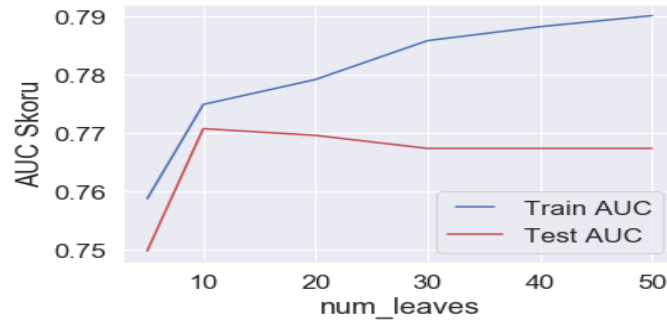
Şekil 3.245. LGBM modelinde min child samples ile AUC skoru



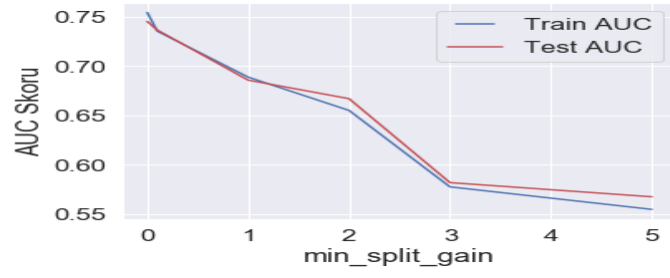
Şekil 3.246. LGBM modelinde subsample ile AUC skoru



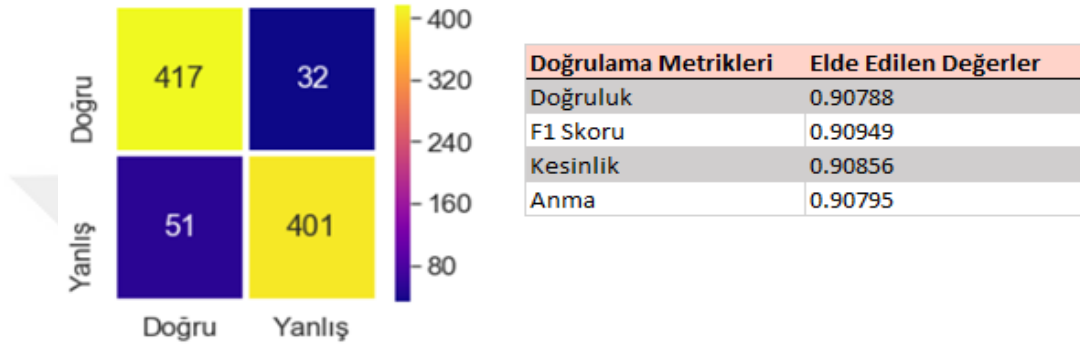
Şekil 3.247. LGBM modelinde min child weight ile AUC skoru



Şekil 3.248. LGBM modelinde num leaves ile AUC skoru

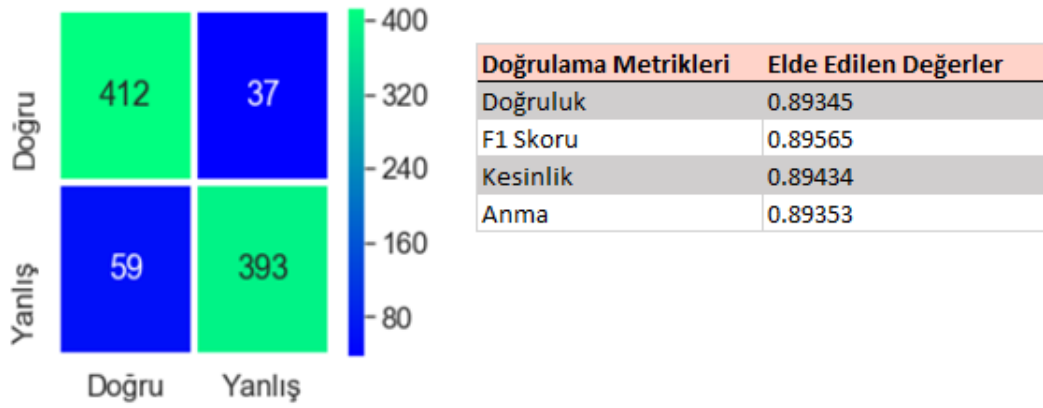


Şekil 3.249. LGBM modelinde min split gain ile AUC skoru



Şekil 3.250. LGBM ile elde edilen değerler

Light gradyan artırma (LGBM) modelinin oluşumuna ilişkin farklı parametrelerde modelin performansını ölçerek, optimum niteliklerin (Hyperparameters – Tuning) belirlenmesi için şekil 3.242’de n estimators 200, şekil 3.243’de öğrenme oranı 0.25, şekil 3.244’de max depth 9, şekil 3.245’de min child samples 2, şekil 3.246’da subsample 1.0 (Stabil olduğundan varsayılan değer alındı.), şekil 3.247’de min child weight 0.25, şekil 3.248’de num leaves 10 ve şekil 3.249’da min split gain 0 olarak alınmıştır. Bu kapsamda şekil 3.250’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (boosting_type='gbdt', num_leaves=10, max_depth=9, learning_rate=0.25, n_estimators=200, subsample_for_bin=200000, objective=None, class_weight=None, min_split_gain=0.0, min_child_weight=0.25, min_child_samples=2, subsample=10, subsample_freq=0, colsample_bytree=0.75, reg_alpha=0.0, reg_lambda=0.0, random_state=42, n_jobs=-1, silent=True, importance_type='split', **kwargs). Öğrenme sonucunda %90.788 doğruluk ve harmonik ortalama olan %90.949 F1 skoru elde edilmiştir.



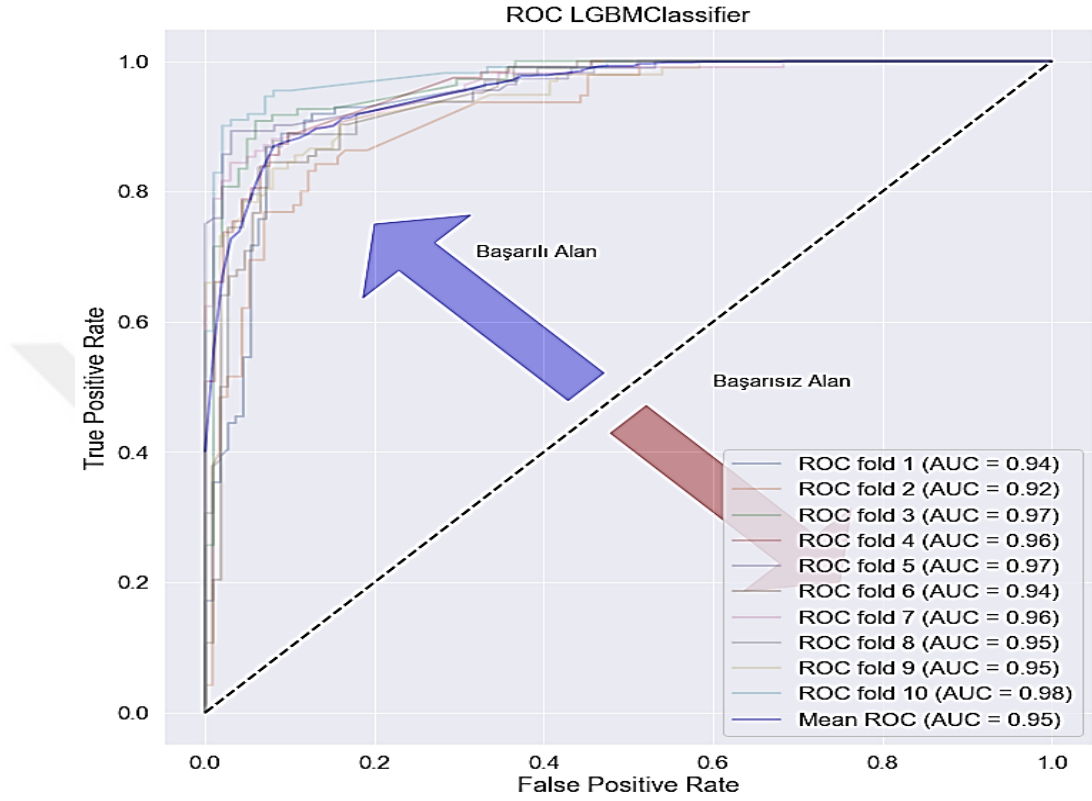
Şekil 3.251. Voting ile elde edilen değerler

Voting modeliyle şekil 3.251’de görüldüğü üzere, test verisi üzerinde elde edilen karışıklık matrisi ve öğrenmeye ilişkin temel parametreler çıkarılmış olup, öğrenmede temel varsayımlar baz alınmıştır (estimators, *, voting='hard', weights=None, n_jobs=None, flatten_transform=True, verbose=False). Öğrenme sonucunda %89.345 doğruluk ve harmonik ortalama olan %89.565 F1 skoru elde edilmiştir.

Çizelge 3.5. Doğruluk değeri bazında modellerin karşılaştırılması

Model	F1 Skoru	Doğruluk
Light Gradient Boosting Model (LGBM)	90.949	90.788
Extra Trees Classifier	90.122	90.122
Voting Classifier Model	89.565	89.345
Decision Tree Model	89.565	89.345
Logistic Regression Model	89.399	89.234
Linear Support Vector Classification	89.109	89.012
Random Forest	88.486	88.013
Bernoulli Naive Bayes	88.44	88.568
Stochastic Gradient Descent (SGD)	87.579	86.903
XGBoost Classifier	87.261	86.681
Perceptron	86.9	86.681
Multinomial Naive Bayes	86.713	87.347
Gradient Boosting Model	86.71	86.459
AdaBoost Classifier	86.283	86.238
Gaussian Naive Bayes	81.978	81.198
Support Vector Machine (SVM)	79.924	76.582
K - Nearest Neighbors Model (KNN)	74.913	68.036
Linear Discriminant Analysis	72.297	72.697
Quadratic Discriminant Analysis	67.637	67.925

Çizelge 3.5’de gösterildiği üzere eğitim setinden test için ayrılan örneklem üzerinde modellerin performansı doğruluk ve F1 skoru bazında mukayeseli olarak bulunmaktadır (Model sıralaması doğruluk değeri bazlıdır.).



Şekil 3.252. Light gradyan artırma modeli ile elde edilen ROC eğrisi

Bu kapsamda eğitimde doğruluk oranı en yüksek olan Light gradyan artırma modelinin (LGBM), K-fold 10 değeri üzerinden doğrulamasının yapılarak, şekil 3.252’de öğrenmenin nasıl gerçekleştiğinin görselleştirilmesi sağlanacaktır.

3.6.3. Makine Öğrenimi Modellerinin Test Veri Seti Üzerine Uygulanması

Test veri seti herhangi bir siber zorbalık etiketi bulunmayan 477779 adet tweetten oluşmakta olup, ön temizleme işlemi sonrası 470000 tweet bulunmaktadır. Söz konusu veri setine, %50’si siber zorbalık olan sosyal medya paylaşımlarından öğrenilenlere göre etiket vurulacaktır.

Çizelge 3.6. Modellerin test verisi üzerindeki sonuçlarının karşılaştırılması

	Model	Etiket 1	Etiket 0	Önişlemede Temizlenen	Toplam	Etiket 1 (%)
0	K - Nearest Neighbors Model (KNN)	10591	461409	7779	479779	2.20747
1	Support Vector Machine (SVM)	12936	459064	7779	479779	2.69624
2	Decision Tree Model	16559	455441	7779	479779	3.45138
3	Light Gradient Boosting Model (LGBM)	17657	454343	7779	479779	3.68024
4	Voting Classifier Model	18955	453045	7779	479779	3.95078
5	Random Forest	20601	451399	7779	479779	4.29385
6	Stochastic Gradient Descent (SGD)	23383	448617	7779	479779	4.8737
7	Extra Trees Classifier	26417	445583	7779	479779	5.50608
8	XGBoost Classifier	27397	444603	7779	479779	5.71034
9	Logistic Regression Model	40488	431512	7779	479779	8.43889
10	Linear Support Vector Classification	41968	430032	7779	479779	8.74736
11	Gradient Boosting Model	44195	427805	7779	479779	9.21153
12	Perceptron	55425	416575	7779	479779	11.5522
13	Bernoulli Naive Bayes	57457	414543	7779	479779	11.9757
14	Gaussian Naive Bayes	57618	414382	7779	479779	12.0093
15	AdaBoost Classifier	59249	412751	7779	479779	12.3492
16	Linear Discriminant Analysis	133745	338255	7779	479779	27.8764
17	Quadratic Discriminant Analysis	135907	336093	7779	479779	28.327
18	Multinomial Naive Bayes	169050	302950	7779	479779	35.235

Çizelge 3.6’da bütün modellerin eğitim setinden öğrendiklerine göre sonuçları gösterilmekte olup, analizde doğruluk oranı en yüksek olan 4 modeli inceleyecek olursak;

- Light gradyan artırma modeliyle (LGBM) %90.788 doğruluk oranıyla 17657 adet içinde @ ve # karakterleri bulunan tweet’in siber zorbalık içerdiğini tespit ettiği ve bu tespitin genel örneklemin %2.207’sini oluşturduğu,
- Ekstra ağaç modeliyle (Extra trees) %90.122 doğruluk oranıyla 26417 adet içinde @ ve # karakterleri bulunan tweet’in siber zorbalık içerdiğini tespit ettiği ve bu tespitin genel örneklemin %5.506’sını oluşturduğu,

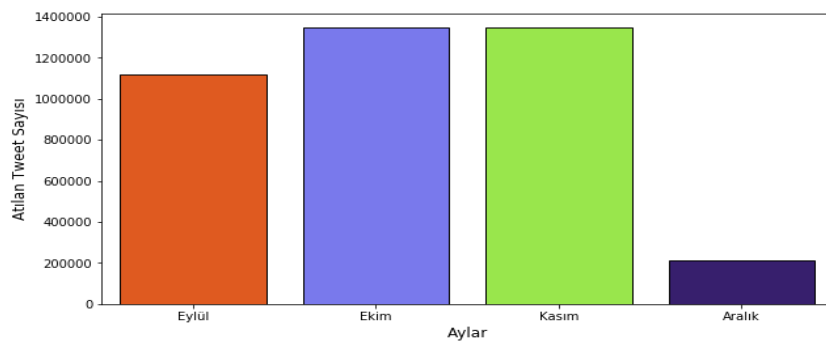
- Voting sınıflandırma modeliyle %89.345 doğruluk oranıyla 18955 adet içinde @ ve # karakterleri bulunan tweet'in siber zorbalık içerdiğini tespit ettiği ve bu tespitin genel örneklemin %3.95'ini oluşturduğu,
- Karar ağacı modeliyle (Decision tree) %89.345 doğruluk oranıyla 16559 adet içinde @ ve # karakterleri bulunan tweet'in siber zorbalık içerdiğini tespit ettiği ve bu tespitin genel örneklemin %3.45'ini oluşturduğu belirlenmiştir.

3.7. Denetimsiz Öğrenme Kapsamında Büyük Veri Üzerinde Keşifsel Veri Analizi ve Açık Kaynak İstihbaratı

Twitter örnekleminde sosyal ağlar üzerinde yapılan paylaşımlardan yola çıkarak gündem hakkında fikir edinmek, gündemi belirleyen kişi ve etiketleri bularak ağ analizi yapmak, paylaşımlardan duygu durumlarını belirleyerek, paylaşımların konu analizinin yapılması çalışmanın amacıdır. Çalışmanın amacı kapsamında elde edilen bulgular alt maddelerde sunulmaktadır.

3.7.1. Keşifsel Veri Analizi

Analiz kapsamında incelemeye konu veri seti ise 4.030.945 adet Tweet'den oluşmaktadır (Twitter API üzerinden İngilizce dil süzgeci içererek, içinde etiketleme imgesi olarak kullanılan “#” simgesi içeren paylaşımlar mevcuttur).



Şekil 3.253. Veri çekilen zaman aralığı ve tweet miktarları

Şekil 3.253’de görüldüğü üzere çalışma kapsamında elde edilen verilerin 04 Eylül – 04 Aralık 2019 tarihleri arasında Twitter API aracılığıyla çekilen verilerdir.

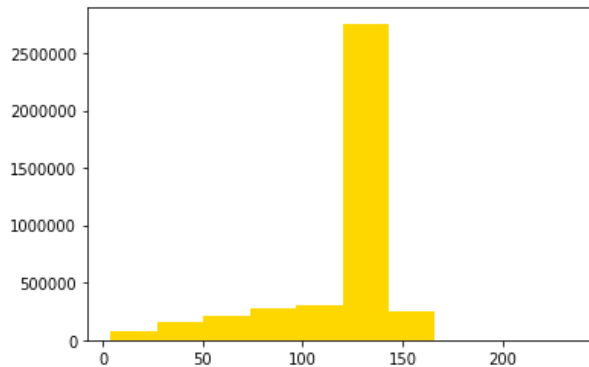
	id	text	created_at	source	retweeted
0	1169037193421213697	Imran Khan's Promise To Sikhs Amid Outrage Ove...	Tue Sep 03 23:59:21 +0000 2019	<a href="https://mobile.twitter.com" rel="nol...	False
1	1169037218746392576	It feels so good to be back at inNBBA TOASTMAS...	Tue Sep 03 23:59:27 +0000 2019	...	False
2	1169037237725667328	RT @SNKPOfficial: Total prize pool over 10 mil...	Tue Sep 03 23:59:31 +0000 2019	<a href="http://twitter.com/download/android" ...	False
3	1169037238379769856	@13_Rosie_96 Went to Sobays my first night her...	Tue Sep 03 23:59:31 +0000 2019	<a href="http://twitter.com/download/iphone" r...	False
4	1169037247712243715	RT @jammytunes: I need more RT on this now...	Tue Sep 03 23:59:34 +0000 2019	<a href="http://twitter.com/download/android" ...	False

```
print("Shape of the dataframe is",data.shape)
print("The number of nulls in each column are \n", data.isna().sum())
```

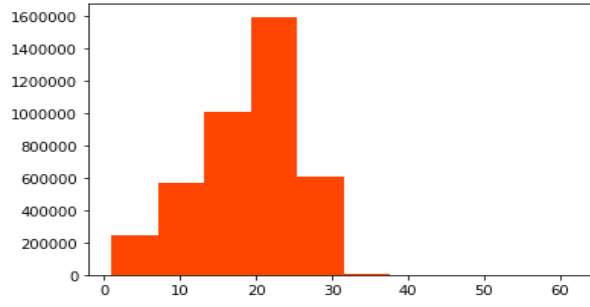
```
Shape of the dataframe is (4030945, 5)
The number of nulls in each column are
id          0
text        0
created_at  0
source      0
retweeted   0
dtype: int64
```

Şekil 3.254. Çekilen verinin örnek görünümü ve boş değer miktarı

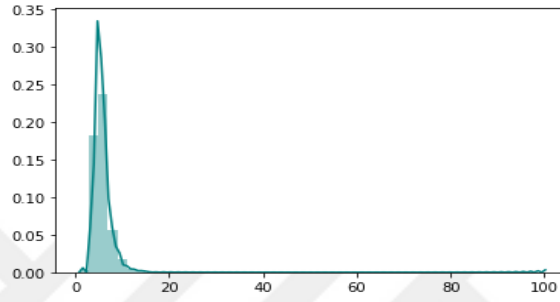
Şekil 3.254’de çekilen verinin pandas formatında görünümü ile veri içerisinde herhangi bir boş satır olmadığı, verinin “id”, “text”, “created_at”, “source” ve “retweeted” sütunlarından oluştuğu görülmektedir.



Şekil 3.255. Tweetler içindeki karakter sayıları

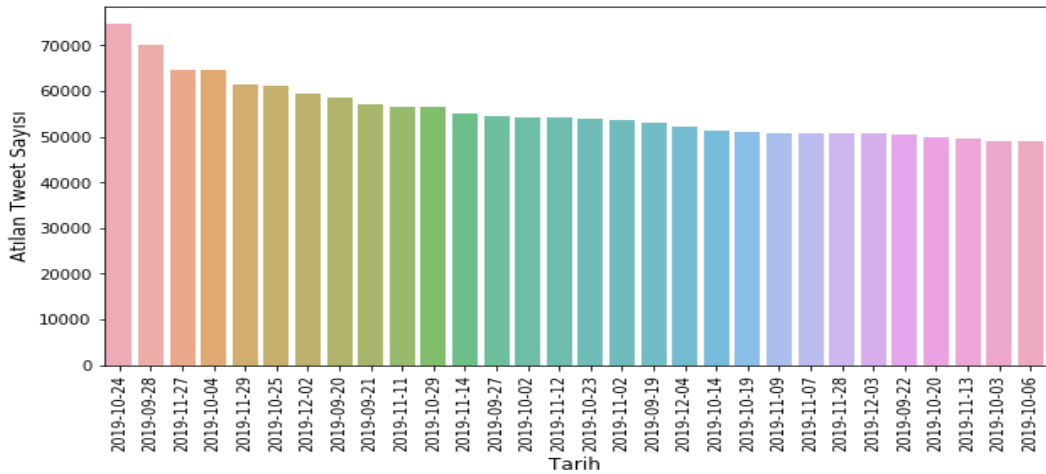


Şekil 3.256. Tweet içindeki kelime sayıları



Şekil 3.257. Tweetler içindeki ortalama kelime uzunlukları

Şekil 3.255’de tweetlerin genelinin 130-150 karakter aralığı bandında olduğu fakat 160 karaktere kadar kullanımın yoğun olduğu, şekil 3.256’da tweet içinde kelime sayılarının 20-25 arasında ağırlık gösterdiği, şekil 3.257’de de tweetler içindeki kelime uzunluğunun genelinin 0-10 arasında görülmektedir.

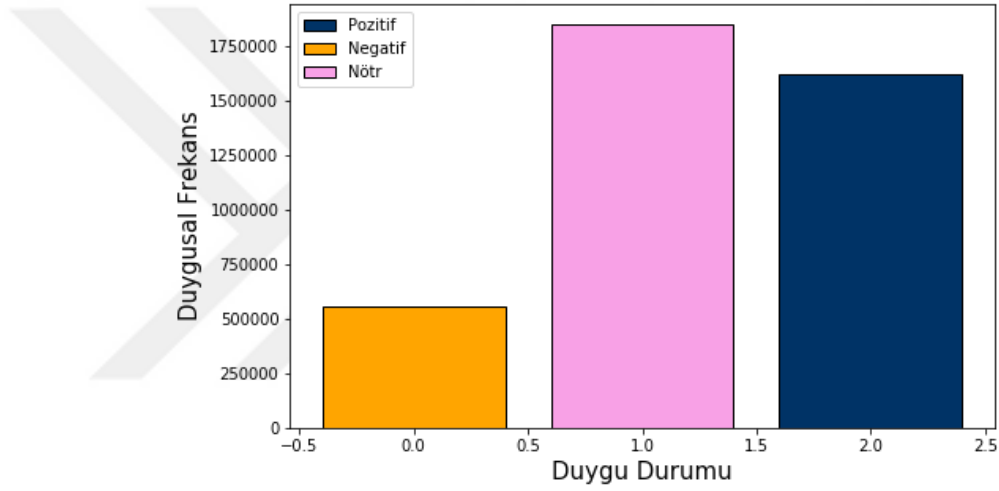


Şekil 3.258. En yoğun verinin olduğu 30 gün ve veri miktarları

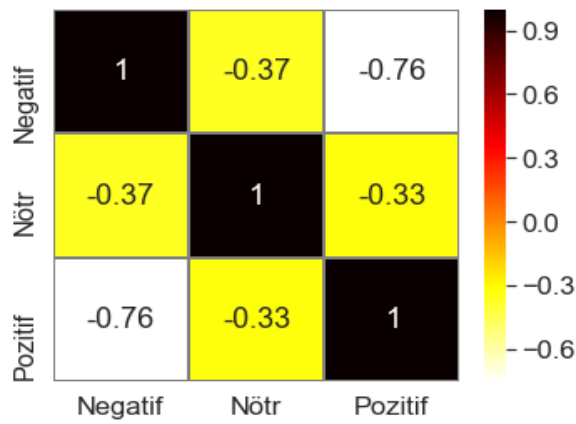
Şekil 3.258’de 91 günlük veri içerisinde en yoğun olan 30 güne ilişkin bilgiler yer almakta olup, günler veri miktarının 50000-75000 arasında değiştiği görülmektedir.

3.7.2. Duygu Analizi

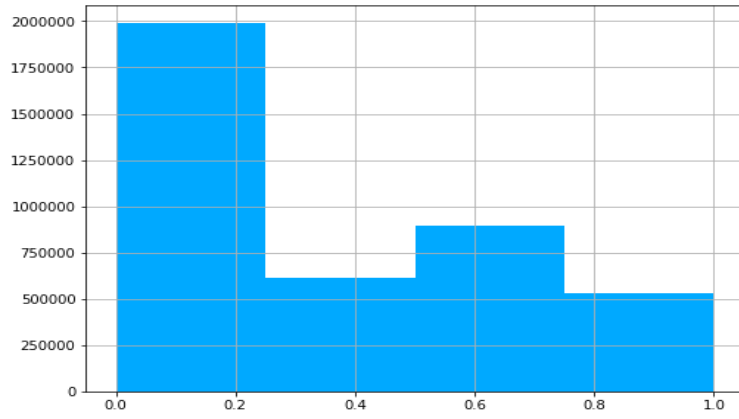
etiketi kapsamında, paylaşım yapılan mesajların duygu durumun belirlenmesine ilişkin tespitler aşağıda belirtilmektedir.



Şekil 3.259. Veri seti üzerinde duygu durum analizi

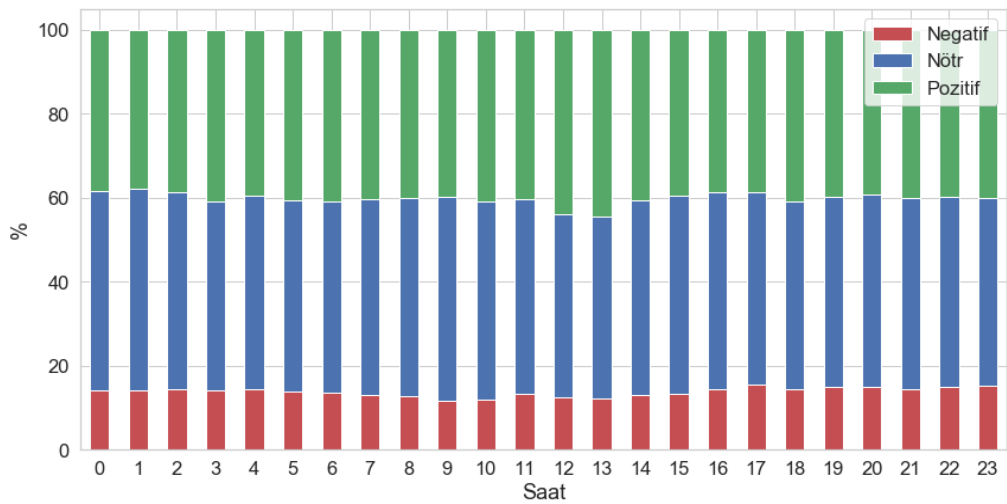


Şekil 3.260. Duygu korelasyon matrisi

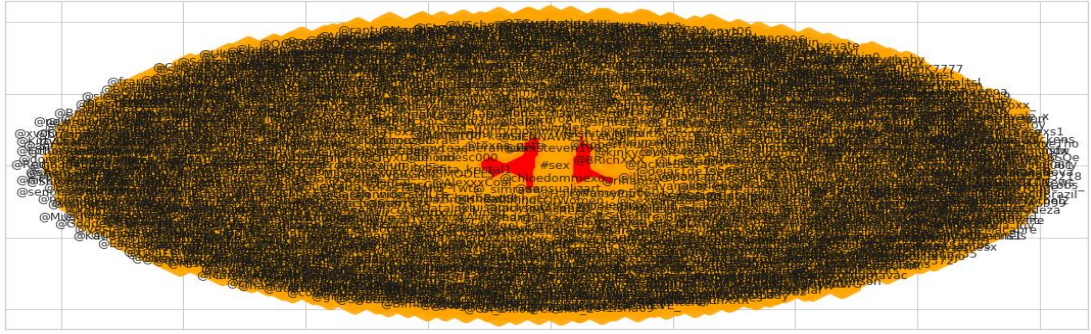


Şekil 3.261. Tespit edilen duygu durumlarının öznelliği

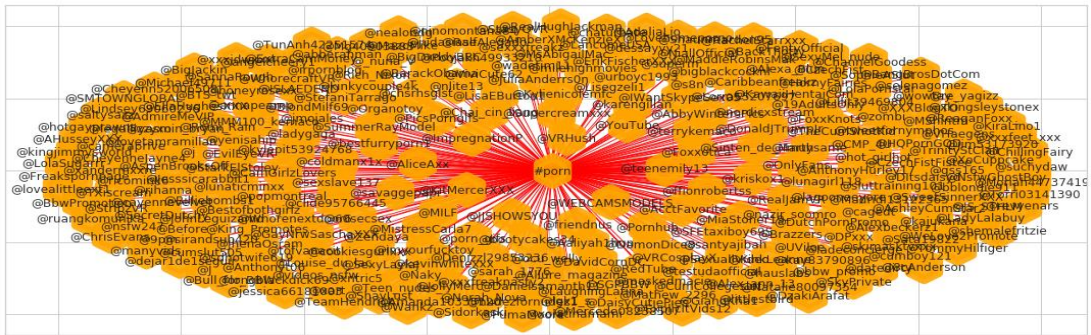
Şekil 3.259’da duygu durumunun 1.850.790 mesajla ağırlıkta nötr mesajlardan oluştuğu, nötr mesajları 1.621.111 mesajla pozitif mesajların izlediği ve 559.044 mesajla da en az duygu durumunun negatif mesajlarda olduğu görülmektedir. Şekil 3.260’da matriste pozitif ve negatif mesajlar arası mesafenin uzak, nötr mesajların da iki duygu durumunun arasında yer aldığı, şekil 3.261’de ise öznelliğin 0-0.25 arasında çoğunlukla yer aldığı görülmekte olup, bu durumda duygu durumunun başarılı şekilde etiketleme yaptığının göstergesidir. Ayrıca şekil 3.260’da görünen negatif ile pozitif mesajlar arasında -0.76 korelasyon olmasını duygu durumunun iyi sınıflandırıldığını belirtir.



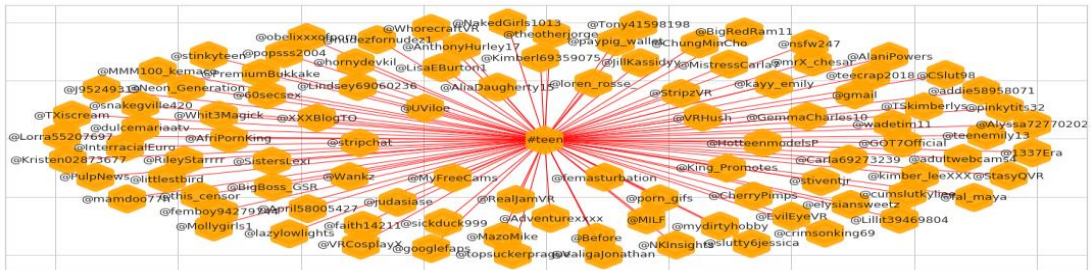
Şekil 3.262. Tespit edilen duygu durumlarının ortalama saat bazlı analizi



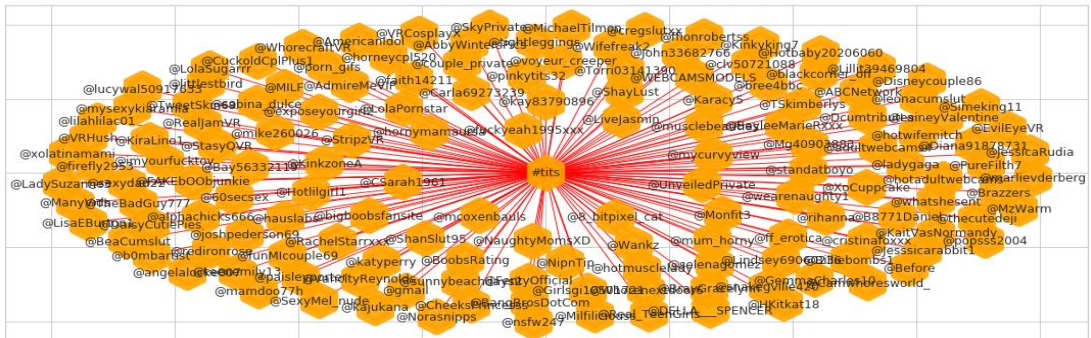
Şekil 3.264. #sex etiketinden hesaplara veri trafiği



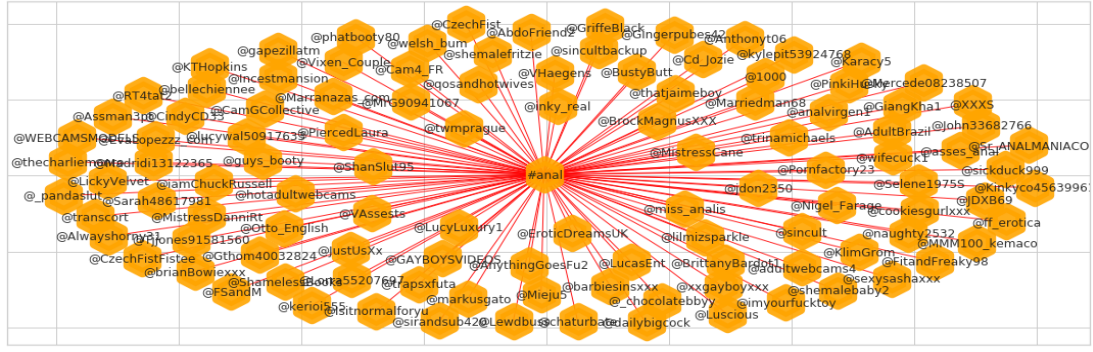
Şekil 3.265. #porn etiketinden hesaplara veri trafiği



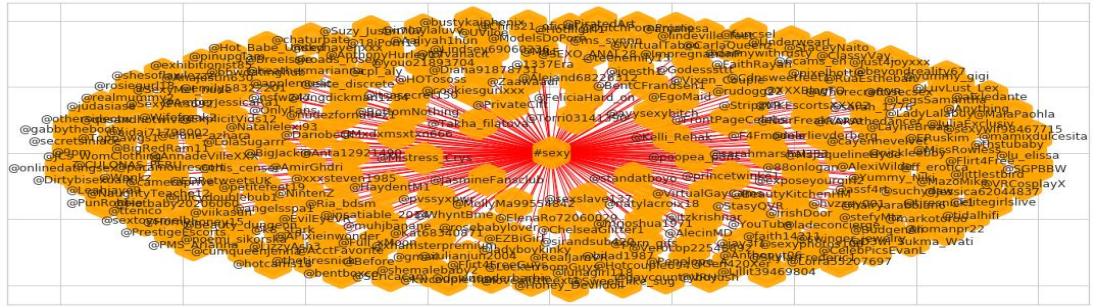
Şekil 3.266. #teen etiketinden hesaplara veri trafiği



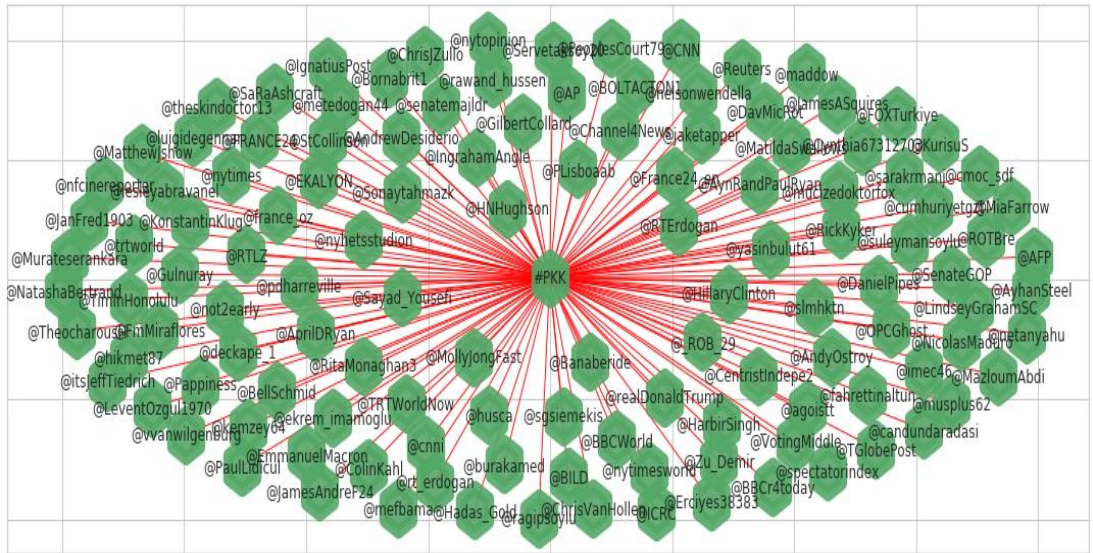
Şekil 3.267. #tits etiketinden hesaplara veri trafiği



Şekil 3.268. #anal etiketinden hesaplara veri trafiği

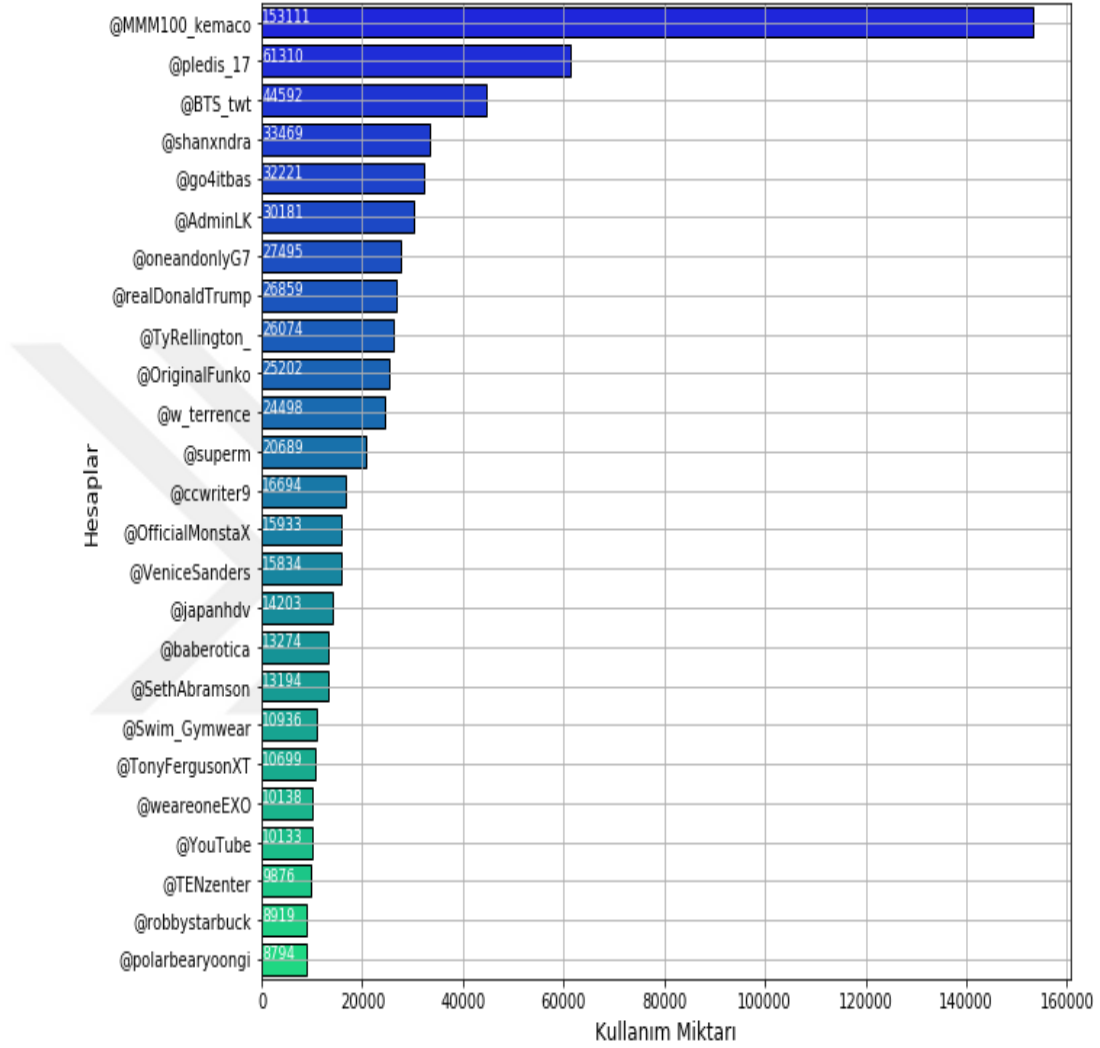


Şekil 3.269. #sexy etiketinden hesaplara veri trafiği



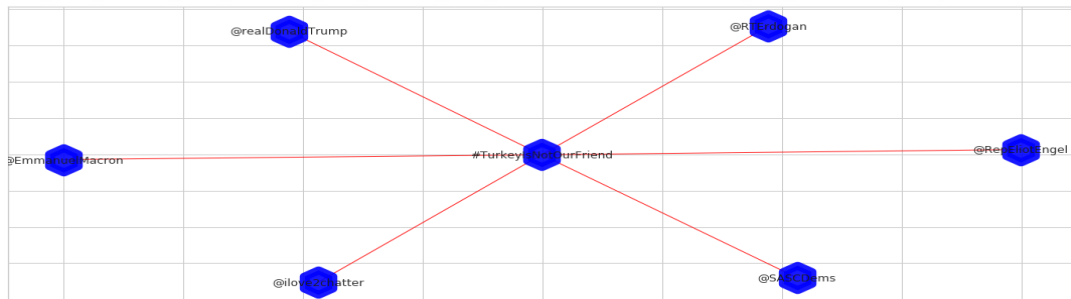
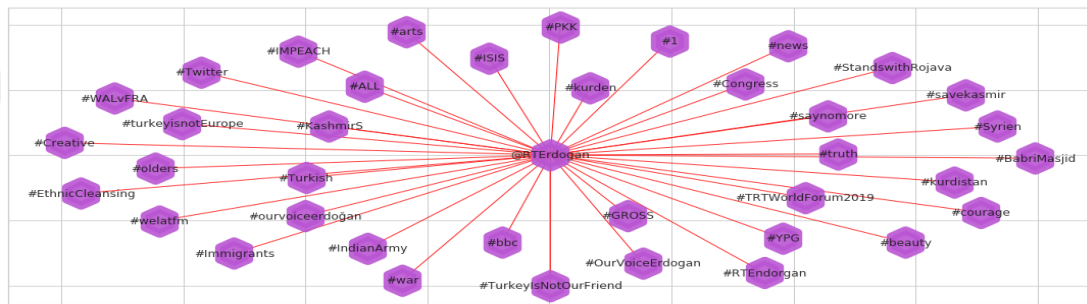
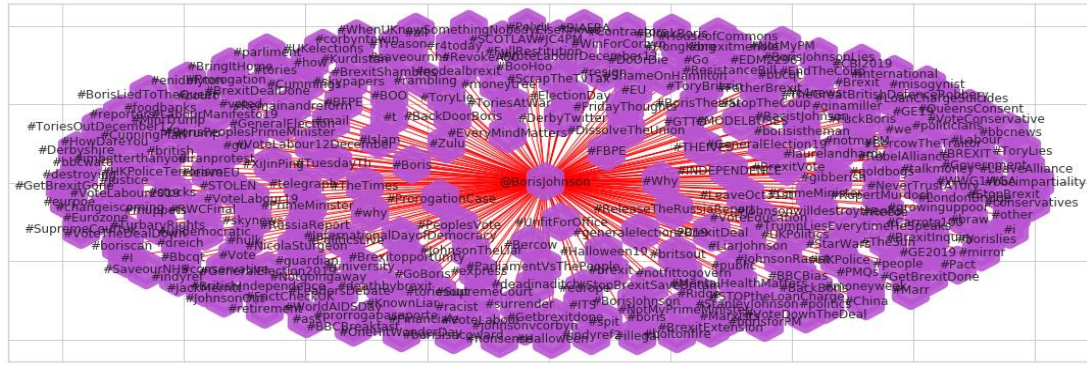
Şekil 3.270. #PKK etiketinden hesaplara veri trafiği

Şekil 3.264, 3.265, 3.266, 3.367, 3.268 ve 3.269’da cinsel kelimeler temelinde ve şekil 3.270’de de “#PKK” etiketinden oluşan konu tanımlamaları hakkında bildirimde bulunan hesap bilgilerine ait ağ trafiği görülmektedir.

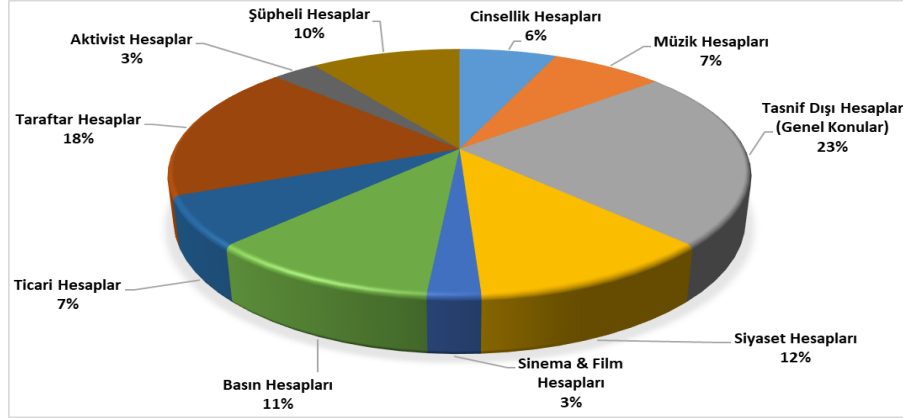


Şekil 3.271. Popüler hesap etiketleri (@)

Sosyal medyada en çok konuşulan hesap bilgileri şekil 3.271’de görülmekte olup, söz konusu hesapların aynı zamanda yönlendirici boyutta olup olmadığını da incelemek mümkündür. Örneğin 72.4 milyon takipçisi olan @youtube’un, 8754 takipçisi olan @baberotica kadar twitter üzerinde etki gücü bulunmamaktadır. Donald Trump da listede 8’inci sırada yer almaktadır.



Şekil 3.272 ve 3.273’de hesap etiketinden konu etiketlerine veri trafiği görülmekte olup, bahse konu şekil 3.273’den yola çıkarak sosyal medyada “#TurkeyisNotOurFriend” konu etiketi üzerinden Türkiye karşıtı propaganda yapıldığı tespitini ve içeriklerini takip etmek mümkündür. Şekil 3.274’den de “#TurkeyisNotOurFriend” etiketinden hangi hesaplara yönelim olduğu görülmektedir.

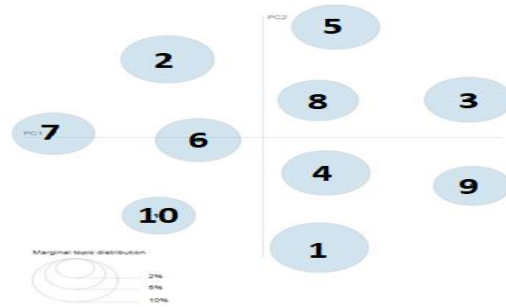


Şekil 3.275. En popüler 200 kullanıcının hesap analizi

Şekil 3.275’de en popüler 200 hesabın, hesap tanıtım bilgileri üzerinden hesap çeşidi analizi görülmektedir. Özellikle %10 olarak şüpheli hesapların ortalama 0-50 aralığında takipçi sayısı olmasına karşın, fakat hesap adlarının kullanıldığı ve bu sebeple de boot tarzı hesap olabileceği ve içlerinde Twitter tarafından askıya alınan hesapların da olduğu belirlenmiştir. En yoğun kullanımın ise taraftar (fan) sayfaları olduğu ve popüler 200 kullanımda %22’lik payı oluşturduğu belirlenmiştir.

3.7.4. Veri Seti Üzerinde Konu Modelleme

İncelemeye konu veri seti içerisinde 4.030.945 tweet içerisinde sadece metin kısımları üzerinde ön işlem sonrası (durağan kelimelerin, URL adreslerinin uzaklaştırılması vb.) 217.205.116 adet kelime elde edilmiştir. Elde edilen değerler vektörel düzeye indirilerek, sayısal formata dönüştürülmüştür.



Şekil 3.276. LDA konu modellemesi

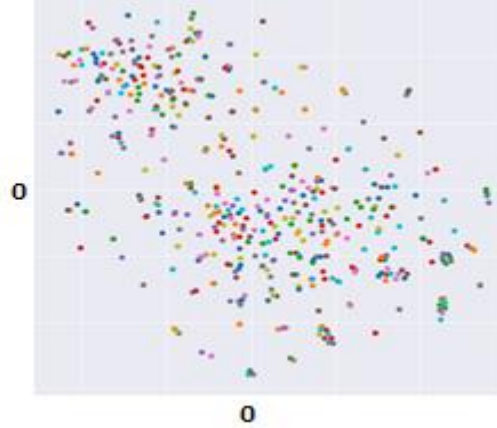
Çizelge 3.7. # konu etiketinde LDA konu modelleri

Konu Nu.	Konular
1	[('yoongi', 100624.51142348068), ('laptop', 93234.74689286109), ('love', 84473.25267023091), ('using', 67953.41110254847), ('want', 67900.29745276715), ('chart', 56741.04028940813), ('good', 52588.144661808336), ('itissuga', 49295.388792052385), ('itissugahyung', 49288.62784392584), ('fans', 48729.86606919177)]
2	[('stole', 93501.31433373239), ('facts', 80860.64308612973), ('know', 75443.16892375097), ('let', 65554.14452657929), ('trend', 64333.86060873585), ('moment', 59268.823872962064), ('impeachment', 50522.45802683704), ('think', 44499.838109864744), ('days', 44398.35460636022), ('artist', 43921.009882012564)]
3	[('vote', 150091.10238613407), ('time', 144254.9015433421), ('nigga', 94252.64938520818), ('make', 70795.59858657444), ('black', 66120.68431595448), ('hashtag', 62651.319450206545), ('history', 47963.76162488016), ('reply', 42523.48870243001), ('giving', 37090.890704377176), ('brain', 36934.44568703457)]
4	[('trump', 56407.55121088141), ('things', 52059.707916975996), ('dems', 44199.42726557309), ('exo', 43605.48650199402), ('tak', 39954.98115648077), ('big', 34143.28843174128), ('news', 33593.424341869955), ('total', 32001.427899711558), ('thank', 31636.091963791605), ('life', 29275.813589777375)]
5	[('amp', 263115.6689352811), ('nov', 55444.683174569625), ('tweet', 54963.30929314515), ('seokjin', 51513.2611770936), ('award', 50303.361528653906), ('line', 47085.66493871672), ('win', 38055.70705632025), ('month', 37303.27352475659), ('young', 36186.441659809825), ('lyric', 34794.52219829098)]
6	[('new', 135095.9908017138), ('trending', 95314.53359307064), ('eligible', 74042.24997928919), ('got', 69458.55447424839), ('year', 50841.36141017999), ('certain', 42432.56210840703), ('mean', 41425.55147128006), ('post', 36684.95241431111), ('drives', 35763.475541600026), ('wednesday', 35065.7763266851)]
7	[('video', 181517.82704824602), ('like', 122066.89747042574), ('use', 80489.1001455386), ('used', 65659.8760947412), ('kst', 60432.732113454214), ('song', 59969.09730350873), ('follow', 57756.40264729795), ('twitter', 48708.5759917676), ('face', 47495.16531726529), ('going', 46094.227476099855)]
8	[('today', 100389.32860883935), ('lmao', 96027.47583485048), ('people', 88393.57774251456), ('america', 82932.24517033529), ('house', 56713.894806751814), ('need', 52107.05398414135), ('devastating', 49903.69349398552), ('effort', 40945.50167089632), ('movie', 40772.98841424858), ('plants', 40162.99701914972)]
9	[('americans', 224294.01950928904), ('help', 59205.25579677226), ('realise', 49753.82676252078), ('world', 43209.6219299245), ('guys', 30541.02295693331), ('years', 30320.164507438418), ('language', 29521.529483057286), ('final', 28439.186830172934), ('mma', 24170.602212742237), ('business', 23696.384093422243)]
10	[('day', 111591.58495339907), ('best', 54237.11675996741), ('tweets', 23208.16036577706), ('heart', 21128.53284218457), ('tag', 20822.407446935944), ('playing', 20670.88293078087), ('seconds', 19486.718548995166), ('november', 19083.640258023242), ('download', 17161.196633286654), ('left', 16758.222856841974)]

Şekil 3.276’da LDA konu modellemesi üzerinden vektörel düzeyde konuların dağılımı, çizelge 3.7’de ise vektörel düzeyde belirlenen numaralarda en fazla kullanılan 10 kelime grubu görülmektedir.

3.7.5. Anahtar Kelime Analizi ve Veriyle İlişkisi

Anahtar kelime analizinin yapılması için metin Word2Vec metoduyla vektörel düzleme geçirilerek, 20 devirde (epoch) eğitilerek 93346 kelime elde edilmiştir. Gösterimlerde t-SNE(t-Distributed Stochastic Neighbor Embedding) kullanılmıştır.



Şekil 3.277. t-SNE modellemesi üzerinden vektörel düzeyde kümeleme

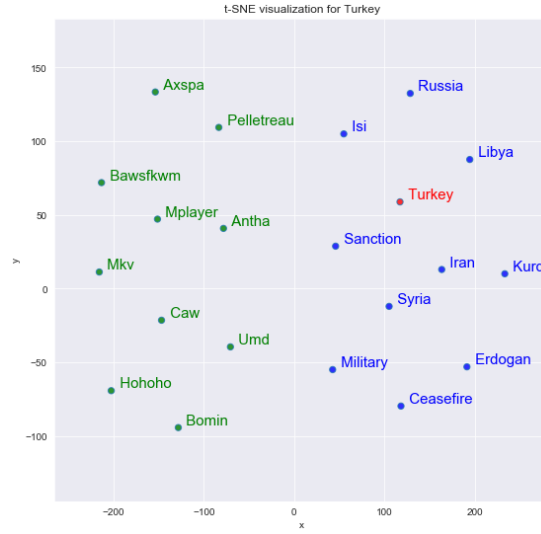
Elde edilen kelime grubunun vektörel düzeyde gösterimi şekil 3.277’de görülmekte olup, 0 üstü değerler ilişkinin pozitif boyutunu, 0 altı değerler ise ilişkinin negatif olan kısmını göstermektedir. Kelimeler arası ilişki analiz için “turkey” ve “pkk” kelimeleri örnek anahtar kelime olarak seçilmiştir.

```
'turkey' in w2v_model.wv.vocab
True

w2v_model.wv.most_similar(positive=['turkey'])
[('syria', 0.6620044708251953),
 ('kurd', 0.655989944934845),
 ('ceasefire', 0.5996520519256592),
 ('erdogan', 0.5983233451843262),
 ('iran', 0.5541303157806396),
 ('sanction', 0.5527450442314148),
 ('russia', 0.549758791923523),
 ('libya', 0.5479387044906616),
 ('isi', 0.5181564092636108),
 ('military', 0.5094724893569946)]

w2v_model.wv.most_similar(negative=['turkey'])
[('mplayer', 0.4319121241569519),
 ('caw', 0.41297364234924316),
 ('antha', 0.39378494024276733),
 ('bawsfkwm', 0.39179572463035583),
 ('pelletreau', 0.36964893341064453),
 ('axspa', 0.36905577778816223),
 ('bomin', 0.36897599697113037),
 ('hohoho', 0.36847788095474243),
 ('mkv', 0.36842238903045654),
 ('umd', 0.36721348762512207)]
```

Şekil 3.278. “turkey” kelimesi üzerinden ilişki analiz



Şekil 3.279. “turkey” kelimesiyle en ilişkili 10 kelime ile en ilgisiz 10 kelimenin t-SNE ile gösterimi

“turkey” anahtar kelimesi üzerinden yapılan sorgu şekil 3.278’de ve t-SNE modellemesinde en ilgili ve en ilgisiz 10 kelimeyle gösterimi ise şekil 3.279’da yer almaktadır.

```
'pkk' in w2v_model.wv.vocab
```

```
True
```

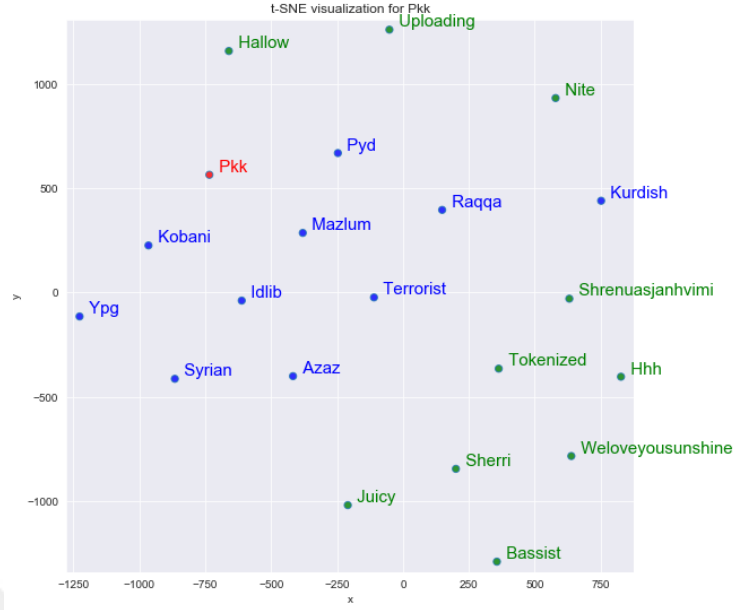
```
w2v_model.wv.most_similar(positive=['pkk'])
```

```
[('ypg', 0.807475209236145),
 ('pyd', 0.7691686749458313),
 ('syrian', 0.6228592991828918),
 ('terrorist', 0.6158443689346313),
 ('idlib', 0.5936588048934937),
 ('mazlum', 0.5903436541557312),
 ('raqqa', 0.5894837379455566),
 ('azaz', 0.5738884210586548),
 ('kurdish', 0.5737199783325195),
 ('kobani', 0.5665556192398071)]
```

```
w2v_model.wv.most_similar(negative=['pkk'])
```

```
[('hhh', 0.37947341799736023),
 ('juicy', 0.3729743957519531),
 ('nite', 0.3695673644542694),
 ('shrenuasjanhvim', 0.3669646084308624),
 ('tokenized', 0.36198174953460693),
 ('hallow', 0.3596702516078949),
 ('weloveyousunshine', 0.35904884338378906),
 ('uploading', 0.3580421209335327),
 ('sherri', 0.35292327404022217),
 ('bassist', 0.3527071177959442)]
```

Şekil 3.280. “pkk” kelimesi üzerinden ilişkisel analiz

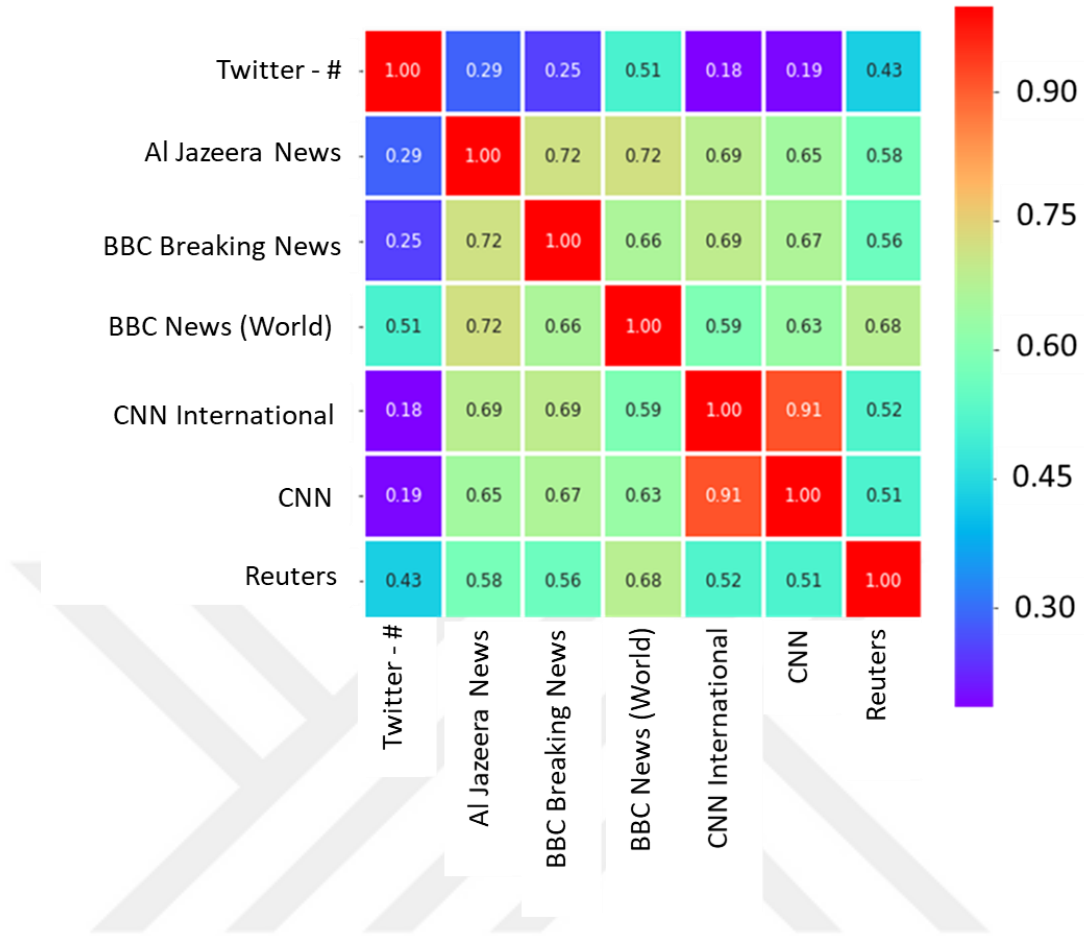


Şekil 3.281. “pkk” kelimesiyle en ilişkili 10 kelime ile en ilgisiz 10 kelimenin t-SNE ile gösterimi

“pkk” anahtar kelimesi üzerinden yapılan sorgu şekil 3.280’de ve t-SNE modellemesinde en ilgili ve en ilgisiz 10 kelimeyle gösterimi ise şekil 3.281’de yer almaktadır. Örneklem olarak yapılan bu modelleme anahtar kelime üzerinden, hangi anahtar kelimelerin ilişkilendirebileceğini ortaya koymaktadır.

3.7.6. Sosyal Medya Gündemi ile Haber Ajansı Paylaşımlarının Kıyaslaması

Suç önleme ve tespit çalışmalarında reaktif ve proaktif tarzda kolluğun reaksiyon göstermesinin temeli suçun olduğunu ya da olabileceğini öğrenmesine bağlıdır. Bu sebeple de her türlü bilginin yer aldığı sosyal ağlardan zaman zaman istihbari çıkarımlar da yapılarak, suç açısından kullanılabilir. Bu sebeple de sosyal ağlarda 04 Eylül – 04 Aralık 2019 tarihleri arasında “@” etiketiyle konuşulan konular ile “Al Jazeera News”, “BBC Breaking News”, “BBC News (World)”, “CNN International”, “CNN” ve “Reuters” üzerinden TF-IDF kullanılarak benzerlik analizi yapılmıştır.



Şekil 3.282. Açık kaynak bilgisi temininde twitter benzerlik matrisi

Şekil 3.282 incelendiğinde;

- Twitter üzerinden konuşulan konulara en benzer haber paylaşımlarının %51 BBC News (World) ve müteakibinde ise %43 ile “Reuters” olduğu,
- En yüksek benzerliğin olduğu %51’lik oran üzerinden konuşulsa dahi sosyal ağlarda konuşulan konuların neredeyse yarısının haber metinlerinde yer almadığı ve bu sebeple de sosyal medyadan haber temini için kullanılabileceği,
- Haber paylaşımları arasında en yüksek benzerliğin %91 ile “CNN International” ve “CNN” arasında olduğu,
- Haber paylaşımları arasında en düşük benzerliğin %51 ile “Reuters” ve “CNN” arasında olduğu anlaşılmaktadır.

4. TARTIŞMA

“Önleyici kolluk faaliyetleri kapsamında sosyal medyada veri madenciliği: suçları önlemek için Twitter Analytic’i kullanma” konu başlığı altında yapılan çalışmaya ilişkin tartışma kısmı aşağıda belirtilen alt başlıklarda açıklanmıştır.

- TDGH kapsamında kolluk kuvvetlerinin paylaşımlarının incelenmesi
- Türkçe sosyal ağ mesajlarında siber zorbalığın tespiti
- Büyük veri üzerinde keşifsel veri analizi ve açık kaynak istihbaratı

4.1. TDGH Kapsamında Kolluk Kuvvetlerinin Paylaşımlarının İncelenmesi

Bilginin yayılımında gönderen, alıcı ve ortam olmak üzere üç temel faktörün etkisi bulunmaktadır (Al-Taie ve Kadry, 2017). Sosyal medya örnekleminde düşünecek olursak, paylaşımı yapan gönderici durumunda, paylaşımdan etkilenmesi beklenen kişi ya da gruplar alıcı konumunda ve sosyal medya platformu ise ortamı teşkil etmektedir. Dünya üzerinde internet kullanan kişilerin günlük ortalama 2 saat 24 dakika süreyle sosyal medyayı kullandıkları bilinmekte olup, bu da sosyal medya ortamlarında gerçekleşen çevrimiçi etkileşimlerin ne kadar yoğun olduğunu ortaya koymaktadır (WeAreSocial, 2020). Etkileşimin yoğun olması, bilginin de hızlı yayılımını tetikleyen temel unsurdur. Özellikle kişi ya da kurum tarafından paylaşılan bir mesajın yeniden (repost, retweet) paylaşılması işlemi, bilginin sosyal medyada yayıldığı ana süreci oluşturmaktadır. Bu sebeple de Twitter örnekleminde bir mesajın beğenilmesi, yorum alması ve tekrar paylaşılması, söz konusu mesajın etki gücünü gösteren temel temsillerdir.

Toplum destekli güvenlik hizmeti, kolluk kuvvetlerinin sürekli vatandaşlarla iletişime geçerek, vatandaşların sorunlarını belirlediği ve belirlenen sorunların çözümüne odaklanan bir kamu güvenliği hizmetidir (Rukus ve ark., 2018; Williams ve ark., 2018). Genel olarak, olay meydana geldikten sonra müdahale etmek yerine,

vatandaşlarla iletişim kurulması ve suçun oluşmasının önlemesi amaçlanmaktadır. Klasik yaklaşımlarda, polis genelde coğrafi konum bazlı oluşturulan bölgelerde hizmet vermektedir (Somerville, 2009). Fakat sosyal medya ve internet kullanımının artmasıyla birlikte, coğrafi sınırların kalktığı bir dönem yaşamaktayız. Özellikle son yıllarda neredeyse tüm jandarma ve polis teşkilatlarının sosyal medya hesapları bulunmaktadır. Bu kapsamda kolluk kuvvetlerinin sosyal medya kullanımlarını incelemek adına, Williams ve ark. (2018) ile Meijer ve Thaens (2013) tarafından Amerika özelinde, Meijer ve Torenvlied (2016) tarafından Hollanda özelinde, Fernandez ve ark. (2017) tarafından ise Birleşik Krallık özelinde, Tang ve ark. (2019) tarafından ise Çin özelinde çalışmalar yapılmıştır. Sosyal medyanın bilginin yayılımındaki gücü bilinmesine rağmen, yapılan paylaşımların çok azının iki yönlü etkileşime açık olduğu belirlenmiştir (Bonsón ve ark., 2012; Hong ve Nadler, 2012; Strauß ve ark., 2015). Yapılan çalışmalardan elde edilen sonuca göre, jandarma ve polis teşkilatlarının sosyal medyanın potansiyelini anladıkları fakat halkla iletişim kurmak için sosyal medyayı yeterince etkin kullanmadıkları belirlenmiştir.

Çalışma kapsamında kullanılan makine çevirisi, doğal dil işlemenin en çok kullanılan uygulamalarından biri olup, doğal dilin bir makine tarafından gerçekleştirilen çevirisidir. Çeviri işlemi, her kelimenin birebir ikamesi değildir. Herhangi iki dil arasında söz dizimsel, gramer ve anlamsal açıdan doğru çeviriler yapmayı amaçlar. Google Translate, Yandex Translate ve Bing Translator gibi uygulamalarla makine çevirileri hala gelişim aşamasında olup, en popüler olanı Google Çeviri'dir (Sarkar, 2019). Aiken (2019) tarafından 50 farklı dilin, Google translate üzerinden çeviri doğruluğu hakkında inceleme yapılmış olup, bu çalışma kapsamında kullanılan Türkçeden – İngilizceye %75, İspanyolcadan – İngilizceye %80, Fransızcadan – İngilizceye %88, İtalyancadan – İngilizceye %90 başarı elde etmiştir. Makine çevirimine ilişkin Roca (2020), Rana ve Atique (2019) ile Shekhawat (2019) tarafından yapılan çalışmalarda da başarılı sonuçlar elde edilmiştir. Bu çalışma kapsamında da TextBlob üzerinden Google translate'den faydalanılmıştır.

Duygu analizi, metinleri analiz etmek ve bu metinlerin içeriğindeki duygu, düşünce ve görüşü tahmin etmek amacıyla kullanılmaktadır. Duygu analizi kavramı

aynı zamanda fikir madenciliği olarak da bilinmekte olup, temel amacı insanların belirli bir varlık hakkındaki duygularını (tepkilerini) analiz etmektir. Sözlüğe dayalı yaklaşım ve makine öğrenimi yaklaşımı şeklinde iki yöntemle duygu analizi yapmak mümkündür. Literatürde Twitter özelinde yapılan bazı duygu analizi çalışmaları incelendiğinde; diyabet hastalığı (Gabarron ve ark., 2019), havayolu şirketi müşteri görüşleri (Prabhakar ve ark., 2019), siyasi seçim tahmini (Huang, 2017), Brexit referandum sonuçları (Grčar ve ark., 2017; Shekhawat, 2019), Covid-19 (Khan ve ark., 2020), Dünya kupası futbol turnuvası (Patel ve Passi, 2020), Endonezya Polisi (Kurniawati ve ark., 2019), Suriyeli mülteciler (Öztürk ve Ayvaz, 2018), Trump'ın Tweet'leri (Sahu ve ark., 2020) ve 5G teknoloji (Seçkin ve Kilimci, 2020) gibi birçok alanda çalışmalar yapıldığı görülmüştür. Yapılan çalışmaların çoğunun sözlük temelli duyarlılık analizi yaklaşımını kullandığı belirlenmiş olup, bu çalışmada da aynı yaklaşım kullanılmıştır.

Çalışmada makine çevirisi ve duygu analizi çalışması sonrası, duyguların ne kadar doğru tasnif edildiğini ortaya koymaya yönelik rastgele orman algoritması kullanılmış olup, çalışmada çeviri de olmasına rağmen %79.565 F1 Skoru elde edilmiştir.

Metin benzerliğinin amacı, metin öğelerinin birbirlerine ne kadar yakın olduğunu analiz etmektir. Genel olarak sözcüksel ve anlamsal olarak iki şekilde metin benzerliği yapmak mümkündür. Literatürde doküman üzerinde yapılan bazı metin benzerliği çalışmaları incelendiğinde; online makale değerlendirilmesi (Lahitani ve ark., 2016), mahkeme dava belgeleri (Novotná, 2020), yasal belgeler (Thenmozhi ve ark., 2017), haberler (Singh ve Singh, 2020), Twitter üzerinden yapılan dizi yorum ve eleştirileri (Prasetyo ve Winarko, 2016), makale benzerliği (Dunn ve ark., 2018), sosyal medya üzerinden acil durum tespiti (Huang ve ark., 2020) gibi çalışmalar mevcuttur. Metin benzerliğinde asıl değerlendirilmesi gereken konunun anlamsal olması gerekirken, her dilin neredeyse sonsuz varyasyonla kullanımı neticesinde bu analizleri yapmak çok kolay değildir. Sözcüksel yöntemlerde de “kaş” ve “kaşık” gibi kelime bazında benzer kelimelerin olması analizi zorlaştırmakta olup, bunun üstesinden gelmek için kelimelerin ağırlıklandırılması gibi daha farklı yöntemlerin

kullanılabileceği değerlendirilmektedir. Yapılan çalışmalarda özellikle doküman benzerliği ile ilgili yapılan çalışmaların birçoğunda kosinüs mesafesi yönteminin, benzerlik hesabında kullanıldığı belirlenmiş olup, bu çalışmada da aynı yöntem kullanılmıştır. Doküman üzerinden yapılan TF-IDF benzerlik de kosinüs temelinde işlev görmektedir.

Konu modelleme bir korpustan ana konuları ve temaları çıkarmak için matematiksel ve istatistiksel modelleme tekniklerinin kullanımını içermektedir. Genellikle olasılıklı istatistiksel modeller olarak da bilinen konu modelleme, metin verilerindeki bağlantılı gizli anlamsal yapıları keşfetmek için gizli anlamsal analiz (LSI), tekil değer ayrışımı (SVD), gizli dirichlet tahsisi (LDA) ve olasılıksal gizli anlamsal analiz (pLSI) gibi çeşitli teknikler kullanır (Lane ve ark., 2019). Söz konusu teknikler arasında en popüler olanı LDA olup, belgelerde denetimsiz olarak kelime dağıtımını yoluyla gizli kelime oluşum kalıplarını tanıyan bir konu modelleme algoritmasıdır (Jacobi ve ark., 2016; Zhou ve Zhang, 2017). İşleyişi bayesian yaklaşımı çerçevesinde olup, konu karışımı Dirichlet dağılımıdır (Jelodar ve ark., 2019). Bu yüzden de doküman, konuların sonlu bir karışımından oluşur ve her bir konunun belge temsili için bir olasılığı vardır. Konu modellerinin pyLDavis ile interaktif şekilde görselleştirmek mümkün olup, t-SNE, PDA ve MDS gibi boyut küçültme teknikleri ile iki boyutlu olarak görselleştirilebilir (Sarkar, 2019). Literatürde konu modelleme üzerinde yapılan bazı metin benzerliği çalışmaları incelendiğinde; Twitter ve Reddit’de yer alan konuların belirlenmesi (Curiskis, 2020), Twitter ile Covid-19 konu modellemesi (Prabhakar ve Prasad, 2020), öneri ve referans konularının tespiti (Basuki ve ark., 2019), Twitter’daki politik cümlelerden konu analizi (Sarddar ve ark., 2020), gazetecilik metinlerinde konuların tespiti (Jacobi ve ark., 2016), Disneyland ile ilgili yorumların konu modellemesi (Luo ve ark., 2020) gibi çalışmalar mevcuttur. Konu modellemesiyle birlikte, çalışma kapsamında veri görselleştirilmesine özellikle önem verilmiştir.

Twitter’ın ilişki merkezli olmaktan ziyade daha çok içerik temelli çalışması diğer sosyal medya uygulamalarından farklı olan yönüdür (Asadi ve Agah, 2017). Aynı zamanda twitter içinde karşılıklı ilişkilere dair göstergelerin uygulama üzerinde

yer almaması, tweet'lerin etkisini ölçmeye zorlaştırmaktadır. Fakat yapılan çalışmalarda kullanıcı takipçi sayısı, tweet'lerin içeriğinin analizi, beğenilmesi, retweet edilmesi ve yorumların yapılması kullanıcının etkili olduğunu gösteren bazı temel öğeler olarak yer almaktadır (Rezaie ve ark., 2020; Dahka ve ark., 2020; Garcia ve ark., 2017). Bu çalışmada yorumlar dışında tüm parametreler metin madenciliği teknikleriyle incelenmiştir. Yorumlar üzerinde inceleme yapılmamasının iki önemli sebebi bulunmaktadır, ilki çalışılan metnin Türkçe, Fransızca, İtalyanca ve İspanyolca içerikler barındırmasının zorluğu ve ikincisi ise kullanıcılar tarafından içeriğinde kısaltılmış ifadelerin çokça kullanıldığı sosyal medya dilinin makine çevirisi yapmadaki zorluğudur.

Jandarma ve polis teşkilatları, genellikle suç ve güvenlik temelinde takipçilerine çeşitli konularda duygu ve düşüncelerini iletmektedir. İletilen mesajların hedef kitleleri tarafından algılanması, iletişimin temel amacıdır. Özellikle İspanya özelinde kolluk birimleri daha fazla tweet atarak halkı bilgilendirmeye çok istekli olmasına karşın, diğer ülkelerdeki jandarma ve polis teşkilatlarının vatandaşları bilgilendirmeye daha az istekli olduğu belirlenmiştir. Fakat İspanya'daki birimlerin günlük ortalama oldukça fazla tweet atmasına karşın, tweetlerin vatandaş üzerinde yeterince etkili olmadığı anlaşılmaktadır. Etki gücünün zayıf olması ise bilginin sosyal medyada yayılım hızını etkileyen en temel unsurlardan biridir. Sık tweet atma ile etkili olma arasındaki parametreler tam olarak bilinmemektedir (Riquelme ve González-Cantergiani, 2016). Fakat çalışmamızda günlük ortalama 2.8 tweet ile en az tweet atan Jandarma'nın ise en yüksek etki değerine sahip olduğu belirlenmiştir.

İletişimde olumsuz uyarıların eşit derecedeki, pozitif uyarılardan daha fazla dikkat çekmesi (Kamp ve ark, 2015) bilinmekte olup, Trevors ve Kendeou (2019) ile Kätsyri ve ark. (2016) tarafından yapılan çalışmalarda da benzer sonuçlar elde edilmiştir. Bu çalışmada da polarite ortalaması düşük yani diğer organizasyonlara göre daha negatif içerikler barındıran paylaşımların, vatandaşlar üzerinde etki gücünün yüksek olduğu belirlenmiştir. Police(Fransa) dışında negatif içerikli mesajların, genelde pozitif korelasyonda beğenilme veya tekrar paylaşım oluşturduğu belirlenmiş olup, Police(Fransa)'nin ise pozitif mesajlardan daha fazla etkileşim aldığı tespit

edilmiştir. Ayrıca sosyal medya üzerinden yapılan paylaşımlarda beğenilme ve tekrar paylaşım korelasyonunun özellikle Policia Nacional(İspanya)’da çok güçlü olduğu belirlenmiştir.

Sosyal medya üzerinden paylaşılan konuların takipçilerin ilgisini çekebilecek şekilde özgün olması etki gücünü etkileyen önemli parametrelerden biridir (Chen ve ark., 2020; Hamzah ve ark., 2020). Yapılan bu çalışmada metin benzerlik oranı en düşük organizasyonun jandarma olduğu belirlenmiş olup, konu modelleri çıkarılmıştır. Konu modellemede sosyal ve kültürel durumların konuları belirlemede etken olduğu tespit edilmiştir.

Çalışmadaki “Güvenlik güçlerinin suça mücadelede toplum desteğini sağlamaya yönelik sosyal ağlarda yaptığı paylaşımların, aynı ülke içinde veya uluslararası boyutta farklılaşmaması beklenmektedir.” hipotezi yapılan analizler sonucu doğrulanmış olup, hipoteze bağlı araştırma soruları ile ilgili olarak;

- Kolluk kuvvetlerinin sosyal ağ kullanım alışkanlıklarını belirlemek adına hangi sıklıkla ve hangi cihazlardan mesajlar paylaştığı belirlenmiş,
- Yapılan paylaşımların duygu durumları ortaya konularak makine öğrenimi temelinde test edilmiş,
- Kolluk tarafından yapılan paylaşımların konu modellemesi ve ilişkili kelime ağları ortaya konmuş,
- Ulusal/uluslararası alanda faaliyet gösteren Jandarma ve Polis tarafından yapılan paylaşımlara halkın verdiği tepki üzerinden mukayeseli korelasyon analizi yapılmış ve şekiller üzerinden anomali tespiti yapmanın mümkün olduğu ortaya konmuş,
- Ulusal/uluslararası alanda faaliyet gösteren Jandarma ve Polis teşkilatlarının sosyal ağdaki etki gücü belirlenmiş,
- Ulusal/uluslararası alanda faaliyet gösteren Jandarma ve Polis teşkilatlarının sosyal ağdaki paylaşımları arasında benzerlik olup olmadığı incelenerek Jandarma (TR)’nın bu analizde ayrıştığı tespit edilmiştir.

Çalışmada farklı ülkelerdeki güvenlik kuvvetlerinin twitter örnekleminde sosyal medya paylaşımları incelenmiş olup, metin madenciliğinin tüm analiz yöntemlerini kullanan benzeri bir çalışma literatürde tespit edilememiştir. Bu konu hem çalışmanın özgünlüğünü hem de çalışmanın literatüre olan katkısını ortaya koymaktadır.

4.2. Türkçe Sosyal Ağ Mesajlarında Siber Zorbalığın Tespiti

Türkçe sosyal medya paylaşımlarından siber zorbalığın tespiti amacı doğrultusunda literatürde bulunan Türkçe doğal dil işleme ve siber zorbalık tespiti yönünde çalışmalar incelenmiştir. Türkçe veri setlerinin yeterince oluşmaması sebebiyle İngilizce başta olmak üzere farklı dillerdeki modellemeler de ortaya konulmuştur.

Raza ve ark. (2020) tarafından İngilizce siber zorbalık mesajlarının tespitine yönelik yapılan çalışmada lojistik regresyon, naive bayes, rastgele orman, adaboost ve voting sınıflandırıcı algoritmaları kullanılmıştır. Kullanılan veri seti etiketlemede dengesiz dağılım göstermektedir. Voting algoritmasında %84,4 doğruluk ile en iyi performans değeri elde edilmiştir.

Van Hee ve ark. (2018) tarafından İngilizce ve Danimarkaca siber zorbalık mesajlarının tespitine yönelik yapılan çalışmada n-gram metodu kullanarak, sadece lineer SVM algoritması kullanılmıştır. Veri seti dengesiz dağılım göstermekte olup, İngilizce için %4.73, Danimarkaca için de %6,97 oranında siber zorbalık mesajları içermektedir. Siber zorbalık mesajları da kendi içerisinde gruplara (Curse, Defamation, Defense, Encouragement, Insult, Sexual, Threat - Küfür, İftira, Savunma, Teşvik, Hakaret, Cinsel, Tehdit) ayrılmış olup, asıl sınıflandırma mesajların siber zorbalık içerip içermediği üzerine yapılmıştır. İngilizce için %64 ile Danimarkaca için de %61 ile en iyi F1 skor elde edilmiştir.

Husain (2020) tarafından Arapça sosyal medya mesajları üzerinde saldırgan dil kullanımı tespit edilmeye çalışılmıştır. Verilerin ön işleminde TF-IDF kullanılmış

olup, emojiler de çalışmada dikkate alınmıştır. Kullanılan veri seti, dengesiz dağılım göstermektedir. Çalışma kapsamında SVM, lojistik regresyon, karar ağacı, torbalama, adaboost ve rastgele orman algoritmaları kullanılmıştır. Torbalama algoritması kullanılarak, %88 F1 Skor ile en iyi performans değeri elde edilmiştir.

Kanan ve ark. (2020) tarafından Arapça siber zorbalık ve siber taciz mesajlarının tespitine yönelik yapılan çalışmada, KNN, SVM, naive bayes, rastgele orman ve karar ağaçları kullanılmıştır. Kullanılan veri seti Facebook ve Twitter verilerinden oluşmakta olup, dengesiz dağılım göstermektedir. Rastgele orman algoritması kullanılarak %94,7 F1 skor ile en iyi performans değeri elde edilmiştir.

Muneer ve Fati (2020) tarafından siber zorbalığın tespitine yönelik yapılan çalışmada lojistik regresyon, LGBM, SGD, rastgele orman, adaboost, naive bayes ve SVM algoritmaları kullanılmıştır. Lojistik regresyon algoritması kullanılarak %92,8 F1 skor ile en iyi performans değeri elde etmiş olup, LGBM algoritması da %92,7 F1 skor elde etmiştir.

Talpur ve O’Sullivan (2020) tarafından siber zorbalığın tespitine yönelik yapılan çalışmada naive bayes, KNN, karar ağacı, rastgele orman ve SVM algoritmaları kullanılmıştır. Risk düzeylerine göre siber zorbalık tweetleri sınıflandırılmış olup, dengesiz sınıf dağılımına karşı SMOTE (Synthetic Minority Oversampling Technique) tekniği kullanılmıştır. Rastgele orman algoritması kullanılarak %92,9 F1 skor ile en iyi performans değeri elde edilmiştir.

Altay ve Alataş (2018) tarafından İngilizce metinlerde siber zorbalığın tespitine yönelik yapılan çalışmada bayesyen lojistik regresyon, rastgele orman algoritması, çok katmanlı algılayıcı, j48 algoritması ve destek vektör makinesi kullanılmıştır. Bayesyen lojistik regresyon algoritması kullanılarak %72,9 F1 skor ile en iyi performans değeri elde edilmiştir.

Ali ve Syed (2020) tarafından İngilizce metinlerde siber zorbalığın tespitine yönelik yapılan çalışmada SVM, naive bayes, rastgele orman ve topluluk öğrenimi yaklaşımları kullanılmıştır. SVM algoritması kullanılarak %79,3 doğruluk ile en iyi performans değeri elde edilmiştir.

Liu ve ark. (2019) tarafından İngilizce metinlerde siber zorbalığın tespitine yönelik yapılan çalışmada naive bayes, karar ağacı, rastgele orman, topluluk, lojistik regresyon ve SVM algoritmaları kullanılmıştır. İçlerinde Facebook, Twitter, Formspring gibi sosyal medya ortamlarının olduğu 9 farklı veri seti kullanılmıştır. Lojistik regresyon algoritması kullanılarak %81,75 doğruluk ile en iyi performans değeri elde edilmiştir.

Zhang ve ark. (2019) tarafından Japonca metinlerde siber zorbalığın tespitine yönelik yapılan çalışmada lineer SVM, lojistik regresyon, karar ağacı rastgele orman, gradyan artıma, lojistik regresyon ve perceptron algoritmaları kullanılmıştır. Veriler Twitter üzerinden alınmış olup, dengeli dağılım göstermektedir. Gradyan artırma algoritması kullanılarak %93,5 F1 skor ile en iyi performans değeri elde edilmiştir.

Yukarıda belirtilen çalışmaların tamamı Türkçe dışındaki dillerde yapılan çalışmalardır. Söz konusu çalışmalarda, siber zorbalığın tespiti için çeşitli sınıflandırma algoritmaları kullanılarak, her dilin kendine özgü ön işleme teknikleri kullanılmıştır. Türkçe dilinin sondan eklenmeli bir dil olması sebebiyle doğal dil işleminin zorluğu, bu alanda kısıtlı çalışma olmasının da en temel gerekçesidir. Bu kapsamda Türkçe dilinde siber zorbalık tespitine yönelik literatürde yer alan sınırlı sayıdaki çalışma incelenmiştir.

Çöltekin (2020) tarafından Türkçe saldırgan dilin kullanımının tespitine yönelik çalışmada, %19'u saldırgan dil olarak etiketlenen sosyal medya mesajları veri seti olarak kullanılmıştır. Saldırgan dilin tespitinde makine öğrenimi kullanılmıştır. Sadece lineer SVM sınıflandırma algoritması kullanılarak, %77.3 F1 skor elde edilmiştir.

Bozyiğit ve ark. (2019) tarafından Türkçe siber zorbalık tespitine yönelik çalışmada, yapay sinir ağları kullanılmıştır. Türkçe doğal dil işleme, mevcut kütüphaneler kullanılmamıştır. “Zararlı Terimler” listesi oluşturularak, terim ile girdi arasında Levenshtein algoritması kullanılarak, yanlış yazımlar düzeltilmeye çalışılmıştır. Metin madenciliğinde TF-IDF kullanılmıştır. Gizli katman sayısının 2 olarak öngörüldüğü çalışmada, %91 F1 skor elde edilmiştir.

Özel ve ark. (2017) tarafından Türkçe siber zorbalığın tespitine yönelik yapılan çalışmada, SVM, karar ağacı, çoklu naive bayes ve KNN algoritmaları kullanılmıştır. Türkçe doğal dil işleme, mevcut kütüphaneler düşük performans göstermeleri sebebiyle kullanılmamıştır. Çoklu naive bayes ile %84 doğruluk elde edilmiştir.

Yukarıda belirtildiği üzere Türkçe siber zorbalık tespiti konusunda yapılan çalışmalar oldukça kısıtlıdır. Yapılan çalışmalarda da sınırlı sayıda makine öğrenimi algoritmaları temelinde çalışma yapılmıştır. Bu sebeple farklı algoritmalar temelinde Türkçe metinlerde siber zorbalığın tespitine yönelik çalışma yapılması gereksinimi oluşmuştur. Siber zorbalığın tespitine yönelik, Bozyiğit (2018) tarafından Twitter üzerinden oluşturulan veri seti kullanılmıştır. Söz konusu veri seti içerisinde 3000 adet twitter mesajı yer almakta olup, 1500’ü siber zorbalık içeren, 1500’ü de siber zorbalık içermeyen şeklinde etiket mevcuttur. Etiketlemenin eşit sayıda yapılması makine öğrenimine dayalı sınıflandırmada daha dengeli bir dağılım yapılması açısından önemlidir.

Çalışmadaki “Makine öğrenimi modelleriyle sosyal ağlardaki Türkçe metinler üzerinde siber zorbalık tespiti yapmak mümkündür.” hipotezi yapılan analizler sonucu doğrulanmış olup, hipoteze bağlı araştırma soruları ile ilgili olarak;

- Siber zorbalık içeren ve içermeyen mesajlar arasında keşifsel analizde anlamlı bir farklılık olmadığı,
- Makine öğrenimi yapılan 19 modelden ilk 14’ünün doğruluk ve F1 skor değerleri %86-%91 arasındadır. Siber zorbalığın Türkçe tweetlerde tespit

edilebilmesine ilişkin yapılan modellemede %90.949 F1 skor ve %90.788 doğrulukla en iyi sonuç LGBM algoritması kullanılarak belirlendiği,

- QDA algoritması ile %67.637 f1 skor ve %67.925 doğruluk değeriyle en kötü sınıflandırma sonucu elde edildiği
- Ağaç temelli ve topluluk öğrenimi (ensemble learning) yapıda algoritmaların performanslarının iyi olduğu,
- LGBM algoritması ile elde edilen %90,949 F1 skor değerinin Türkçe literatürde klasik makine öğrenimi modelinin kullanıldığı bilinen en başarılı sonuç olduğu,
- Elde edilen sonucun Bozyiğit ve ark. (2019) tarafından iki katmanlı yapay sinir ağlarında elde edilen F1 skor ile benzerlik taşıdığı,
- Özellikle son yıllarda geliştirilen XGBoost ve LGBM algoritmalarının Türkçe siber zorbalık metinlerinde bilinen ilk performans değerlendirmesi olduğu, bunun yanı sıra farklı dillerde metinlerde kullanımının çok sık olmadığı ve yapılan çalışmaların sonuçlarının araştırmamızdaki performans ile benzerlik taşıdığı,
- Mevcut sonuçlar, literatürde yer alarak yukarıda açıklanan farklı dillerde yapılan araştırmalarda elde edilen sonuçlarla kıyaslandığında Türkçe dilinde de anlamlı sonuçların elde edilebildiği belirlenmiştir.

4.3. Büyük Veri Üzerinde Keşifsel Veri Analizi ve Açık Kaynak İstihbaratı

Son yıllarda sosyal medya, çeşitli konularda bilgi edinmek ve yaymak için önemli bir iletişim platformu haline gelmiştir. Geleneksel iletişim araçlarına (gazete, televizyon, radyo gibi) göre popülerliği daha yüksektir (Simeone ve Russo, 2017). Sosyal medyayı geleneksel iletişim araçlarından daha popüler yapan en önemli konu, kişi ya da grupların fikirlerini özgürce paylaşabilmeleri ve başkalarıyla çevrimiçi bağlantı kurabilmeleridir. (Ahmed ve ark., 2019).

Sosyal medya platformu üzerinden kullanıcılar, kendilerini takipçilerine tanıtacak bir dizi mesajlar yayınlamaktadır. Söz konusu mesajlar, takipçiler tarafından beğenilmekte ve tekrar paylaşılabilir. Beğeni yapıldığı takdirde, çeşitli

algoritmalar yardımıyla benzer yönelim gösteren kullanıcılara önerilmekte olup, tekrar paylaşıldığı takdirde ise başka kullanıcılar tarafından da görülebilmekte ve sosyal ağ içinde hızla yayılmaktadır. Sosyal ağlar içindeki iletişim kişisel ağlar içinde olabileceği gibi, fiziksel mesafeler ve asimetrik programlar nedeniyle çevrimdışı etkileşimlerin imkânsız olacağı ve daha önceden hiçbir bağı olmayan bireyler arasında da olabilmektedir (Barberá, 2015).

Sosyal ağ içinde hızla yayılan öğeler zaman zaman suç oluşturmaktadır. Kolluk kuvvetlerinin söz konusu paylaşımları tespiti, ihbar/şikâyet edilmesi ya da proaktif önleyici devriyelerle mümkün olmaktadır. Ateş ve Tokay (2018) tarafından yapılan çalışmada, sosyal ağlardaki suçlarla ilgili kolluk makamlarına yeterince ihbarın gelmediği belirlenmiş olup, bu tespitle kolluğun yapacağı proaktif devriyelerin suçu önlemedeki önemi artmaktadır. Bunun yanı sıra proaktif tutum, suç eğilimleri konusunda da kolluk görevlilerine bilgi sağlayıp, suçun önlenmesi adına kullanımı yapılabilmektedir. Ayrıca sosyal medyadaki örgülerden faydalanılarak açık kaynak istihbaratı da yapmak mümkündür. Literatürde sosyal medya üzerinde suç incelemesinde farklı metotlarda pek çok tespit bulunmakta olup, önemli görülenleri aşağıda açıklanmıştır.

Andrews ve ark. (2018) tarafından yapılan çalışmada, sosyal medyada, vatandaşlar tarafından sağlanan yorum ve gözlemlerin varlığı nedeniyle "zayıf sinyalleri" tanımlamak için bir kaynak olarak kullanılabilmesi ve sosyal medyadaki gönderilerin haber kaynakları, gazeteciler ve diğer ilgi gruplarından toplanan verilerle karşılaştırıldığında özellikle daha nadir bilgiler içerdiği belirlenmiştir. Sunulan yaklaşımda anahtar kelime temelli ifadelerin varlığı olarak "zayıf sinyaller" fikri ortaya koymuş ve bu da sosyal medyanın açık kaynaklardan (OSINT) istihbarat geliştirmenin temeli olacağı değerlendirilmiştir.

Vadapalli ve ark. (2018) tarafından Twitter özelinde yapılan çalışmada, siber atakların önlenmesi için açık kaynak istihbaratı önerilmiştir. Twitter'dan anahtar kelimeler üzerinden elde edilen veriler, doğal dil işleme ile analiz edilmiş ve frekans bilgilerinin de yer aldığı ve adına "TwitterOISNT" dedikleri yazılıma kaydedilmiştir.

Rovner (2019) tarafından yapılan çalışmada, 2011 yılı sonrası açık kaynak istihbaratının Dünya üzerindeki birçok siyasi karışıklıkta kullanılmaya başlandığı ve mevcut durumda devletlerin yanı sıra özel şirketlerin de sosyal medyadan istihbarat temin ettiği vurgulanmıştır.

Tundis ve ark. (2020) tarafından yapılan çalışmada, internette bulunan tehditlerle ilgili, istihbarat kaynaklarından zamanında ve doğru bilgilerin alınmasının önemine değinilerek, otomasyona dayalı bir model tasarlanmıştır. Twitter'daki kaynakların alaka düzeyini tahmin etmek için regresyon modeli eğitilmiş, sosyal medya içeriklerine alaka düzeyi temelinde puan verilerek, istihbarat doğrulaması için bekleme süresini azaltmayı ve böylece erken tehdit tespitinin zaman avantajını artırdığı belirlenmiştir.

Juhlin ve Richardson (2016) tarafından yapılan çalışmada, sosyal medyanın istihbarat servislerinin gözardı edemeyeceği ölçüde bilgi sağladığına değinilerek, Danimarka Kraliyet Savunma Koleji tarafından yürütülen “Project Avatar” sisteminin modellenmesi tanıtılmıştır. Bu proje kapsamında örneğin Twitter’da elde edilen bilgi, Google’da doğrulanmakta ve özellikle İŞİD temelinde zaman zaman yanlış bilgilerin yayıldığı belirlendiği ifade edilmiştir.

Petrova ve ark. (2018) tarafından yapılan çalışmada, sosyal medyadaki büyük veri üzerinde, veri madenciliği teknikleri kullanılarak analizi amaçlamıştır. Keşifsel veri analizi türü bir çalışma olup, hangi ülkelerden ne kadar kullanım olduğu ve siyasi liderlerin sosyal medya kullanımı ve aldığı tepkiler analiz edilmiştir.

Delavallade ve ark. (2017) tarafından yapılan çalışmada, mevcut veya devam eden suç faaliyetini sosyal medyadaki kullanım frekanslarından tespiti yönelik bir çalışma sunulmuş, genel bir istihbarat odaklı süreç kullanılmıştır. Çalışma organize suçla yakından ilgili olan demiryolu sinyali kablolarının çalınması ve sosyal medyada cihatçı propagandanın itici güçlerini anlamlandırma temelinde yapılmıştır, elde edilen sonuçlar sosyal medyanın suç göstergeleriyle ilişkili olduğunu göstermiştir.

Müller ve Schwarz (2019) tarafından Facebook özelinde yapılan çalışmada, sosyal medya üzerinden nefret suçlarının ortaya konması amaçlanmış olup, özellikle sosyal medyanın aşırıcı düşünenlerin görüşlerini yayması sebebiyle, şiddet suçlarının hızlı yayılmasını sağladığı belirlenmiştir.

Curiel ve ark. (2020) tarafından İspanyolca konuşulan ülkelerde Twitter özelinde yapılan çalışmada suçla ilgili tweetlerin sıklığı daha sonra anketlerde kaydedilen cinayet sayısı, cinayet oranı veya suç korkusu düzeyiyle karşılaştırılmıştır. Elde edilen sonuçlara göre, gazeteler ve diğer kitle iletişim araçları gibi, sosyal medyanın da şiddet veya cinsel suçlara yönelik güçlü bir önyargıdan muzdarip olduğunu ve insanların sosyal medyada suç korkusu yaşadıkları belirlenmiştir.

Suçla mücadele kapsamında büyük veri yığınları arasından enformasyona dayalı bilginin ortaya konması hem operasyonel hem de stratejik seviyede analiz etmeyi mümkün kılmaktadır. Bu kapsamda söz konusu bu çalışma, suçla ilgili nelerin ortaya konulabileceğini modellemek amacıyla hazırlanmıştır. Sosyal medyadaki gündem konular üzerinden # etiketi kullanılmış olup, literatürdeki çalışmalardan en büyük farkı budur.

Yapılan çalışma sonucunda, sosyal medyada cinsellik ifadelerinin oldukça sık kullanıldığı belirlenerek, muhteviyatında suç olabileceği değerlendirilen #teen gibi konu etiketlerinin veri hareketliliğinin takip edilmesinin suç mağduriyetinin önüne geçeceği değerlendirilmektedir.

Sosyal medyanın kimler tarafından yönlendirildiğini irdelemek adına, mesajlar içinde en fazla kullanılan kişi etiketleri toplanarak, bunların ilk 200'sinin tasnifi çıkarılmıştır. Elde edilen tasnif sonucunda hemen hemen her konuda kullanım mevcut olup, yaklaşık %10'u şüpheli hesap olarak tasnif edilmiştir. Üç aylık süre zarfı içinde açılıp kapatılmış veya takipçi sayısı oldukça düşük olmasına rağmen anormal veri etiketi olduğu gözlemlenmiştir. Gündemi belirleyen sayfaların yaklaşık %6'sının ise cinsellikle ilgili sayfalar olduğu, bu durumun sık konuşulan konularda cinselliğin olmasını doğruladığı değerlendirilmektedir.

Sosyal ağın genel duygu değerlendirilmesinde %45.9'luk mesajların nötr, %40.2'lik mesajların pozitif ve %13.9'luk mesajların ise negatif olarak gruplandırılmıştır. Spesifik konu olmadığı için nötr mesajların fazla olması, negatif mesajların ise az olması normaldir. Barnaghi ve ark. (2016) tarafından yapılan çalışmada da spesifik olaylarla ve polarite arasında ilişki olduğu gözlemlenmiş, bu sebeple de yaklaşık üç aylık süreyi kapsayan mesajlarda net bir ilişki olması beklenmemektedir.

Çalışma kapsamında büyük veri oluşturulan sisteme öğretilmiş ve anahtar kelimeler üzerinden ilişkili kelimelerin düzeyleriyle birlikte ortaya konması sağlanmıştır. Bulgular kısmında belirtilen örnek sorgulamalarda oldukça başarılı sonuçlar ortaya konulduğu değerlendirilmektedir. Unnisa ve ark. (2016) ve Dragoni ve ark. (2019) tarafından yapılan çalışmalarda da benzer şekilde başarılı modellemeler oluşturulmuştur.

Suç verilerinin konu modellemesi kapsamında; Vomfell ve ark., (2018), Aghababaei ve Makrehchi (2018), Hossny ve ark., (2018) ve Tundis ve ark. (2019) tarafından yapılan çalışmalarda başarılı çıktılar elde edilmiştir. Fakat bu çalışma kapsamında # konu etiketi üzerinden genel bir konu modellemesi yapılarak, konu modellerinin pyLDavis ile interaktif şekilde görselleştirme yapılmış ve görsel analiz için t-SNE boyut küçültme tekniği kullanılmıştır.

Andrews ve ark. (2018) tarafından sosyal medyadaki gönderilerin haber kaynakları ile karşılaştırıldığında daha nadir olduğu tespitine istinaden “Al Jazeera News”, “BBC Breaking News”, “BBC News (World)”, “CNN International”, “CNN” ve “Reuters” haber ajanslarının paylaşımları ile sosyal medyada # etiketiyle konuşulan diğer konular mukayese edilmiştir. Elde edilen sonuçlarda en fazla benzerliğin %51 ile BBC News (World)’de olduğu, diğer sonuçların oldukça düşük olduğu gözlemlenmiştir. Bu durum da istihbari açıdan önemi tartışmalı olsa da açık kaynak istihbaratı yapılarak nadir bilgilere erişebileceğinin göstergesi konumundadır.

Çalışmadaki “Veri madenciliği teknikleriyle sosyal ağlar üzerinde keşifsel veri analizi yapılarak, açık kaynak istihbaratı elde edilebilir.” hipotezi yapılan analizler sonucu doğrulanmış olup, hipoteze bağlı araştırma soruları ile ilgili olarak;

- Herhangi bir konuya ilişkin atılan mesajların duygu durumu tespit edilebilir ve bu durum zaman zaman anomalilerin tespitini sağlayabileceği,
- Olumsuz mesajların (terör, hakaret, zorbalık vb. içeren) UTF sisteminde kullandığımız saat ile ilişkisi olup olmadığı yönünde net bir şekilde analiz yapılamamış olup, yapılan saat bazlı incelemede az da olsa 17:00-24:00 saatleri arası negatif mesajların fazla olduğu belirlenmiş,
- Herhangi bir konuda en etkin # ve @ etiketlerini görmek mümkün olup, söz konusu veriler arası ağ trafikleri ortaya koyulmuş,
- Denetimsiz öğrenme kapsamında konu modellemesi yapılabileceği, özellikle gizli örüntülerin ortaya konması açısından önem taşımakta olduğu belirlenmiş,
- Anahtar kelime analizi araması yapılarak, büyük veriden anahtar kelimeyle bağlantılı ilişkili veriler öğrenilebilir, bu konuda bulgular kısmında “Turkey” ve “PKK” örnekleminde denetimsiz öğrenme sonucu oldukça başarılı sonuçlar elde edilmiş,
- Haber ajanslarından elde edilen bilgi, insanların sosyal medyada yazdıklarının ne kadarını karşıladığı belirlenmiş olup, yapılan tespitte sosyal medyanın istihbarat için yadsınamaz öneme haiz olduğu belirlenmiştir.

5. SONUÇ VE ÖNERİLER

“Önleyici kolluk faaliyetleri kapsamında sosyal medyada veri madenciliği: suçları önlemek için Twitter Analytic’i kullanma” konu başlığı altında suçla mücadele odak noktası alınarak, tezde üç farklı konu işlenmiştir.

Elde edilen sonuçların ilki kapsamında; günümüzde sosyal medyanın yaygın kullanımı ile bireyler ve kurumlar; fikirlerini, düşünceleri ve hatta tutumlarını özgürce ifade edebilmektedir. Bu anlamda Jandarma ve polis teşkilatları da sosyal medya üzerinden duygu ve düşüncelerini ifade etmekte olup, zaman zaman yaptıkları paylaşımların vatandaşlardan yeterince ilgi görmediği belirlenmiştir. Sosyal medyada paylaşımın ilgi görmesi yani diğer insanlar üzerinde etki bırakması, bilginin yayılımı için oldukça önemlidir. Bu kapsamda çalışmada veri madenciliğinin birçok analiz yöntemi (Makine çevirisi, cümle polarite analizi, makine öğrenmesi, korelasyon, metin benzerliği, konu modelleme) kullanılmış ve elde edilen sonuçlar ulusal/uluslararası boyutta mukayese edilmiştir. Yapılan çalışmanın suçla mücadele alanında çalışan jandarma ve polis teşkilatları özelinde olması sebebiyle, metin madenciliği alanında yapılan pek çok çalışmadan ayrıştığı değerlendirilmektedir. Çalışma kapsamında Türkiye (Jandarma - Polis), İtalya (Carabinieri - Polizia), Fransa (Gendarmerie - Police) ve İspanya’da (Guardia Civil - Policia) faaliyet gösteren jandarma ve polis teşkilatlarının twitter örnekleminde karşılaştırmalı olarak, sosyal medyayı kullanım durumları ve takipçileri üzerinde ne ölçüde etkili olabildikleri belirlenmeye çalışmıştır. Yapılan analiz sonucunda tweetlerin pozitif polarite değeri en düşük olan, twitter’ı en az kullanan ve diğer tüm kolluk birimlerinden en farklı içerik paylaşımları yapan Jandarma’nın (Türkiye) çok etkili bir şekilde Twitter’ı kullandığı belirlenmiştir. Ayrıca analizler kapsamında, Türkçe, İtalyanca, Fransızca ve İspanyolca tweetler makine çevirisiyle İngilizceye çevrilmiş olup, sentimental analiz sonrası oluşan duygu durumları makine öğrenimiyle test edilmiştir. Makine çevirisi olmasına rağmen, elde edilen sonuçların rastgele orman sınıflandırma algoritmasıyla ortalama %79.565 F1 skor elde edilmesi dikkate değerdir.

Sonuçların ikincisi kapsamında; sosyal medya, kullanıcılarına çeşitli konular hakkında duygu ve düşüncelerini ifade etme şansı vermektedir. Fakat bazı kullanıcılar siber zorbalık gibi eylemlerle sosyal medyayı kötü amaçlarla kullanmakta ve diğer kullanıcılara karşı hakaret, tehdit ve nefret gibi söylemlerde bulunmaktadır. Bu kapsamda çalışmada sosyal medyanın kötü amaçlarla kullanımın tespiti için makine öğrenimi modelleri kullanılmıştır. Siber zorbalık tespit çalışmasının Türkçe dili özelinde yapılması ve son yıllarda geliştirilen birçok güncel makine öğrenim algoritmalarını da kullanması sebebiyle uluslararası alanda yapılan pek çok siber zorbalık tespit çalışmasından ayrıştığı değerlendirilmektedir. Çalışma kapsamında Twitter örneğinde Türkçe sosyal medya mesajları üzerinden 19 farklı makine öğrenimi algoritması kullanılarak siber zorbalık tespit performansları belirlenmeye çalışılmıştır. %90.949 F1 skor ve %90.788 doğrulukla en iyi sonuç LGBM algoritması kullanılarak elde edilmiştir. Söz konusu sonuç, uluslararası alanda pek araştırma yapılmayan Türkçe dilinde de makine öğrenimine dayalı siber zorbalığın başarılı şekilde tespit edilebildiğini göstermektedir. Ayrıca yarı denetimli öğrenme kapsamında, siber zorbalık tespit sistemlerini kullanmanın yaklaşık %98'e yakın (LGBM) işgücünü azaltacağı belirlenmiştir.

Elde edilen sonuçların üçüncüsü kapsamında ise; sosyal medyada yapılan paylaşımların denetimsiz öğrenme ve açık kaynak istihbaratı açısından ele alınması amaçlanmıştır. Sosyal medya kullanımının artmasına paralel olarak, farklı konularda dinamik biçimde sürekli bir paylaşım mevcut olup, söz konusu veri yığınları arasından anlamlı bilginin tespit edilmesi önemlidir. Bu kapsamda çalışmada veri madenciliğinin birçok analiz yöntemi kullanılmış ve elde edilen sonuçlar ulusal/uluslararası boyutta mukayese edilmiştir. Yapılan çalışmanın suç özelinde olması sebebiyle, metin madenciliği alanında yapılan pek çok çalışmadan ayrıştığı değerlendirilmektedir. Çalışma kapsamında sosyal medya veri trafikleri incelenmiş ve anomali analizi kapsamında tespitler yapılmış ve bazı örnek veri trafikleri paylaşılmıştır. Sosyal medyada oldukça etkin olan bazı hesapların, yaklaşık üç ay gibi kısa sürede kapanması, paylaşımlarının silinmesi ve takipçisi olmaması gibi normal dışı şüpheli hesaplar belirlenmiştir. Elde edilen veriler üzerinden kelime analizi yapılarak, denetimsiz öğrenme üzerinde anlamlı tespitler yapılmıştır. Çalışmanın sonunda ise

açık kaynak istihbaratı kapsamında genel konuşulan konular ile haber ajanslarının paylaşımları mukayese edilmiş ve elde edilen sonuçlara göre sosyal medyanın başarılı doğal dil işleme araçları ile açık kaynak istihbaratı kapsamında oldukça faydalı sonuçlar verebileceği değerlendirilmiştir.

Yapılan bu çalışmaların geliştirilerek;

- Türkçe metin dil işlemeye yönelik başarılı kütüphanelerin oluşturulması,
- Siber zorbalığın tespitine yönelik sosyal medya akış verilerinden anlık modellemenin yapılması,
- Twitter dışında Facebook ve Instagram gibi popüler diğer hesaplarda da incelemeler yapılması,
- Arayüz geliştirilerek, anlık sorgulamaya izin veren bir sistem modeli geliştirilmesi önerilmektedir.

ÖZET

Önleyici Kolluk Faaliyetleri Kapsamında Sosyal Medyada Veri Madenciliği: Suçları Önlemek İçin Twitter Analytic'i Kullanma

Sosyal ağların yaygın kullanımı ile birlikte, geçmişte fiziksel mekânlarda olan iletişim ortamlarının giderek sanal hale gelmeye başladığı bir dönemdeyiz. Sosyal ağ kullanımıyla yaşanan değişimler sadece iletişim alanında olmayıp, neredeyse hayatın her alanında mevcuttur. Söz konusu değişimlerin bazıları oldukça iyi, bazıları oldukça kötüdür. Suç kavramı, sosyal medya ile değişimden çoğunlukla olumsuz etkilenen alanlardan biri olup, yapılan çalışmalar sosyal medyadaki suçların giderek arttığını göstermektedir.

Tez çalışması ile sosyal ağların, suç ve suçla mücadelede kullanımı bütüncül yaklaşımla ele alınmıştır. Bütüncül yaklaşım içerisinde, sosyal medyanın suç önlemeye yönelik rolü ve suç oluşumu sonrası tespiti için çalışmalar bulunmaktadır. Çalışmadan elde edilen verilerin, farklı analiz yöntemleriyle birlikte ortaya konması amaçlanmıştır.

Tez çalışmasında kullanılan veri madenciliği yöntemleriyle birlikte, elde edilen bilgilerin analiz edilmesi bakımından tanımlayıcı/keşfedici sınıftan bir yöntem kullanılmıştır. Çalışmada kullanılan veriler, Twitter üzerinden alınmıştır. Sınırlı zaman periyodundaki veriler üzerinden analiz yapılması, çalışmanın temel kısıtıdır. Verilerin elde edilmesi ve analizi aşamasında Python 3 programlama dili kullanılmıştır. Tez kapsamında; polarite analizi, keşifsel veri analizi, konu modelleme, makine öğrenmesi, ağ analizi, makine çevirisi, veri görselleştirme ve metin benzerliği çalışmaları yapılmıştır.

Çalışmada elde edilen verilerden yola çıkarak, sonuçların ilki suçla mücadelede toplum destekli güvenlik hizmeti kapsamında kolluğun paylaşımlarının önemli olduğudur. Etki gücü yüksek mesajlar, sosyal medyada daha çok ilgi görmek ve hızlıca yayılmaktadır. Sonuçların ikincisi sosyal medya üzerinden makine öğrenmesi yöntemleri kullanılarak oldukça başarılı şekilde siber zorbalık içeren mesajların sınıflandırabildiğidir. Çalışma kapsamında 19 farklı makine öğrenimi algoritması kullanılmış olup, en başarılı sonuç veren LGBM algoritması ile %90.2 F1 Skoru elde edilmiştir. LGBM algoritmasının denetimsiz öğrenme kapsamında suçla mücadelede kullanılması halinde, sosyal medyadaki suçların tespiti için gerekli işgücünü yaklaşık %98'e yakın azaltacağı belirlenmiştir. Sonuçların üçüncüsü ise denetimsiz öğrenme kapsamında büyük veri üzerinde keşifsel veri analizi yaparak, açık kaynak istihbaratı yapılabildiğidir. Veriler üzerinde anomali tespiti yapılarak, olağandışı hesap trafikleri analiz edilmiştir. Bunun yanı sıra sosyal medyanın açık kaynak istihbaratında oldukça nadir bilgileri içerebilmesi açısından önemli olduğu belirlenmiştir.

Çalışma sonucu elde edilen verilerden yola çıkarak, sosyal ağların her yönüyle suç ve suçla mücadelede kullanılabileceği tespit edilmiştir. Yapılan bu tez çalışması ile sosyal ağdaki suç kavramı, adli bilişim, kriminoloji ve kriminalistik disiplinleri açısından incelenmiştir. Elde edilen sonuçların, sosyal ağda suç sorununun kökeninin belirlenmesi ile önleyici ve adli hizmetlerde alınacak tedbirler için fikir verdiği değerlendirilmektedir.

Anahtar Kelimeler: Adli Bilimler, Makine Öğrenmesi, Suç, Twitter, Veri Madenciliği.

SUMMARY

Data Mining in Social Media for Preventive Policing Activities: Using Twitter Analytics for Crime Prevention

We are now in a period where communication media that were in physical spaces in the past are gradually becoming virtual with the widespread use of social networks. The changes experienced with the use of social networks do not only exist in the field of communication, but also in almost every aspect of life. Some of these changes in question are quite good, while some are quite bad. The concept of crime is one of the areas that are usually negatively affected by the change through social media, and studies conducted indicate that social media crimes are gradually increasing.

Through this thesis study, the use of social networks both in crime and the fight against crime was discussed with a holistic approach. The holistic approach includes studies conducted on the role of social media in preventing crime and post-crime detection. It was aimed to reveal the data obtained from the study through different analysis methods.

Besides the data mining methods used in the thesis study, a method from the descriptive/exploratory category was used in terms of analyzing the obtained information. The data used in the study were taken from Twitter. The main limitation of the study is performing analysis on the data, which are in a limited time period. Python 3 programming language was used during the data retrieval and analysis phase. Polarity analysis, exploratory data analysis, topic modeling, machine learning, network analysis, machine translation, data visualization, and text similarity studies were conducted within the scope of the thesis.

The first result based on the data obtained in the study is that the posts shared by law enforcement within the scope of community-supported security service in the fight against crime are significant. Messages with a strong power of influence attract more attention on social media and spread rapidly. The second result is that messages containing cyberbullying can be classified quite successfully by using machine learning methods on social media. 19 different machine learning algorithms were used within the scope of the study, and a 90.2% F1 Score was obtained with the LGBM algorithm, which gave the most successful results. It was determined that the workforce required for the detection of social media crimes would decrease by approximately 98% if the LGBM algorithm was used in the fight against crime within the scope of unsupervised learning. The third result is that open-source intelligence can be performed by conducting exploratory data analysis on big data within the scope of unsupervised learning. Abnormal account traffic was analyzed by detecting anomalies in the data. In addition, it was determined that social media is of importance in open-source intelligence in terms of its ability to contain quite rare information.

Based on the data obtained as a result of the study, it was determined that social networks can be used both in crime and the fight against crime in all aspects. The concept of crime in the social network was examined in terms of digital forensics, criminology, and criminalistic disciplines through this thesis study. It is considered that the obtained results

provide insight for the determination of the origin of the crime problem in the social network and precautions to be taken in preventive and legal services.

Keywords: Crime, Data Mining, Forensic Science, Machine Learning, Twitter.



KAYNAKLAR

- AAMODT A, PLAZA E (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, **7(1)**: 39-59.
- ABDEL-BASSET M, MOHAMED M, SMARANDACHE F, CHANG V (2018). Neutrosophic association rule mining algorithm for big data analysis. *Symmetry*, **10(4)**: 106.
- ABDULLAH N, ISMAIL SA, SOPHIAYATI S, SAM SM (2015). Data quality in big data: a review. *Int. J. Advance Soft Compu. Appl*, **7(3)**: 17-27.
- AGHABABAEI S, MAKREHCHI M (2018). Mining Twitter data for crime trend prediction. *Intelligent Data Analysis*, **22(1)**: 117-141.
- AHMED A (2020). "From Data to Wisdom" Using Machine Learning Capabilities in Accounting and Finance Professionals. *Talent Development & Excellence*, 12.
- AHMED YA, AHMAD MN, AHMAD N, ZAKARIA NH (2019). Social media for knowledge-sharing: A systematic literature review. *Telematics and informatics*, **37**: 72-112.
- AIKEN M. (2019). An Updated Evaluation of Google Translate Accuracy. *Studies in linguistics and literature*, **3(3)**: 253-260.
- AKÇA EB, SAYIMER İ (2017). Siber zorbalık kavramı, türleri ve ilişkili olduğu faktörler: Mevcut araştırmalar üzerinden bir değerlendirme. *AJIT-e: Bilişim Teknolojileri Online Dergisi*, **8(30)**: 7-19.
- ALBAWI S, MOHAMMED TA, AL-ZAWI S (2017, August). Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET) (pp. 1-6). IEEE.
- ALELYANI S, TANG J, LIU H (2018). Feature selection for clustering: A review. In Data Clustering (pp. 29-60). Chapman and Hall/CRC.

- ALI A, SYED AM (2020). Cyberbullying Detection using Machine Learning. *Pakistan Journal of Engineering and Technology*, **3(2)**: 45-50.
- ALLAHYARI M, POURIYEH S, ASSEFI M, SAFAEI S, TRIPPE ED, GUTIERREZ JB, KOCHUT K (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919.
- AL-TAIE MZ, KADRY S (2017). Information diffusion in social networks. In *Python for Graph and Network Analysis* (pp. 165-184). Springer, Cham.
- ALTAY EV, ALATAŞ B (2019). Detection of Cyberbullying in Social Networks Using Machine Learning Methods, in: *International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism, IBIGDELFT 2018 - Proceedings*. <https://doi.org/10.1109/IBIGDELFT.2018.8625321>
- AMRA I AA, MAGHARI AY (2017, May). Students performance prediction using KNN and Naïve Bayesian. In *2017 8th International Conference on Information Technology (ICIT)* (pp. 909-913). IEEE.
- ANDREWS S, BREWSTER B, DAY T (2018). Organised crime and social media: a system for detecting, corroborating and visualising weak signals of organised crime online. *Security Informatics*, **7(1)**: 3.
- ANTONCIC M (2020). A paradigm shift in the board room: Incorporating sustainability into corporate governance and strategic decision-making using big data and artificial intelligence. *Journal of Risk Management in Financial Institutions*, **13(4)**: 290-294.
- ARNABOLDI M, BUSCO C, CUGANESAN S (2017). Accounting, accountability, social media and big data: revolution or hype? *Accounting, auditing & accountability journal*. **30(4)**: 762-776.
- ARORA B (2016). Exploring and analyzing internet crimes and their behaviours. *Perspectives in Science*, **8**: 540-542.
- ARYONO A, PRASTYANTI RA (2020). Protection of Children from Violence in Social Media in The New Normal Era. *Veteran Justice Journal*, **2(1)**.
- ASADI M, AGAH A (2017, November). Characterizing user influence within Twitter. In *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing* (pp. 122-132). Springer, Cham.

- ASONGU S, NWACHUKWU J, ORIM SM, PYKE C (2019). Crime and social media. *Information Technology & People*. **32(5)**: 1215-1233.
- ATALAY M, ÇELİK E (2017). Büyük Veri Analizinde Yapay Zekâ ve Makine Öğrenmesi Uygulamaları. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, **9(22)**: 155-172.
- ATEŞ EC, BOSTANCI E (2018). Determination of the Child Suspects and Victims's Profiles in Crimes Involving the Use, Sale and Purchase of Drug and Stimulant Substances, International Data Science Engineering Workshop IDSEW 2018, Karabük.
- ATEŞ EC, BOSTANCI E, GÜZEL MS (2020) Security Evaluation of Industry 4.0: Understanding Industry 4.0 on the Basis of Crime, Big Data, Internet of Thing (IoT) and Cyber Physical Systems. *Güvenlik Bilimleri Dergisi, (International Security Congress Special Issue)*, 29-50.
- ATEŞ EC, TOKAY A (2018). Sosyal Ağ (Medya) Kullanımı ve Suç Mağduriyeti. *Online Journal of Technology Addiction & Cyberbullying*, **5(1)**: 1-33.
- AZEVEDO AIRL, SANTOS MF (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.
- BANUMATHI S, ALOYSIUS A (2017). Predictive analytics concepts in big data-a survey. *International Journal of Advanced Research in Computer Science*, **8(8)**.
- BARBERÁ P (2015). How social media reduces mass political polarization. Evidence from Germany, Spain, and the US. Job Market Paper, New York University, 1-46.
- BARNAGHI P, GHAFARI P, BRESLIN JG (2016, March). Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. In IEEE second international conference on big data computing service and applications (Big Data Service) (pp. 52-57). IEEE.
- BARTLETT J, MILLER C, CRUMP J, MIDDLETON L (2013). Policing in an information age. Demos Publishing, London.
- BASUKI S, AZHAR Y, MINARNO AE, ADITYA CSK, SUMADI FDS, RAMADHAN AI (2019, July). Detection of Reference Topics and Suggestions using Latent Dirichlet Allocation (LDA). In 12th International Conference on Information & Communication Technology and System (ICTS) (pp. 1-5). IEEE.

- BELL J (2014). Machine learning: hands-on for developers and technical professionals. John Wiley & Sons.
- BENGFORT B, BILBRO R, OJEDA T (2018). Applied Text Analysis with Python: Enabling Language-aware Data Products with Machine Learning. O'Reilly Media, Inc.
- BENIWAL S, ARORA J (2012). Classification and feature selection techniques in data mining. *International journal of engineering research & technology (IJERT)*, **1(6)**: 1-6.
- BERMAN E, FELTER JH, SHAPIRO JN (2020). Small wars, big data: the information revolution in modern conflict. Princeton University Press.
- BLEI DM (2012). Probabilistic topic models. *Communications of the ACM*, **55(4)**: 77-84.
- BLEI DM, LAFFERTY JD (2009). Topic models. Text mining: classification, clustering, and applications (pp. 71-93). Chapman and Hall/CRC.
- BONSÓN E, ROYO S, RATKAI M (2015). Citizens' engagement on local governments' Facebook sites. An empirical analysis: The impact of different media and content types in Western Europe. *Government information quarterly*, **32(1)**: 52-62.
- BONZANINI M (2016). Mastering social media mining with Python. Packt Publishing Ltd.
- BORLEA ID, PRECUP RE, DRAGAN F, BORLEA AB (2017). Centroid update approach to K-means clustering. *Advances in Electrical and Computer Engineering*, **17(4)**: 3-11.
- BOZYIĞIT A (2018). Turkish Cyberbullying Dataset. Erişim Adresi: [<https://www.kaggle.com/abozyigit/turkish-cyberbullying>]. Erişim Tarihi: 5/1/2019.
- BOZYIĞIT A, UTKU S, NASİBOĞLU E (2019). Cyberbullying Detection by Using Artificial Neural Network Models, in: 4th International Conference on Computer Science and Engineering (UBMK). IEEE, pp. 520–524.
- BRAILOVSKAIA J, TEISMANN T, MARGRAF J (2018). Cyberbullying, positive mental health and suicide ideation/behavior. *Psychiatry research*, **267**: 240-242.
- BRANDL SG (2019). Criminal investigation. 4th Edition. SAGE Publications.
- BURCHER M, WHELAN C (2019). Intelligence-led policing in practice: Reflections from intelligence analysts. *Police quarterly*, **22(2)**: 139-160.

- BUTT SS, JAMIL F, KHALID R (2019). Cyberbullying, self-esteem and interpersonal trust in young adults. *Pakistan Journal of Social and Clinical Psychology*, **17(1)**: 38-46.
- CADWALLADR C, GRAHAM-HARRISON E (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*, 17, 22.
- CAMP NH (2016). Restoring trust: A grounded theory study of cyberbullying among young women (Doctoral dissertation, Loyola University Chicago).
- CAMPBELL C, YING Y (2011). Learning with support vector machines. *Synthesis lectures on artificial intelligence and machine learning*, **5(1)**: 1-95.
- CAO L (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, **50(3)**: 43.
- CHEN J, HOSSAIN MS, ZHANG H (2020). Analyzing the sentiment correlation between regular tweets and retweets. *Social Network Analysis and Mining*, **10(1)**: 13.
- CHEN M, MAO S, LIU Y (2014). Big data: A survey. *Mobile networks and applications*, **19(2)**: 171-209.
- CIELEN D, MEYSMAN A, ALI M (2016). Introducing data science: big data, machine learning, and more, using Python tools. Manning Publications Co.
- CLARKE C (2006). Proactive policing: Standing on the shoulders of community-based policing. *Police Practice and Research*, **7(1)**: 3-17.
- ÇÖLTEKİN Ç (2020, May). A Corpus of Turkish Offensive Language on Social Media. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020) (pp. 6174-6184).
- CONWAY D (2010). The Data Science Venn Diagram. Erişim Adresi: [<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>]. Erişim Tarihi: 3/12/2019.
- COUPE RT, ARIEL B, MUELLER-JOHNSON K (2019). Crime Solvability Factors: Police Resources and Crime Detection. Springer Nature.
- COX SM, MARCHIONNA S, KOSKI CM, FITCH BD (2017). Introduction to policing. Sage Publications.

- CURIEL RP, CRESCI S, MUNTEAN CI, BISHOP SR (2020). Crime and its fear in social media. *Palgrave Communications*, **6(1)**: 1-12.
- CURISKIS SA, DRAKE B, OSBORN TR, KENNEDY PJ (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management*, **57(2)**.
- DAHKA ZY, HAJIHEYDARI N, ROUHANI S (2020). User response to e-WOM in social networks: how to predict a content influence in Twitter. *International Journal of Internet Marketing and Advertising*, **14(1)**: 91-111.
- DANJUMA KJ (2015). Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients. arXiv preprint arXiv:1504.04646.
- DAWSON M, LIEBLE M, ADEBOJE A (2018). Open source intelligence: Performing data mining and link analysis to track terrorist activities. In *Information Technology-New Generations* (pp. 159-163). Springer, Cham.
- DELAVALLE T, BERTRAND P, THOUVENOT V (2017). Extracting future crime indicators from social media. In *Using open data to detect organized crime threats* (pp. 167-198). Springer, Cham.
- DESHPANDE SP, THAKARE VM (2010). Data mining system and applications: A review. *International Journal of Distributed and Parallel systems (IJDPS)*, **1(1)**: 32-44.
- DEY A (2016). Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies*, **7(3)**: 1174-1179.
- DING Q, LI Z, HAERI S, TRAJKOVIĆ L (2018). Application of machine learning techniques to detecting anomalies in communication networks: Datasets and feature selection algorithms. In *Cyber threat intelligence* (pp. 47-70). Springer, Cham.
- DRAGONI M, FEDERICI M, REXHA A (2019). An unsupervised aspect extraction strategy for monitoring real-time reviews stream. *Information processing & management*, **56(3)**: 1103-1118.
- DUBEY A, KAMATH S, KANAKIA D (2016). Learning data mining techniques. *International Journal of Computer Applications*, **136(11)**.

- DUNN AG, COIERA E, BOURGEOIS FT, (2018). Unreported links between trial registrations and published articles were identified using document similarity measures in a cross-sectional analysis of Clinical Trials. gov. *Journal of Clinical Epidemiology*, **95**: 94-101.
- DWIVEDI S (2020). Theoretical exploration of the new borders of globalisation. Re-imagining Border Studies in South Asia, 99.
- ECONOMIST (2017). The world's most valuable resource is no longer oil, but data. Erişim Adresi: [<https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data>]. Erişim Tarihi: 25/10/2019.
- EMRE İE, EROL ÇS (2017). Veri Analizinde İstatistik mi Veri Madenciliği mi? *International Journal of Informatics Technologies*, **10(2)**.
- ENGLANDER E, DONNERSTEIN E, KOWALSKI R, LIN CA, PARTI K (2017). Defining cyberbullying. *Pediatrics*, **140(S-2)**: 148-151.
- ERTEL W (2017). Introduction to Artificial Intelligence, 1-356, Springer & UTICS.
- ETHNOLOGUE (2020). Erişim Adresi: [<https://www.ethnologue.com/about>]. Erişim Tarihi: 10/09/2020.
- FERGUSON AG (2016). Policing predictive policing. Wash. UL Rev., 94, 1109.
- FERNANDEZ M, DICKINSON T, ALANI H (2017, September). An analysis of UK policing engagement via social media. In International Conference on Social Informatics (pp. 289-304). Springer, Cham.
- FERRARI E (2018). Fake accounts, real activism: Political faking and user-generated satire as activist intervention. *New Media & Society*, 20(6), 2208-2223.
- FERREIRA PC, SIMÃO AV, FERREIRA A, SOUZA S, FRANCISCO S (2016). Student bystander behavior and cultural issues in cyberbullying: When actions speak louder than words. *Computers in Human Behavior*, **60**: 301-311.
- FISH JT, MILLER LS, BRASWELL MC, WALLACE EW (2013). Crime scene investigation. Routledge.

- GABARRON E, DORRONZORO E, RIVERA-ROMERO O, WYNN R (2019). Diabetes on Twitter: a sentiment analysis. *Journal of diabetes science and technology*, **13(3)**: 439-444.
- GARCIA D, MAVRODIEV P, CASATI D, SCHWEITZER F (2017). Understanding popularity, reputation, and social influence in the twitter society. *Policy & Internet*, **9(3)**: 343-364.
- GE Z, SONG Z, DING SX, HUANG B (2017). Data mining and analytics in the process industry: The role of machine learning. *IEEE Access*, 5, 20590-20616.
- GEDİK D (2019). Bilişim Suçlarında IP Tespiti ile Ekran Görüntüleri Çıktılarının İspat Değeri. *Bilişim Hukuku Dergisi*, **1(1)**: 51-84.
- GENTNER S (2016). Industry 4.0: reality, future or just science fiction? How to convince today's management to invest in tomorrow's future! Successful strategies for industry 4.0 and manufacturing IT. *CHIMIA International Journal for Chemistry*, **70(9)**: 628-633.
- GHEISARI M, WANG G, BHUIYAN MZA (2017, JULY). A survey on deep learning in big data. In *IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC)* (Vol. 2, pp. 173-180). IEEE.
- GLEAVE A, STEINRUECKEN C (2017). Making compression algorithms for Unicode text. *arXiv preprint arXiv:1701.04047*.
- GOMAA WH, FAHMY AA (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, **68(13)**: 13-18.
- GOMEZ-ALANIS A, PEINADO AM, GONZALEZ JA, GOMEZ AM (2019). A Gated Recurrent Convolutional Neural Network for Robust Spoofing Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12), 1985-1999.
- GOODFELLOW I, BENGIO Y, COURVILLE A (2016). *Deep learning*. MIT press.
- GÖVER İH, ŞENOL D (2017). Toplumsal Bir Olgu Olarak Suç: Sosyolojik Bir Analiz. *Journal of Current Researches on Social Sciences*, **7(2)**: 191-208.

- GRČAR M, CHEREPNALKOSKI D, MOZETIČ I, NOVAK PK (2017). Stance and influence of Twitter users regarding the Brexit referendum. *Computational social networks*, **4(1)**: 6.
- GROSS AG, HARMON JE (2016). The Internet revolution in the sciences and humanities. Oxford University Press.
- GUDMUNDSSON J, MENKES F (2018). Swedish Natural Language Processing with Long Short-term Memory Neural Networks: A Machine Learning-powered Grammar and Spell-checker for the Swedish Language. Erişim Adresi: [<https://www.diva-portal.org/smash/get/diva2:1232482/FULLTEXT01.pdf>]. Erişim Tarihi: 25/12/2019.
- GUO Y, LIU Y, OERLEMANS A, LAO S, WU S, LEW MS (2015). Deep learning for visual understanding: A review. *Neurocomputing*, 187, 27-48.
- GÜRHAN N (2017). Her Yönü ile Akran Zorbalığı. *Türkiye Klinikleri Psychiatric Nursing-Special Topics*, **3(2)**: 175-181.
- HAGAN FE, DAIGLE LE (2018). Introduction to criminology. SAGE Publications.
- HALEVI G, MOED HF (2012). The evolution of big data as a research and scientific topic: overview of the literature. *Research trends*, **30(1)**: 3-6.
- HAMM JA, TRINKNER R, CARR J D (2017). Fair process, trust, and cooperation: Moving toward an integrated framework of police legitimacy. *Criminal justice and behavior*, **44(9)**: 1183-1212.
- HAMZAH A, HIDAYATULLAH AF, PERSADA AG (2020). Discovering trends of mobile learning research using topic modelling approach. *Int. J. Interact. Mob. Technol.* <https://doi.org/10.3991/ijim.v14i09.11069>
- HAN J, PEI J, KAMBER M (2011). Data mining: concepts and techniques. Elsevier.
- HAND DJ, ADAMS NM (2014). Data Mining. Wiley StatsRef: Statistics Reference Online, 1-7.
- HARDENIYA N, PERKINS J, CHOPRA D, JOSHI N, MATHUR I (2016). Natural Language Processing: Python and NLTK. Packt Publishing Ltd.

- HASSAN SU, SHABBIR M, IQBAL S, SAID A, KAMIRAN F, NAWAZ R, SAIF U (2019). Leveraging Deep Learning and SNA approaches for Smart City Policing in the Developing World. *International Journal of Information Management*, 102045.
- HEMALATHA M, MEGALA S (2011). Mining Techniques in Health Care: A Survey of Immunization. *Journal of Theoretical and Applied Information Technology*, **25(2)**: 65-70.
- HESS KM, ORTHMANN CH, CHO HL (2016). Criminal investigation. Cengage learning.
- HEY J (2004). The data, information, knowledge, wisdom chain: the metaphorical link. *Intergovernmental Oceanographic Commission*, **26**: 1-18.
- HILBE JM. (2016). Practical guide to logistic regression. CRC Press.
- HOLT TJ, BOSSLER AM, SEIGFRIED-SPELLAR KC (2015). Cybercrime and digital forensics: An introduction. Routledge.
- HONG S, NADLER D (2012). Which candidates do the public discuss online in an election campaign? The use of social media by 2012 presidential candidates and its impact on candidate salience. *Gov. Inf. Q.* <https://doi.org/10.1016/j.giq.2012.06.004>
- HOSSNY AH, MOSCHUO T, OSBORNE G, MITCHELL L, LOTHIAN N (2018). Enhancing keyword correlation for event detection in social networks using SVD and k-means: Twitter case study. *Social Network Analysis and Mining*, **8(1)**: 49.
- HOUCK MM, SIEGEL JA (2009). Fundamentals of forensic science. Academic Press.
- HUANG J Y (2017). Web mining for the mayoral election prediction in Taiwan. *Aslib Journal of Information Management* <https://doi.org/10.1108/AJIM-02-2017-0035>
- HUANG L, LIU G, CHEN T, YUAN H, SHI P, MIAO Y (2020). Similarity-Based Emergency Event Detection in Social Media. *Journal of Safety Science and Resilience*, 1-32. doi: <https://doi.org/10.1016/j.jnlssr.2020.11.003>
- HUBER S, WIEMER H, SCHNEIDER D, IHLENFELDT S (2019). DMME: Data mining methodology for engineering applications—a holistic extension to the CRISP-DM model. *Procedia Cirp*, **79**: 403-408.
- HUSAIN F (2020). Arabic Offensive Language Detection Using Machine Learning and Ensemble Machine Learning Approaches. arXiv Prepr. arXiv2005.08946.

- IDRIS IK (2018). Government social media in Indonesia: Just another information dissemination tool. *Jurnal Komunikasi: Malaysian Journal of Communication*, **34(4)**.
- IRSHAD S, SOOMRO TR (2018). Identity theft and social media. *International Journal of Computer Science and Network Security*, **18(1)**: 43-55.
- JACOBI C, VAN ATTEVELDT W, WELBERS K (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, **4(1)**: 89-106.
- JAMES G, WITTEN D, HASTIE T, TIBSHIRANI R (2013). Statistical learning. An Introduction to Statistical Learning, Springer: 15-57.
- JELODAR H, WANG Y, YUAN C, FENG X, JIANG X, LI Y, ZHAO L (2019). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, **78(11)**: 15169-15211.
- JIAO ZL (2017). Internet Revolution and Characteristics of Logistic Service Development in China. In *Contemporary Logistics in China* (pp. 193-211). Springer, Singapore.
- JINDAL P, KUMAR D (2017). A review on dimensionality reduction techniques. *International journal of computer applications*, **173(2)**: 42-46.
- JO S, TRUMMER I, YU W, WANG X, YU C, LIU D, MEHTA N (2019). AggChecker: a fact-checking system for text summaries of relational data sets. *Proceedings of the VLDB Endowment*, **12(12)**, 1938-1941.
- JOSEPH SR, HLOMANI H, LETSHOLO K (2016). Data mining algorithms: an overview. *Neuroscience*, **12(3)**: 719-743.
- JOSHI AV (2020). Machine Learning and Artificial Intelligence. Springer International Publishing.
- JUHLIN JA, RICHARDSON J (2016, May). Social media for intelligence: practical examples of analysis for understanding. In *Next-Generation Analyst IV* (Vol. 9851, p. 98510I). International Society for Optics and Photonics.
- KAGGLE (2020). Kernel Language. Erisim: Adresi: [https://www.kaggle.com/search?q=machine+learning+in+%3Anotebooks]. Erisim Tarihi: 18/05/2020.

- KAMP SM, POTTS GF, DONCHIN E (2015). On the roles of distinctiveness and semantic expectancies in episodic encoding of emotional words. *Psychophysiology*. <https://doi.org/10.1111/psyp.12537>
- KANAN T, ALDAAJA A, HAWASHIN B (2020). Cyber-Bullying and Cyber-Harassment Detection Using Supervised Machine Learning Techniques in Arabic Social Media Contents. **J. Internet Technol.** **21**: 1409–1421.
- KÄTSYRI J, KINNUNEN T, KUSUMOTO K, OITTINEN P, RAVAJA N (2016). Negativity bias in media multitasking: The effects of negative social media messages on attention to television news broadcasts. *PLoS One*. <https://doi.org/10.1371/journal.pone.0153712>
- KEET CM, ŁAWRYNOWICZ A, D'AMATO C, KALOUSIS A, NGUYEN P, PALMA R, STEVENS R, HILARIO M (2015). The data mining optimization ontology. *Journal of web semantics*, **32**: 43-53.
- KEMPE D, KLEINBERG J, TARDOS É (2005, July). Influential nodes in a diffusion model for social networks. In *International Colloquium on Automata, Languages, and Programming* (pp. 1127-1138). Springer, Berlin, Heidelberg.
- KHAN N, YAQOOB I, HASHEM IAT, INAYAT Z, ALI M, KAMALELDIN W, ALAM M, SHIRAZ M, GANI A (2014). Big data: survey, technologies, opportunities, and challenges. *The scientific world journal*, 1-18.
- KHAN R, SHRIVASTAVA P, KAPOOR A, TIWARI A, MITTAL A 2020. Social media analysis with AI: sentiment analysis techniques for the analysis of twitter covid-19 data. *J. Critical Rev.* **7(9)**: 2761–2774.
- KHARE AR, SHRIVASTA P (2018). Data mining for the internet of things. In *Exploring the Convergence of Big Data and the Internet of Things* (pp. 181-191). IGI Global.
- KOPER CS, LUM C, WU X, FRITZ N (2020). Proactive policing in the United States: a national survey. *Policing: An International Journal*, **46(5)**: 861-875.
- KOYUNCUGİL A, ÖZGÜLBAŞ N (2009). Veri madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları. *Bilişim Teknolojileri Dergisi*, **2(2)**.
- KULKARNİ A, SHIVANANDA A (2019). Natural language processing recipes. Apress.

- KURNIAWATI D, PRAYITNO E, SARI DF, PUTRA SN (2019). Sentiment Analysis of Twitter Use on Policy Institution Services using Naïve Bayes Classifier Method. In Journal of International Conference Proceedings, 2(1), 33.
- LAHITANI AR, PERMANASARI AE, SETIAWAN NA (2016). Cosine similarity to determine similarity measure: Study case in online essay assessment, in: Proceedings of 4th International Conference on Cyber and IT Service Management, CITSM 2016. <https://doi.org/10.1109/CITSM.2016.7577578>
- LAINE S, AILA T (2016). Temporal ensembling for semi-supervised learning. arXiv preprint arXiv:1610.02242.
- LANE H, HOWARD C, HAPKE MH (2019). Natural Language Processing in Action, Manning.
- LANTZ B (2013). Machine Learning with R. Olton, United Kingdom, Packt Publishing Ltd.
- LAROSE DT, LAROSE CD (2014). Discovering knowledge in data: an introduction to data mining (Vol. 4). John Wiley & Sons.
- LAXMAN S, SASTRY PS (2006). A survey of temporal data mining. Sadhana, 31(2), 173-198.
- LEE WM (2019). Python Machine Learning. John Wiley & Sons.
- LEVENTAKIS G, HABERFELD MR (2018). Community-Oriented Policing and Technological Innovations. Springer International Publishing.
- LEVY R, SREBNICK AG (2017). Crime and Culture: An Historical Perspective. Routledge.
- LEWIS L (2016). What Your Audience is Doing When They're Not Listening to You. Erisim: Adresi: [[https:// www.allaccess.com/merge/archive/24227/what-your-audience-is-doing-when-they-re-not](https://www.allaccess.com/merge/archive/24227/what-your-audience-is-doing-when-they-re-not)]. Erisim Tarihi: 19/02/2020.
- LEWIS L (2018). 2018 Update: What Happens In An Internet Minute. Erisim: Adresi: [<https://www.allaccess.com/merge/archive/28030/2018-update-what-happens-in-an-internet-minute>]. Erisim Tarihi: 19/02/2020.
- LEWIS L, CALLAHAN C (2020). Infographic: What Happens in an Internet Minute 2020? Erisim: Adresi: [<https://www.allaccess.com/merge/archive/31294/infographic-what-happens-in-an-internet-minute>]. Erisim Tarihi: 11/09/2020.

- LIU Y, ZAVARSKY P, MALIK Y (2019). Non-linguistic Features for Cyberbullying Detection on a Social Media Platform Using Machine Learning, in: International Symposium on Cyberspace Safety and Security. Springer, Cham, pp. 391–406.
- LUO J M, VU HQ, LI G, LAW R (2020). Topic modelling for theme park online reviews: analysis of Disneyland. *J. Travel Tour. Mark.* <https://doi.org/10.1080/10548408.2020.1740138>
- MA L, ZHANG Y (2015, October). Using Word2Vec to process big text data. In 2015 IEEE International Conference on Big Data (Big Data), 2895-2897. IEEE.
- MAINZER K (2020). Artificial intelligence-When do machines take over? Springer Nature.
- MANOGARAN G, THOTA C, LOPEZ D, VIJAYAKUMAR V, ABBAS K M, SUNDARSEKAR R (2017). Big data knowledge system in healthcare. In Internet of things and big data technologies for next generation healthcare, 133-157. Springer, Cham.
- MARISCAL G, MARBAN O, FERNANDEZ C (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, **25(2)**: 137-166.
- MCHUGH ML (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, **22(3)**: 276-282.
- MEIJER A, THAENS M (2013). Social media strategies: Understanding the differences between North American police departments. *Gov. Inf. Q.* <https://doi.org/10.1016/j.giq.2013.05.023>
- MEIJER AJ, TORENVLIED R (2016). Social Media and the New Organization of Government Communications. *Am. Rev. Public Adm.* <https://doi.org/10.1177/0275074014551381>
- MILLER JD (2017). Statistics for Data Science: Leverage the power of statistics for Data Analysis, Classification, Regression, Machine Learning, and Neural Networks. Packt Publishing Ltd.
- MISTEK E, FIKIET MA, KHANDASAMMY SR, LEDNEV IK (2018). Toward locard's exchange principle: recent developments in forensic trace evidence analysis. *Analytical chemistry*, **91(1)**: 637-654.

- MOHAMAD IB, USMAN D (2013). Standardization and its effects on K-means clustering algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, **6(17)**: 3299-3303.
- MOHAMMED M, KHAN MB, BASHIER EBM (2016). Machine learning: algorithms and applications. CRC Press.
- MOHLER GO, SHORT MB, MALINOWSKI S, JOHNSON M, TITA GE, BERTOZZI AL, BRANTINGHAM PJ (2015). Randomized controlled field trials of predictive policing. *Journal of the American statistical association*, **110(512)**: 1399-1411.
- MOHRI M, ROSTAMIZADEH A, TALWALKAR A (2018). Foundations of Machine Learning, 2th Edition, The MIT Press, London, England.
- MUKHOPADHYAY A, MAULIK U, BANDYOPADHYAY S, COELLO CAC (2013). A survey of multiobjective evolutionary algorithms for data mining: Part I. *IEEE Transactions on Evolutionary Computation*, **18(1)**, 4-19.
- MÜLLER K, SCHWARZ C (2019). Fanning the flames of hate: Social media and hate crime. Available at SSRN 3082972.
- MULYADI U, FITRIANA L (2018). Hashtag (#) as message identity in virtual community. *Jurnal The Messenger*, **10(1)**: 44-53.
- MUNEER A, FATI SM (2020). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Futur. Internet* **12**: 187.
- MURALIDHARAN S, LA FERLE C (2018). Religious symbolism in the digital realm: A social advertising approach to motivate bystanders to aid victims of cyberbullying. *International Journal of Consumer Studies*, **42(6)**: 804-812.
- NAMAZI H (2018). Can we mathematically correlate brain memory and complexity. *ARC J. Neurosci*, **3(2)**: 10-12.
- NEWBURN T (2017). Criminology, Routledge Press.
- NGAI EWT, XIU L, CHAU DCK (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Syst. Appl.* <https://doi.org/10.1016/j.eswa.2008.02.021>

- NONGHULOO MS, KRISHNAMURTHI K (2017). Spell checker for Khasi language. *Int. J. Softw. Eng*, **7(1)**: 1-12.
- NOVOTNÁ T (2020). Document Similarity of Czech Supreme Court Decisions. *Masaryk University Journal of Law and Technology*, **14(1)**: 105-122.
- OLSON DL, LAUHOFF G (2019). Descriptive data mining. In *Descriptive Data Mining* (pp. 129-130). Springer, Singapore.
- OMAND D, BARTLETT J, MILLER C (2012). Introducing social media intelligence (SOCMINT). *Intelligence and National Security*, **27(6)**: 801-823.
- OSCE (2017). Guidebook Intelligence-Led Policing, ISBN 978-3-903128-04-0, TNTD/SPMU Publication Series Vol. 13, Vienna.
- OTTER DW, MEDINA JR, KALITA JK (2018). A survey of the usages of deep learning in natural language processing. arXiv preprint arXiv:1807.10854.
- ÖZDEMİR S (2016). Principles of Data Science. Packt Publishing Ltd.
- ÖZEL SA, AKDEMİR S, SARAÇ E, AKSU H (2017). Detection of cyberbullying on social media messages in Turkish, in: 2nd International Conference on Computer Science and Engineering, UBMK 2017. <https://doi.org/10.1109/UBMK.2017.8093411>
- ÖZGÜL F (2016). Python 3 için Türkçe Kılavuz Sürüm 3, (E-Kitap) Erisim: Adresi: [<https://xn--e-kitapcm-2pb.com/kitap/python-3-ic-in-tu-rkc-e-kilavuz.html>]. Erisim Tarihi: 11/09/2019.
- ÖZKAN M, BORAN L (2014). Veri Madenciliğinin Finansal Kararlarda Kullanımı. *Çankırı Karatekin Üniversitesi İİBF Dergisi*, **4(1)**: 59-82.
- ÖZTÜRK N, AYVAZ S (2018). Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telemat. Informatics*. <https://doi.org/10.1016/j.tele.2017.10.006>
- PAGLIARDINI M, GUPTA P, JAGGI M (2017). Unsupervised learning of sentence embeddings using compositional n-gram features. arXiv preprint arXiv:1703.02507.
- PARYASTO M, ALAMSYAH A, RAHARDJO B (2014, May). Big-data security management issues. In *IEEE 2nd International Conference on Information and Communication Technology (ICoICT)*, 59-63.

- PATEL R, PASSI K (2020). Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning. *IoT* 1, 218–239.
- PAULO D, FISCHL B, MARKOW T, MARTIN M, SHAKARIAN P (2013, August). Social network intelligence analysis to combat street gang violence. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, 1042-1049.
- PERLIGER A, PEDAHZUR A (2011). Social network analysis in the study of terrorism and political violence. *PS: Political Science & Politics*, **44(1)**: 45-50.
- PERRY WL (2013). Predictive policing: The role of crime forecasting in law enforcement operations. Rand Corporation.
- PETHERICK W, TURVEY BE, FERGUSON CE (2009). *Forensic Criminology*. Academic Press.
- PETROVA MM, SUSHCHENKO O, TRUNINA I, DEKHTYAR N (2018, October). Big data tools in processing information from open sources. In *2018 IEEE First International Conference on System Analysis & Intelligent Computing (SAIC)*, 1-5.
- PIATETSKY G (2020). Python leads the 11 top Data Science, Machine Learning platforms: Trends and Analysis. KDnuggets. Erisim Adresi: [http: [https://www.kdnuggets.com/2019/05/poll-top-data-science-machine learning-platforms.html/2](https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html/2)]. Erisim Tarihi: 25/05/2020.
- PONSAERS P (2001). Reading about “community (oriented) policing” and police models. *Policing: an international journal of police strategies & management*.
- PRABHAKAR E, SANTHOSH M, KRISHNAN AH, KUMAR T, SUDHAKAR R (2019). Sentiment Analysis of US Airline Twitter Data using New Adaboost Approach. *Int. J. Eng. Res. Technol.* 7: 1–6.
- PRABHAKAR KD, PRASAD DA (2020). Informational flow on Twitter–Corona virus outbreak–topic modelling approach. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, **11(3)**.
- PRASETYA DD, WIBAWA AP, HIRASHIMA T (2018). The performance of text similarity algorithms. *International Journal of Advances in Intelligent Informatics*, **4(1)**: 63-69.

- PRASETYO VR, WINARKO E (2017). Rating of Indonesian sinetron based on public opinion in Twitter using Cosine similarity, in: Proceedings - 2016 2nd International Conference on Science and Technology-Computer, ICST 2016. <https://doi.org/10.1109/ICSTC.2016.7877374>
- PUASCHUNDER JM (2020). On Freedom in the Artificial Age. RAIS Conference Proceedings - Research Association for Interdisciplinary Studies. 75-80.
- PUGHAZENDI N, PUNITHAVALLI M (2011). Discovering Patterns from Temporal Databases Using Temporal Association Rule“. *Global Journal of Computer Science and Technology*, **11(14)**: 1-5.
- QIU X, ZHANG L, SUGANTHAN PN, AMARATUNGA GA (2017). Oblique random forest ensemble via Least Square Estimation for time series forecasting. *Information Sciences*, **420**: 249-262.
- QUANDT T, FRISCHLICH L, BOBERG S, SCHATTO-ECKRODT T (2019). Fake news. *The international encyclopedia of Journalism Studies*, 1-6.
- RAMESH V (2017). An Efficient Survey of Managing Big Data Analytics using Swarm Intelligence. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2**: 933–936.
- RANA M, ATIQUE M (2019). Language Translation: Enhancing Bi-Lingual Machine Translation Approach Using Python. *i-Manager's J. Comput. Sci.* **7**: 36.
- RATCLIFFE JH (2016). Intelligence-Led Policing, 2nd Edition, Routledge.
- RAZA MO, MEMON M, BHATTI S, BUX R (2020). Detecting Cyberbullying in Social Commentary Using Supervised Machine Learning, in: Advances in Intelligent Systems and Computing. https://doi.org/10.1007/978-3-030-39442-4_45
- REHUREK R, SOJKA P (2010). Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.
- REINER R (2016). Crime, the mystery of the common-sense concept. John Wiley & Sons.
- REZAIE B, ZAHEDI M, MASHAYEKHI H (2020). Measuring time-sensitive user influence in Twitter. *Knowledge and Information Systems*, 1-28.
- RIQUELME F, GONZÁLEZ-CANTERGIANI P (2016). Measuring user influence on Twitter: A survey. *Information processing & management*, **52(5)**: 949-975.

- ROCA J (2020). Lexicon Based Approaches to Sentiment Analysis of Spanish Tweets: A Comparative Study.
- ROKACH L, MAIMON O (2015). Data mining with decision trees: theory and applications (2nd Edition), World scientific, Singapore, ISBN:978-9814590075.
- RONG X (2014). Word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.
- ROVNER J (2019). Intelligence in the age of Twitter. In Handbook of Terrorism and Counter Terrorism Post 9/11. Edward Elgar Publishing.
- RUKUS J, WARNER ME, ZHANG X (2018). Community policing: Least effective where need is greatest. *Crime & Delinquency*, **64(14)**: 1858-1881.
- SAAR J (2017). A Suitable Amount of Crime and a Cultural-Civilisational Approach. *Juridica Int'l*, 25, 3.
- SAGI O, ROKACH L (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery*, **8(4)**: 1249.
- SAHOO D, LIU C, HOI SC (2017). Malicious URL detection using machine learning: A survey. arXiv preprint arXiv:1701.07179.
- SAHU K, BAI Y, CHOI Y (2020, January). Supervised Sentiment Analysis of Twitter Handle of President Trump with Data Visualization Technique. In 2020 IEEE 10th Annual Computing and Communication Workshop and Conference (CCWC), 0640-0646.
- SALLOUM SA, ALHAMAD AQ, AL-EMRAN M, SHAALAN K (2018). A survey of Arabic text mining. In Intelligent Natural Language Processing: Trends and Applications, pp. 417-431. Springer, Cham.
- SAMPSON F (2017). Intelligent evidence: Using open source intelligence (OSINT) in criminal proceedings. *The Police Journal*, **90(1)**: 55-69.
- SANTORINI B (1990). Part-of-speech tagging guidelines for the penn treebank project (3rd revision). Technical Reports (CIS), 570.
- SAPPELLI M, DE BOER MH, SMIT SK, BOMHOF F (2017). A Vision on Prescriptive Analytics. ALLDATA, 54.

- SARDDAR D, DEY RK, BOSE R, ROY S (2020). Topic Modeling as a Tool to Gauge Political Sentiments from Twitter Feeds. *International Journal of Natural Computing Research (IJNCR)*, **9(2)**: 14-35.
- SARKAR D (2019). Text analytics with Python: a practitioner's guide to natural language processing. Apress.
- SCHROEDER L, SJOQUIST DL, STEPHAN PE (2016). Understanding regression analysis: An introductory guide (Vol. 57). Sage Publications.
- SCOTT MS, ECK JE, KNUTSSON J, GOLDSTEIN H (2016). 11W Problem-oriented policing. *Environmental Criminology and Crime Analysis*, 227.
- SEÇKIN T, KILIMCI ZH (2020). The Evaluation of 5G technology from Sentiment Analysis Perspective in Twitter, in: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU). pp. 1–6.
- SHAFIQUE U, QAISER H (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, **12(1)**: 217-222.
- SHAO L, DUAN Y, SUN X, GAO H, ZHU D, MIAO W (2017, July). Answering Who/When, What, How, Why through Constructing Data Graph, Information Graph, Knowledge Graph and Wisdom Graph. In SEKE (pp. 1-6).
- SHAPIRO A (2017). Reform predictive policing. *Nature News*, **541(7638)**: 458.
- SHAPIRO A (2019). Predictive policing for reform? Indeterminacy and intervention in big data policing. *Surveillance & Society*, **17(3/4)**: 456-472.
- SHARMA P, SINGH A (2017, July). Era of deep neural networks: A review. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.
- SHEKHAWAT BS (2019). Sentiment Classification of Current Public Opinion on BREXIT: Naïve Bayes Classifier Model vs Python's TextBlob Approach. *Natl. Coll. Irel.* .
- SIEGEL JA, MIRAKOVITS K (2015). Forensic science: the basics. CRC Press.

- SIMEONE M, RUSSO C (2017). The growing influence of social and digital media: Impact on consumer choice and market equilibrium. *Br. Food J.* <https://doi.org/10.1108/BFJ-05-2017-0283>
- SINGH A, THAKUR N, SHARMA A (2016, March). A review of supervised machine learning algorithms. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1310-1315). IEEE.
- SINGH R (2020). Importance of Data Analytics in International Trade: A Case of Indian Cotton. In *Application of Big Data and Business Analytics*. Emerald Publishing Limited.
- SINGH R, SINGH S (2020). Text Similarity Measures in News Articles by Vector Space Model Using NLP. *J. Inst. Eng. Ser. B.* <https://doi.org/10.1007/s40031-020-00501-5>
- SOARES C, SIMÃO E (2019). Immersive multimedia in information revolution. In *Trends, Experiences, and Perspectives in Immersive Multimedia and Augmented Reality* (pp. 192-210). IGI Global.
- SOMERVILLE P (2009). Understanding community policing. *Policing: an international journal of police strategies & management*.
- SOORAJ S, MANJUSHA K, ANAND KM, SOMAN KP (2018). Deep learning based spell checker for Malayalam language. *Journal of Intelligent & Fuzzy Systems*, **34(3)**: 1427-1434.
- SPARROW MK (2018). Problem-oriented policing: matching the science to the art. *Crime Science*, **7(1)**: 14.
- SRINIVAS K, RANI BK, GOVRDHAN A (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, **2(02)**: 250-255.
- STALANS LJ, FINN MA (2016). Understanding how the internet facilitates crime and deviance. *An International Journal of Evidence-based Research, Policy, and Practice*. **11(4)**: 501-508.
- STRAUß N, KRUIKEMEIER S, VAN DER MEULEN H, VAN NOORT G (2015). Digital diplomacy in GCC countries: Strategic communication of Western embassies on Twitter. *Gov. Inf. Q.* <https://doi.org/10.1016/j.giq.2015.08.001>

- SUMACHEV AV (2018). Criminal Punishment and Counteraction to Crime: Theoretical Applied Analysis. *Review of law sciences*, **2(4)**: 16.
- SUTTON RS, BARTO AG (2018). Reinforcement learning: An introduction. MIT press.
- TAJUDEEN FP, JAAFAR NI, AININ S (2018). Understanding the impact of social media usage among organizations. *Information & Management*, **55(3)**: 308-321.
- TALIB R, HANIF MK, AYESHA S, FATIMA F (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, **7(11)**: 414-418.
- TALPUR BA, O’SULLIVAN D 2020. Cyberbullying severity detection: A machine learning approach. PLoS One. <https://doi.org/10.1371/journal.pone.0240924>
- TANG XB, LI S, GU N, TAN ML (2019). Exploring repost features of police-generated microblogs through topic and sentiment analysis. The Electronic Library. <https://doi.org/10.1108/EL-02-2019-0044>
- TAYEBI MA, GLÄSSER U (2016). Social network analysis in predictive policing. In *Social Network Analysis in Predictive Policing* (pp. 7-14). Springer, Cham.
- TAYLOR S (2015). *Crime and Criminality: A multidisciplinary approach*. Routledge.
- THENMOZHI D, KANNAN K, ARAVINDAN C (2017, December). A Text Similarity Approach for Precedence Retrieval from Legal Documents. In *FIRE (Working Notes)* (pp. 90-91).
- TREVORS G, KENDEOU P (2019). The Effects of Positive and Negative Emotions on Knowledge Revision. *SSRN Electron. J.* <https://doi.org/10.2139/ssrn.3274111>
- TUNDIS A, JAIN A, BHATIA G, MUHLHAUSER M (2019, July). Similarity analysis of criminals on social networks: An example on Twitter. In *2019 28th International Conference on Computer Communication and Networks (ICCCN)* (pp. 1-9). IEEE.
- TUNDIS A, RUPPERT S, MÜHLHÄUSER M (2020, June). On the Automated Assessment of Open-Source Cyber Threat Intelligence Sources. In *International Conference on Computational Science* (pp. 453-467). Springer, Cham.
- TURING A (1950). Computing machinery and intelligence: *Mind*, 59, 433–460.

- TWITTER (2020). Erişim Adresi: [<https://developer.twitter.com/en/docs>] Erişim Tarihi:20/04/2020.
- UGUR Y (2019). An information-theoretic approach to distributed learning: distributed source coding under logarithmic loss (Doctoral dissertation, Université Paris-Est).
- UNNISA M, AMEEN A, RAZIUDDIN S (2016). Opinion mining on Twitter data using unsupervised learning technique. *International Journal of Computer Applications*, **148(12)**: 12-19.
- VADAPALLI SR, HSIEH G, NAUER KS (2018). TwitterOISNT: automated cybersecurity threat intelligence collection and analysis using twitter data. In Proceedings of the International Conference on Security and Management (SAM) (pp. 220-226).
- VAIDHYANATHAN S, BULOCK C (2014). Knowledge and dignity in the era of “big data”. *The Serials Librarian*, **66(1-4)**: 49-64.
- VAN HEE C, JACOBS G, EMMERY C, DESMET B, LEFEVER E, VERHOEVEN B, DE PAUW G, DAELEMANS W, HOSTE V (2018). Automatic detection of cyberbullying in social media text. PLoS One. <https://doi.org/10.1371/journal.pone.0203794>
- VILLARS RL, OLOFSON CW, EASTWOOD M (2011). Big data: What it is and why you should care. White Paper, IDC, 14, 1-14.
- VOMFELL L, HÄRDLE WK, LESSMANN S (2018). Improving crime count forecasts using Twitter and taxi data. *Decision Support Systems*, **113**: 73-85.
- WAMBA SF, AKTER S, EDWARDS A, CHOPIN G, GNANZOU D (2015). How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, **165**: 234-246.
- WANG N, WANG P, ZHANG B (2010, June). An improved TF-IDF weights function based on information theory. In 2010 International Conference on Computer and Communication Technologies in Agriculture Engineering (Vol. 3, pp. 439-441). IEEE.
- WANG X, WANG S, SUN G, CHEN X, WANG X, LIU Z (2019, June). Apriori Algorithm-Based Association Rules of Event Language Expression Discovery. In International Conference on Applications and Techniques in Cyber Security and Intelligence (pp. 1225-1234). Springer, Cham.

- WEARESOCIAL (2020). Digital 2020: Global digital yearbook, July Global Statshot Report.
Erişim Adresi: [<https://wearesocial.com/digital-2020>] Erişim Tarihi:20/10/2020.
- WIELKI J (2013). Implementation of the big data concept in organizations-possibilities, impediments and challenges. In 2013 Federated Conference on Computer Science and Information Systems (pp. 985-989). IEEE.
- WILLIAMS CB, FEDOROWICZ J, KAVANAUGH A, MENTZER K, THATCHER JB, XU J (2018). Leveraging social media to achieve a community policing agenda. Gov. Inf. Q. <https://doi.org/10.1016/j.giq.2018.03.001>
- WILLIAMS GJ (2009). Rattle: a data mining GUI for R. *The R Journal*, **1(2)**: 45-55.
- WILSON HJ, DAUGHERTY P, BIANZINO N (2017). The jobs that artificial intelligence will create. *MIT Sloan Management Review*, **58(4)**: 14.
- WIRTH R, HIPPE J (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). London, UK: Springer-Verlag.
- XU L, JIANG C, WANG J, YUAN J, REN Y (2014). Information security in big data: privacy and data mining. *IEEE Access*, 2, 1149-1176.
- YANG XS (2019). Introduction to Algorithms for Data Mining and Machine Learning. Academic Press.
- YILMAZ M (2009). Enformasyon ve bilgi kavramları bağlamında enformasyon yönetimi ve bilgi yönetimi. *Ankara Üniversitesi Dil ve Tarih-Coğrafya Fakültesi Dergisi*, **49(1)**: 95-118
- YOUNG T, HAZARIKA D, PORIA S, CAMBRIA E (2018). Recent trends in deep learning based natural language processing. *IEEE Computational intelligence magazine*, **13(3)**: 55-75.
- YUAN X (2017). An improved Apriori algorithm for mining association rules. In AIP conference proceedings (Vol. 1820, No. 1, p. 080005). AIP Publishing LLC.
- ZANERO S (2017). Cyber-physical systems. *Computer*, **50(4)**: 14-16.
- ZHANG J, OTOMO T, LI L, NAKAJIMA S (2019). Cyberbullying Detection on Twitter using Multiple Textual Features, in: 2019 IEEE 10th International Conference on

Awareness Science and Technology, ICAST 2019 - Proceedings.
<https://doi.org/10.1109/ICAwST.2019.8923186>.

ZHANG X D (2020). Machine learning. In A Matrix Algebra Approach to Artificial Intelligence (pp. 223-440). Springer, Singapore.

ZHOU Q, ZHANG C (2017). Emotion evolutions of sub-topics about popular events on microblogs. The Electronic Library. <https://doi.org/10.1108/EL-09-2016-0184>

ZINS C (2007). Conceptual approaches for defining data, information, and knowledge. *J. Am. Soc. Inf. Sci. Technol.* **58**: 479–493.

ŽIŽKA J, DAŘENA F, SVOBODA A (2019). Text Mining with Machine Learning: Principles and Techniques. CRC Press.