**ANKARA ÜNİVERSİTESİ**
**EĞİTİM BİLİMLERİ ENSTİTÜSÜ**

**EĞİTİM BİLİMLERİ ANABİLİM DALI**
**EĞİTİMDE ÖLÇME VE DEĞERLENDİRME PROGRAMI**

# AÇIK UÇLU MADDELERİN OTOMATİK PUANLANMASINDA DOĞAL DİL İŞLEME YÖNTEMİNDEN YARARLANILARAK MAKİNE ÖĞRENMESİ ALGORİTMALARININ KARŞILAŞTIRILMASI

**DOKTORA TEZİ**

**KÜBRA YILMAZ**

**ANKARA**
**ŞUBAT, 2025**


ANKARA ÜNİVERSİTESİ
1946

**ANKARA ÜNİVERSİTESİ**
**EĞİTİM BİLİMLERİ ENSTİTÜSÜ**

**EĞİTİM BİLİMLERİ ANABİLİM DALI**
**EĞİTİMDE ÖLÇME VE DEĞERLENDİRME PROGRAMI**

# AÇIK UÇLU MADDELERİN OTOMATİK PUANLANMASINDA DOĞAL DİL İŞLEME YÖNTEMİNDEN YARARLANILARAK MAKİNE ÖĞRENMESİ ALGORİTMALARININ KARŞILAŞTIRILMASI

**DOKTORA TEZİ**

**KÜBRA YILMAZ**

**DANIŞMAN: PROF.DR. KAAN ZÜLFİKAR DENİZ**

**ANKARA**
**ŞUBAT, 2025**

Ankara Üniversitesi Eğitim Bilimleri Enstitüsü Müdürlüğüne,

Kübra Yılmaz adlı öğrencinin hazırladığı "Açık Uçlu Maddelerin Otomatik Puanlanmasında Doğal Dil İşleme Yönteminden Yararlanılarak Makine Öğrenmesi Algoritmalarının Karşılaştırılması" başlıklı bu çalışma Eğitim Bilimleri Anabilim Dalı / Eğitimde Ölçme ve Değerlendirme Programı'nda jüri üyelerince oy birliği / oy çokluğu ile **Doktora Tezi** olarak kabul edilmiştir.

| | Jüri Üyeleri | İmza |
|---|---|---|
| Başkan | …… | …….. |
| Üye | …… | …….. |
| Üye | …… | …….. |
| Üye | …… | …….. |
| Üye | …… | …….. |

ONAY

Bu tez Ankara Üniversitesi Lisansüstü Eğitim Öğretim Yönetmeliği'nin ilgili maddeleri uyarınca, jüri üyeleri tarafından …/…/20... tarihinde, Enstitü Yönetim Kurulu tarafından ise …/…/20... tarihinde kabul edilmiştir.

……

Prof. Dr. Mehmet İkbal YETİŞİR

Eğitim Bilimleri Enstitüsü Müdürü

## ETİK İLKELERE UYGUNLUK BİLDİRİMİ

Tez içindeki bütün bilgileri akademik yazım kurallarına uygun biçimde raporlaştırdığımı ve bunları etik ilkelere (atıfta bulunulan tüm yapıtlara kaynaklarda yer verilmesi, tezde kullanılan bilgi ve belgelere resmi yollarla ulaşılması ve bunların aslı bozulmadan kullanılması vb.) uygun olarak elde ettiğimi ve sunduğumu bildiririm.

Kübra YILMAZ

# ÖZET

## AÇIK UÇLU MADDELERİN OTOMATİK PUANLANMASINDA DOĞAL DİL İŞLEME YÖNTEMİNDEN YARARLANILARAK MAKİNE ÖĞRENMESİ ALGORİTMALARININ KARŞILAŞTIRILMASI

YILMAZ, Kübra
Doktora Tezi, Eğitim Bilimleri Anabilim Dalı
Tez Danışmanı: Prof. Dr. Kaan Zülfikar DENİZ
Şubat, 2025, xii + 84 Sayfa

Bu tez çalışmasında makine öğrenmesi ve doğal dil işleme yöntemlerinden yararlanılarak Türkçe açık uçlu maddelerin otomatik puanlanmasında denetimli makine öğrenmesi modellerinden Naif Bayes (Naive Bayes-NB), Lojistik Regresyon (Logistic Regression-LR), Karar Ağacı (Decision Tree-DT), Rastgele Orman (Random Forest-RF), Gradyan Artırma (Gradient Boosting-GB) ile BERT (Bidirectional Encoder Representations from Transformers - Çift Yönlü Kodlayıcı Temsilleri Dönüştürücü) model karşılaştırılmıştır. Çalışma ABİDE (Akademik Becerilerin İzlenmesi ve Değerlendirilmesi) 2016 Türkçe açık uçlu maddeleri ve nihai puanlayıcılardan elde edilen puanlar üzerinde yürütülmüştür. Çalışmada 8.sınıf öğrencilerine ait cevaplar yer almaktadır. Çalışmanın makine öğrenmesi modelleri ile ilgili analizlerinde Python programlama dili ve Pycharm geliştirme ortamından yararlanılmıştır. BERT modeli için ise Google Colaboratory ortamı kullanılmıştır. Metin ön işleme adımlarında MintLemon Türkçe paketi, Pandas ve Scikit-Learn kütüphaneleri kullanılmıştır. Makine öğrenmesi ve BERT algoritmaları ile ayrı ayrı puanlama yapılmış ve makine insan uyumu Ağırlıklandırılmış Kappa (QWK) ile hesaplanmıştır. Buna göre en iyi uyumu BERT modelin sağladığı sonucuna ulaşılmıştır (0,72-0,93). Gradyan Artırma algoritmasının da BERT modele yakın değerler vererek oldukça iyi performans sergilediği görülmüştür (0,67-0,90). Naif Bayes ise tüm maddeler için en düşük performans sağlayan algoritma olmuştur (0,44-0,78).

**Anahtar Sözcükler:** Makine Öğrenmesi, BERT Modeli, Otomatik Puanlama, Doğal Dil İşleme, Eğitimde Yapay Zeka

## ABSTRACT

### COMPARISON OF MACHINE LEARNING ALGORITHMS USING NATURAL LANGUAGE PROCESSING METHOD FOR AUTOMATIC SCORING OF OPEN-ENDED QUESTIONS

YILMAZ, Kübra

Dissertation, Department of Educational Sciences

Supervisor: Prof. Dr. Kaan Zülfikar DENİZ

February, 2025, xii + 84 Pages

In this thesis, Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB) and BERT (Bidirectional Encoder Representations from Transformers) models were compared for automatic scoring of Turkish open-ended items by using machine learning and natural language processing methods. The study was conducted on ABIDE (Monitoring and Evaluation of Academic Skills) 2016 Turkish open-ended items and scores obtained from final raters. The study includes the answers of 8th grade students. Python programming language and Pycharm development environment were used in the analyses related to machine learning models. Google Colaboratory environment was used for the BERT model. MintLemon Turkish package, Pandas and Scikit-Learn libraries were used in text preprocessing steps. Machine learning and BERT algorithms were scored separately and machine-human agreement was calculated using Weighted Kappa (QWK). Accordingly, it was concluded that the BERT model provided the best fit (0.72-0.93). The Gradient Boosting algorithm was also found to perform very well, giving values close to the BERT model (0.67-0.90). Naive Bayes was the lowest performing algorithm for all items (0.44-0.78).



**Key Words:** Machine Learning, BERT Model, Automatic Scoring, Natural Language Processing, Artificial Intelligence in Education

# ÖNSÖZ

Bu tez Türkçe açık uçlu maddelerin otomatik puanlanması üzerine yapılan bir araştırmanın sonucunda ortaya çıkmıştır. Çalışmamın her aşamasında bana yol gösteren, bilgi ve tecrübelerini paylaşan kıymetli danışmanım Prof. Dr. Kaan Zülfikar DENİZ'e sonsuz teşekkür ederim.

Bu çalışmayı yürütürken desteklerini esirgemeyen değerli katkılarıyla çalışmayı zenginleştiren tez izleme kurulu üyeleri Sayın Prof. Dr. Nuri DOĞAN ve Sayın Doç. Dr. Seher YALÇIN hocalarıma teşekkürlerimi sunarım. Ayrıca tez jüri üyeleri Sayın Prof. Dr. Selahattin GELBAL ve Sayın Prof.Dr. Celal Deha DOĞAN'a yapıcı tutumları ve önemli katkıları için en içten teşekkürlerimi sunuyorum.

Tez çalışmam için gerekli veri talebimi karşılayan Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü Veri İzleme ve Değerlendirme Başkanlığı'na teşekkür ediyorum.

Tezimin hazırlanmasında bana manevi destek sağlayan ve motivasyonumu her zaman yüksek tutan sevgili eşim M. Bilal YILMAZ'a, sürecin her aşamasını benimle yaşayan sevgili oğluma, minicik kızıma, eğitim hayatım boyunca hep yanımda olan sevgili anneme, kardeşlerime tüm aileme en içten teşekkürlerimi sunarım. Bu süreç boyunca sabır ve hoşgörü gösteren tüm dostlarıma teşekkür ederim.

Son olarak, araştırmam süresince karşılaştığım zorlukların üstesinden gelmemde katkı sağlayan herkese bir kez daha teşekkür ederim.

*Aileme…*

# İÇİNDEKİLER

**Sayfa**

ETİK İLKELERE UYGUNLUK BİLDİRİMİ ... iii
ÖZET ... iv
ABSTRACT ... v
ÖNSÖZ ... vi
İÇİNDEKİLER ... viii
TABLOLAR DİZİNİ ... x
ŞEKİLLER DİZİNİ ... xi
KISALTMALAR/SİMGELER ... xii
BÖLÜM 1 ... 1
GİRİŞ ... 1
Problem ... 3
Amaç ... 5
Önem ... 6
Sınırlılıklar ... 7
Tanımlar ... 7
BÖLÜM 2 ... 9
KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR ... 9
Makine Öğrenmesi Algoritmaları ... 9
Denetimli Makine Öğrenmesi (Supervised Algorithms) ... 10
Denetimsiz Makine Öğrenmesi (Unsupervised Algorithms) ... 11
Yarı Denetimli Algoritmalar (Semi-Supervised Algorithms) ... 12
Takviyeli Makine Öğrenmesi (Reinforcement Machine Learning) ... 13
Çok Terimli Naif Bayes (Multinomial Naive Bayes) ... 14
Karar Ağacı (DecisionTree) ... 15
Rastgele Orman (Random Forest) ... 16
Gradyan Artırma (Gradient Boosting) ... 17
Lojistik Regresyon (Logistic Regression) ... 18
BERT (Bidirectional Encoder Representations from Transformers- Çift Yönlü Kodlayıcı Temsilleri) ... 18
Doğal Dil İşleme (DDİ-Natural Language Processing-NLP) ... 20
Türkçe Doğal Dil İşleme Yaklaşımları ve Karşılaşılan Zorluklar ... 25
BÖLÜM 3 ... 42
YÖNTEM ... 42
Araştırmanın Modeli ... 42
Çalışma Grubu ... 42
Verilerin Toplanması ... 42
Verilerin Çözümlenmesi ... 43

BÖLÜM 4 ..... 48
BULGULAR VE YORUMLAR ..... 48
Birinci Alt Amaca İlişkin Bulgular ve Yorumlar: TF-IDF ve BERT Vektörleştirme Yöntemleri Kullanılarak Açık Uçlu Maddeler Puanlandığında, Farklı Makine Öğrenmesi Modellerinin Doğruluk Oranları ..... 48
İkinci Alt Amaca İlişkin Bulgular ve Yorumlar : Klasik Makine Öğrenmesi Modelleri İle BERT Modeli, Sınıflandırma Doğruluk Metrikleri Açısından Karşılaştırılması ..... 54
Üçüncü Alt Amaca İlişkin Bulgular ve Yorumlar: TF-IDF Ve BERT Yöntemleriyle Vektörleştirilen Açık Uçlu Maddelerde İnsan ve Bilgisayar Puanlayıcıları Arasındaki QWK Uyumu ..... 57
BÖLÜM 5 ..... 61
SONUÇLAR VE ÖNERİLER ..... 61
Sonuçlar ..... 61
Öneriler ..... 63
Uygulamaya Yönelik Öneriler ..... 63
Araştırmacılara Yönelik Öneriler ..... 63
KAYNAKLAR ..... 65
EKLER ..... 75
EK 1. Etik Kurul Onayı ..... 76
EK 2. Millî Eğitim Bakanlığı Araştırma İzni ..... 77
EK 3. Uygulama ile İlgili Bilgiler ..... 78
EK 4. Kod Örnekleri ..... 80
EK 5. İnsan Puanlayıcı ve Makine Puanlayıcı Puanlama Örnekleri ..... 81
BENZERLİK BİLDİRİMİ ..... 82
ÖZGEÇMİŞ ..... 83

# TABLOLAR DİZİNİ

**Tablo** **Sayfa**

Tablo 1. Kappa Uyum Değer Aralığı ..... 45
Tablo 2. İkinci Maddeye İlişkin Makine Öğrenmesi Algoritma Performansları Bulguları ..... 48
Tablo 3. Yedinci Maddeye İlişkin Bulgular (3 Kategorili) ..... 50
Tablo 4. Sekizinci Maddeye İlişkin Bulgular ..... 51
Tablo 5. Onuncu Maddeye İlişkin Makine Öğrenmesi Algoritmaları Bulgular ..... 52
Tablo 6. On Birinci Maddeye İlişkin Bulgular ..... 53

# ŞEKİLLER DİZİNİ

**Şekil** **Sayfa**

Şekil 1. Denetimli Makine Öğrenmesi Sürecine İlişkin Görsel (Turhost,2020). ..... 11
Şekil 2. Denetimsiz Makine Öğrenmesi Sürecine İlişkin Görsel (Turhost,2020). ..... 12
Şekil 3. Yarı Denetimli Makine Öğrenmesi Sürecine İlişkin Görsel (Turhost,2020). ... 12
Şekil 4. Takviyeli Makine Öğrenmesi Sürecine İlişkin Görsel (Turhost, 2020). ..... 13
Şekil 5. Karar Ağacı Yapısının Şematik Diyagramı (Zhang ve Yuan, 2022). ..... 15
Şekil 6. Karar Ağacı Yapısı (Çelik ve Altunay,2018). ..... 16
Şekil 7. Eğitim Alanındaki Makine Öğrenmesi İle İlişkili Çalışmalarda Kullanılan Yöntemler (Alanoğlu, 2022). ..... 22
Şekil 8. İkinci Maddeye İlişkin Makine Öğrenmesi Modelleri ve BERT Model Karşılaştırması ..... 54
Şekil 9. Yedinci Maddeye İlişkin Makine Öğrenmesi Modelleri Ve BERT Model Karşılaştırması ..... 55
Şekil 10. Sekizinci Maddeye İlişkin Makine Öğrenmesi Modelleri ve BERT Model Karşılaştırması ..... 56
Şekil 11. Onuncu Maddeye İlişkin Makine Öğrenmesi Modelleri ve BERT Model Karşılaştırması ..... 56
Şekil 12. On Birinci Maddeye İlişkin Makine Öğrenmesi Modelleri ve BERT Model Karşılaştırması ..... 57
Şekil 13. İkinci Maddeye İlişkin QWK Model Performansları ..... 58
Şekil 14. Yedinci maddeye ilişkin QWK model performansları ..... 58
Şekil 15. Sekizinci Maddeye İlişkin QWK Model Performansları ..... 59
Şekil 16. Onuncu Maddeye İlişkin QWK Model Performansları ..... 59
Şekil 17. On Birinci Maddeye İlişkin QWK Model Performansları ..... 60

## KISALTMALAR/SİMGELER

| | |
|---|---|
| DDİ | Doğal Dil İşleme (Natural Language Processing-NLP) |
| AES | Otomatik Yazılı Değerlendirme Sistemi (Automated Essay Scoring) |
| NB | Naive Bayes (Naif Bayes) |
| LR | Logistic Regression (Lojistik Regresyon) |
| DT | Decision Tree (Karar Ağacı) |
| RF | Random Forest (Rastgele Orman) |
| GB | Gradient Boosting (Gradyan Artırma) |
| BERT | (Bidirectional Encoder Representations from Transformers- Çift Yönlü Kodlayıcı Temsilleri Dönüştürücü) |

# BÖLÜM 1

# GİRİŞ

Bu bölümde; problem, amaç, önem, sınırlılıklar ve çalışmada yer alan kavramların tanımları yer almaktadır.

Teknolojik gelişmelerin hız kazanmasıyla birlikte doğal dil işleme (DDİ), makine öğrenmesi, yapay zeka gibi kavramlar sık karşılaşılan kavramlar haline gelmiştir. Yapay zeka kavramı diğer kavramları kapsayan bir anlamda kullanılmaktadır. Bu teknolojik gelişmeler pek çok alanda dönüşümleri de beraberinde getirmiştir. Sağlık, hukuk, mimarlık ve eğitim vb. farklı disiplinlerde kullanım alanı bulmaya başlamıştır (Chaudhry & Kazim, 2021; Coelho, H. Silva, L. Silva, Andrade, & Rodrigues, 2023; Göloğlu Demir & Yılmaz, 2018; Li, 2023; Mulianingsih, Anwar, Shintasiwi, & Rahma, 2020; Ramachandran & Rana, 2024; Sytnyk & Podlinyayeva, 2024).

Yapay zeka bilgisayarların ve makinelerin insan karar alma ve verme süreçlerinin benzerini yapmayı sağlayan yazılımsal teknikleridir (Newell & Simon, 1956). Diğer bir tanımda yapay zeka akıllı makineler yapma bilimi ve mühendisliği olarak tanımlanmıştır (McCarthy, 2007). Yapay zeka, insan zekasını bilgisayar modelleri yardımıyla inceleyen ve bunları taklit ederek yapay sistemlere uygulamayı amaçlayan disiplinler arası bir araştırma alanıdır. Yapay zeka insan benzeri düşünme ve insan gibi davranma gibi hedeflere sahiptir (Kühl vd., 2019). Yapay zeka ile ilgili yapılan tanımlar, insan gibi düşünme, insan gibi davranma, akılcı düşünme, akılcı davranma şeklinde sınıflandırılabilmektedir (Russel & Norvig, 2010).

Algoritmaları eğitmenin makine öğrenmesi, sinir ağları, derin öğrenme gibi yolları bulunmaktadır (Kış, 2019). Eğitim setleriyle (training set) eğitilen yapay zeka yazılımının istatistiksel algoritmaları eğitim verilerini işlemekte ve hedef tahmine ulaştırmaktadır. Bu tahminler yapay zeka kullanıcısı için öneri olarak kullanılmaktadır. Algoritma sonucu üretilen bu tahmin (veya karar) mekanizması ile kendi karar alma mekanizması işleyen sistemler meydana gelmiştir (Kış, 2019).

Arthur Samuel tarafından 1959'da ortaya atılan bir terim olan makine öğrenmesi (Burkov, 2019) üç temelden oluşmaktadır. Bunlar veri, model ve öğrenmedir (Deisenroth vd., 2020). Veriler bilgisayar ortamına aktarıldıktan sonra bilgisayarın anlayabileceği

şekilde modeller oluşturulmaktadır. Bu modeller eğitim verileriyle eğitilip test verileriyle eğitilen model doğruluğu test edilmektedir. Makine öğrenmesi, verilerden model oluşturma sürecidir. Veri madenciliğinin temelini oluşturmaktadır (Leskovec, Rajaraman & Ullman 2020).

Yapay zekanın başka bir boyutu olan Doğal Dil İşleme (DDİ-Natural Language Processing-NLP), metin analizi yapmaya yönelik bir yaklaşımdır. Teorilere ve teknolojilere dayalı bir kavramdır. Ancak üzerinde anlaşmaya varılmış tek bir tanım bulunmamaktadır (Liddy, 2001).

Yapay zeka teknolojisinin gelişimi 1950'lere kadar dayanmasına rağmen eğitimde yapay zeka örnekleri sınırlı kalmıştır. Son yıllarda hem yurtdışı hem yurtiçi çalışmalarda eğitimde yapay zekanın kullanıldığı örnekler artış göstermeye başlamıştır. Yapay zeka teknolojileri eğitim kalitesini artırma ve uygun öğretim yöntemleri geliştirme noktasında potansiyele sahiptir (Sadiku, Ashaolu, Ajayi-Majebi & Musa, 2021). Günümüzde otomatik madde üretimi (Bezirhan & Davier, 2023; Shin, 2021), otomatik puanlama (Bulut, 2022; Gierl, Zhou & Alves, 2008; Gierl & Lai, 2016; Uysal, 2019) gibi çalışmalara ek olarak duygu analizi, öğrenciye geri bildirim (Yılmaz & Deniz, 2024) gibi konularda da yapay zeka teknolojileri kullanılabileceğine dair çalışmalar yapılmaktadır.

Eğitimde ölçme ve değerlendirme alanında madde yazımı ve madde puanlama süreçleri oldukça önemlidir. Bireyler çeşitli amaçlarla ulusal ve uluslararası pek çok sınava katılmaktadır. Geleneksel madde yazma süreci, bireysel test maddelerini manuel tasarlayan ve değerlendiren uzmanlar tarafından gerçekleştirilir. Maddelerin insanlar tarafında yazılması yoğun ve zahmetli bir çalışma gerektirmektedir (Alikaniotis, Yannakoudakis, & Rei, 2016). Soru yazarlarının metnin ana fikrini ve ana konuları tanımlamak için binlerce makaleyi incelemesi, okuması ve anlaması gerekmektedir. Karmaşık ve zaman alıcı olan bu görevin üstesinden gelebilmek için yapay zeka teknolojileri kullanılabilmektedir. Metinlerden ana fikir cümlesi ve metnin özeti gibi anlamsal özellikleri tanımlayıp kullanarak test maddeleri üretilebilmektedir (Shin, 2021). Otomatik madde üretimi, bilgisayar teknolojileri yardımıyla oluşturulan maddelerin üretimine rehberlik etmek için bilişsel ve psikometrik modelleme uygulamalarını birleştirmektedir (Gierl & Lai, 2016). Sınav sayısı arttıkça soru üretimi gereksinimi de artmaktadır. Bu sınavlar için metin soruları üretmek yapay zeka teknolojileriyle artık mümkün hale gelmiştir (Bezirhan & Davier, 2023).

Çok sayıda bireyin katılımı ile gerçekleşen bu sınavların puanlanması da ayrı bir gereksinimdir. Bu durum sınavların puanlanması noktasında da çeşitli bilgisayar

teknolojilerinin kullanımını gerekli kılmaktadır. Bu anlamda otomatik puanlama sistemlerinin geliştirilmesi önem arz etmektedir (Ahmed, Joorabchi & Hayes 2022; Fang, Lee & Zhai, 2023; Kakkonen, Myller, Timonen & Sutinen, 2005; Marinho, Anchieta, & Moura, 2021; Nael, Elmanyalawy & Sharaf, 2022; Sun & Wang, 2024; Davier, Khorramdel & Tyack, 2022; Yang, Raković, Li, Guan, Gašević & Chen, 2024). Bu tez çalışmasında Türkçe öğrenci cevaplarının otomatik puanlanması için kullanılan makine öğrenmesi algoritmalarının performansı karşılaştırılmıştır. Türk dilinde otomatik puanlama için uygun algoritmalar belirlenmiştir.

## Problem

Günümüzde bireyler çeşitli amaçlar için hayatları boyunca pek çok sınava girmektedir. KPSS (Kamu Personeli Seçme Sınavı), ALES (Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı), ABİDE (Akademik Becerilerin İzlenmesi ve Değerlendirilmesi) vb. ulusal sınavların yanı sıra PISA (Uluslararası Öğrenci Değerlendirme Programı-Programme for International Student Assessment), TIMSS (Uluslararası Matematik ve Fen Eğilimleri Araştırması - Trends in International Mathematics and Science Study) gibi geniş ölçekli testlere de birçok birey katılmaktadır. Dolayısıyla daha fazla aday, daha fazla sınav, daha fazla sınavsa değerlendirilmesi gereken birçok maddenin olduğu anlamına gelmektedir.

Testlerde sadece çoktan seçmeli maddelerin kullanılması öğrenme ve öğretme süreçlerini etkilemektedir. Bloom taksonomisinin güncel haline göre hatırlama, anlama, uygulama, analiz, değerlendirme ve yaratma olmak üzere altı farklı öğrenme düzeyi bulunmaktadır (Wilson, 2016). Öğrencilerin bilgiyi analiz etme ve sentezleme yeteneği Bloom Taksonomisine göre üst seviyelerde yer almaktadır (Munzenmaier & Rubin, 2013). Bu nedenle sınavlarda hem çoktan seçmeli hem açık uçlu maddelerin yer alması önemlidir. Çoktan seçmeli maddelerin puanlanması açık uçlu soruların puanlanmasına göre daha kolaydır. Çoktan seçmeli maddeler bilgisayar programları tarafından kısa sürede puanlanabilmektedir. Geniş ölçekli sınavların yaygınlaşmasıyla birlikte açık uçlu maddelerin de sınav süreçlerine daha fazla katılması söz konusu olabilmektedir. Çoktan seçmeli maddelere ilişkin puanlama kolaylığının açık uçlu maddeler için de sağlanması gerekliliği ortaya çıkmaktadır.

Açık uçlu maddelerde soru sayısı ve testi alan kişi sayısı arttıkça daha fazla puanlayıcıya ihtiyaç duyulabilmektedir. Bu durum puanlayıcılardan kaynaklı hatalar (puanlamada katılık, cömertlik vb.) puanlamanın adil yapılmasına ilişkin sorunları da beraberinde getirmektedir. Açık uçlu maddeleri uygulamak ve puanlamak uzun zaman ve emek gerektiren oldukça zahmetli bir görevdir. Özellikle ülke genelinde yapılan ve binlerce öğrencinin katılımıyla gerçekleşen sınavlarda puanlama yapmak oldukça zaman almaktadır. Bu noktada teknolojik gelişmelerden yararlanılarak doğal dil işleme teknikleri ile otomatik madde puanlaması sağlanabilirse hem puanlayıcıdan kaynaklı yanlılık gibi olası riskler de ortadan kaldırılmış olacak hem de zaman tasarrufu ve maliyet konularında avantaj sağlanmış olacaktır. Bu noktada dikkat edilmesi gereken husus, bu sistemler geliştirilirken gerçek puanlayıcıların puanlama noktasında iyi eğitilmiş olmasıdır. Çünkü geliştirilen sistemler gerçek puanlayıcılardan elde edilen puanlar ile eğitilmektedir.

Bu çalışmada Türk dilinde doğal dil işleme yoluyla otomatik puanlama sisteminin geliştirilebilmesi için uygun algoritmalar karşılaştırılmıştır. Doğal dilin işlenmesi noktasında dilin yapısal özellikleri önemlidir. Türk dili Altay dil ailesinde yer almaktadır. Altay Dilleri, Türkçe, Mançu-Tunguzca, Moğolca, Japonca ve Korece'den meydana gelen bir dil grubudur (Akar, 2017). Altmış milyondan fazla kişinin anadili olmasına rağmen Türkçe, doğal dil işleme çalışmalarında nispeten geç kalmış ve Türkçe'nin doğal dil işlemesine yönelik araç ve kaynakların geliştirilmesi ancak son yirmi yılda denenmiştir (Oflazer ve Saraçlar, 2018). Otomatik puanlama noktasında ülkemizde az sayıda çalışma bulunmaktadır. Açık uçlu maddelerin geniş uygulama alanlarında kullanılması, bu noktada bireysel iş gücünün azaltılması ve gelişen teknolojinin eğitime entegrasyonunun sağlanması, dilin yapısal özelliklerine özgü sistemlerin geliştirilmesi önemli görülmektedir. Tüm bu sebepler doğrultusunda Türkçe açık uçlu sorular için otomatik puanlamaya ihtiyaç duyulmaktadır.

Açık uçlu maddelerin puanlanmasında TF-IDF (Term Frequency/ Inverse Document Frequency- Terim ve Ters Terim Frekansı) ve BERT (Bidirectional Encoder Representations from Transformers- Çift Yönlü Kodlayıcı Temsilleri Dönüştürücü) vektörleştirme yöntemleri kullanıldığında, klasik makine öğrenmesi modelleri ile BERT modelinin doğruluk (accuracy, precision, recall, F1-score) ve uyum (Ağırlıklandırılmış Kappa - QWK) açısından nasıl bir performans gösterdiği ve hangi yöntemlerin puanlama noktasında daha uygun olduğu belirsizdir. Bu bağlamda bu tez çalışmasında açık uçlu maddelerin otomatik puanlanmasında doğal dil işleme yönteminden yararlanılarak

makine öğrenmesi algoritmaları karşılaştırılarak incelenen makine öğrenmesi ve BERT algoritmaları arasında Türk diline en uygun makine öğrenmesi algoritmaları belirlenmiştir.

## Amaç

Bu çalışmanın amacı; Türkçe doğal dil işleme yöntemlerinden yararlanılarak Türk dilinde doğal dil işleme yardımıyla otomatik puanlama için uygun algoritmaların keşfedilmesi ve otomatik puanlanın ne derece doğru yapıldığının ortaya konulmasıdır. Ayrıca bu tez çalışmasında makine-insan puanları arası uyum puanlarıyla hangi algoritmanın insan benzeri puanlamalar yaparak otomatik puanlama sistemlerinde kullanılabileceği araştırılması amaçlanmaktadır.

Doğal dil işleme yöntemleri kullanılarak ABİDE sınavı açık uçlu maddelerin otomatik puanlanmasında çeşitli makine öğrenmesi tekniklerinden yararlanılacaktır. Türk dili üzerinde Naif Bayes, Lojistik Regresyon, Karar Ağacı, Rastgele Orman, Gradyan Artırma ve BERT model yöntemlerinden elde edilen sonuçların karşılaştırılması, bu modellerin nasıl bir performans gösterdiği ve hangi yöntemlerin puanlama noktasında daha uygun olduğunun belirlenmesi amaçlanmaktadır.

Araştırma kapsamında belirlenen alt amaçlar şu şekildedir:

1. TF-IDF ve BERT vektörleştirme yöntemleri kullanılarak açık uçlu maddeler puanlandığında, farklı makine öğrenmesi modellerinin doğruluk oranları nasıldır?
2. Klasik makine öğrenmesi modelleri ile BERT modeli, sınıflandırma doğruluk metrikleri açısından nasıl karşılaştırılmaktadır?
3. TF-IDF ve BERT yöntemleriyle vektörleştirilen açık uçlu maddelerde insan ve bilgisayar puanlayıcıları arasındaki uyum Ağırlıklandırılmış Kappa katsayısı (QWK) açısından nasıldır?

Bu amaç ve alt amaçlar kapsamında incelenen makine öğrenmesi ve BERT modellerinin öğrenci cevabı ile verilen puanı ne derece doğru sınıflandığı ve otomatik puanlama noktasında hangisinin daha iyi sonuçlar verdiği ortaya konulmuş olacaktır.

## Önem

Teknolojik gelişmelerle birlikte eğitim alanında da daha fazla verinin toplanması sonucunda, yapay zeka, makine öğrenmesi, doğal dil işleme yöntemleri eğitim alanlarında da kullanım alanı bulmaktadır. Örneğin; cümle benzerliği, öğrenciye geri bildirim, duygu analizi, soru üretimi gibi alanlarda bu teknolojilerden yararlanılmaktadır (Yılmaz & Deniz, 2024). Eğitimde ölçme ve değerlendirme alanında önemli bir diğer konu da sınavların otomatik puanlanmasıdır. Otomatik puanlama fikri Page (1966) tarafından ortaya atılmıştır. Otomatik puanlamada, yazılı bir metin bilgisayar algoritmaları yardımıyla değerlendirilebilmektedir (Shermis, 2010). Otomatik puanlama sistemleri hem kısa cevap parçaları hem de uzun denemeler gibi farklı cevaplar üzerinde çalışmaya imkan sağlamaktadır (Gierl, Latifi, Lai, Boulais, & Champlain, 2014).

Gardner, O'Leary ve Yuan (2021) yapay zeka tabanlı puanlama sistemlerinin önemli bir potansiyele sahip olduğunu vurgulamaktadır. Bununla birlikte geliştirilmesi gereken noktaların olduğunu da ifade etmektedir. Otomatik puanlama sistemleri öğrencilere kişiselleştirilmiş geri bildirimler sunma ve yazma süreçlerinin geliştirme gibi katkılar sağlayabilecektir. Bu sistemlerin gelecekte daha doğru puanlamalar sunabileceği de öngörülmektedir.

Otomatik puanlamayı etkileyebilecek bir faktör dil yapılarındaki farklardır. Otomatik puanlama yaklaşımlarının Türkçe dışındaki dillerde gerçekleştirildiği görülmektedir (Ishioka & Kameda, 2006; Jang, Kang, Noh, Kim, Sung & Seong, 2014; Kakkonen vd., 2005; Krishnamurthy,Gayakwad & Kailasanathan, 2019; Nael vd., 2022). Ancak Türkçe otomatik puanlama çalışmalarının sınırlı sayıda olduğu görülmektedir. Bu nedenle, Türk dilinde otomatik puanlama araştırılmalıdır (Uysal ve Doğan, 2021).

Açık uçlu sorular bireylerin üst düzey düşünme becerilerinin ölçülmesinde önemli rol oynamaktadır. Puanlama güçlüğü ortadan kaldırılabilirse açık uçlu sorular geniş kitlelere uygulanabilir (Çınar, İnce, Gezer & Yılmaz, 2020; Jang vd., 2014; Davier vd., 2022). Milli Eğitim Bakanlığı'nın (MEB) ülke geneli ortak sınavlarda, öğrencilerin cevapları kendi cümleleriyle yazacakları açık uçlu ve kısa cevaplı sorularla yapması ve bu sınavların değerlendirilmesine de katkı sağlayabileceği ön görülmektedir. Bu çalışmada Doğal Dil İşleme ve Makine Öğrenmesi tekniklerinin eğitime entegrasyonunun yapılması, Doğal Dil İşleme, Makine Öğrenmesi ve BERT tekniklerinden yararlanılarak açık uçlu sınavların otomatik puanlanması sağlanarak zaman, insan gücü, yanlılık riskleri gibi olası sorunlara çözüm üretilmesi hedeflenmektedir. Bu tez çalışması farklı metin

vektörleştirme yaklaşımlarının Türkçe açık uçlu sorularda kullanılması ve bilgisayar-insan puanlayıcı arasındaki uyumun incelenmesi ve metinlerin otomatik puanlamasının işlevselliğini ortaya koyması bakımından önemli görülmektedir. Ayrıca araştırmacı tarafından geliştirilen mevcut makine öğrenmesi algoritmalarını karşılaştırma sisteminin de veriye özel algoritma seçimi noktasında yararlı olacağı ön görülmektedir.

## Sınırlılıklar

1. Bu çalışma ABİDE sınavının açık uçlu sorulara verilen Türkçe cevapların TF-IDF ve BERT model vektörleştirme yöntemleri ile sınırlıdır.
2. Çalışma NB, LR, DT, RF, GB makine öğrenmesi algoritmaları ve Bert modeli ile sınırlıdır.
3. Çalışma öğrencilerin yanıtladığı 5 madde ile sınırlıdır.
4. Model eğitim süresinin uzunluğu ve işlemci kapasitesinin yetersizliği sebebiyle BERT modeli için 5 kat çapraz doğrulama yapılmıştır.
5. Bu çalışmada, zaman ve kaynak kısıtlamaları nedeniyle hiperparametre optimizasyonu yapılmamıştır. Modellerin varsayılan hiperparametre değerleri ile çalışılmış ve performanslar bu şekilde değerlendirilmiştir.

## Tanımlar

**Makine Öğrenmesi:** Bu çalışmada makine öğrenmesi, Türkçe metin sınıflandırma modellerini eğitmek için kullanılan denetimli öğrenme algoritmaları olarak ele alınmıştır.

**Doğal Dil İşleme:** Bu çalışmada Doğal Dil İşleme, Türkçe metinlerin ön işlenmesi, vektörleştirilmesi ve sınıflandırılması süreçlerini kapsayan bir dizi teknik ve algoritma bütünü olarak ele alınmaktadır.

**Ağırlıklandırılmış Kappa (QWK):** Bu çalışmada QWK, otomatik puanlama sisteminin insan değerlendiricilerle ne kadar uyumlu olduğunu ölçmek için kullanılmaktadır.

**Doğruluk:** Modelin doğru sınıflandırdığı örneklerin toplam örnek sayısına oranı olarak hesaplanmaktadır.

**Kesinlik:** Modelin pozitif olarak tahmin ettiği örneklerin gerçekten pozitif olma oranıdır.

**Hassasiyet/Duyarlılık:** Bu çalışmada, hassasiyet/duyarlılık değeri, modelin doğru sınıflandırdığı pozitif örneklerin, toplam gerçek pozitif örneklere oranı olarak hesaplanmıştır**.**

**F1-Skoru:** Bu çalışmada F1-Skoru, modelin kesinlik (precision) ve duyarlılık (recall) değerlerini dengeli bir şekilde ölçmek için kullanılmaktadır**.**

# BÖLÜM 2

# KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

Bu bölümde makine öğrenmesi algoritmaları, doğal dil işleme, Türkçe doğal dil işleme yaklaşımları ve karşılaşılan zorluklara yer verilmiştir.

## Makine Öğrenmesi Algoritmaları

Yapay zeka ve makine öğrenmesinin temel taşları ilk olarak Turing (1936) tarafından hesaplanabilirlik teorisi ile oluşmaya başlamıştır. Turing (1936) makalesinde hesaplanabilir olan her türlü fonksiyonun bir makine tarafından da yapılabileceğini ortaya atmıştır. Daha sonra McCulloch ve Pitts (1943) sinir ağlarını gündeme getirmiş ve insan beynine benzer sinirsel yapılarla problem çözme becerilerinin bilgisayar tarafından yapılabilirliğini araştırmıştır. Ancak dönemin şartlarında veri işleme yetersizliği sebebiyle bir süre sinir ağları kavramı gölgede kalmıştır.

Turing (1950)'in ortaya attığı "Makineler düşünebilir mi?" sorusu Dartmouth Konferansı'nda McCarthy, Minsky, Rochester ve Shannon (1955) tarafından yanıtlanarak Yapay zeka kavramı yazılı bir çerçeve olarak ele alınmıştır. Yapay zeka kavramı genellikle algoritmalarla eş değer tutulmuştur (Sheikh, Prins & Schrijvers, 2023). Yapay zeka akıllı makineler oluşturma sanatı olarak tanımlanmıştır (McCarthy, 2007).

Newell ve Simon (1956) tarafından "Logic Theorist'" adında ilk yapay zeka programı geliştirilmiştir. Rosenblatt (1958) geliştirdiği "Perceptron" modeli ile yapay sinir ağı modellerinin gelişmesine katkı sağlamıştır. Samuel (1959) dama oynayan bir bilgisayar geliştirilerek insan rakibine karşı oldukça başarılı olduğunu kanıtlanmıştır. Böylece makine öğrenmesi kavramı bu çalışma ile 1959'da Arthur Samuel tarafından ortaya atılmıştır (Burkov, 2019). 1966'da Weizenbaum tarafından "Eliza" adında doğal dil işleme tabanlı bir program geliştirilmiştir. Çevresinin algılayabilen ilk mobil robot olan "Shakey" geliştirilmiştir (Nilsson, 1984). "Yapay Zeka Kışı" olarak adlandırılan dönem 1970-1980 yılları arası gerçekleşmiştir. Bu dönemde yayınlanan Lighthill (1973)

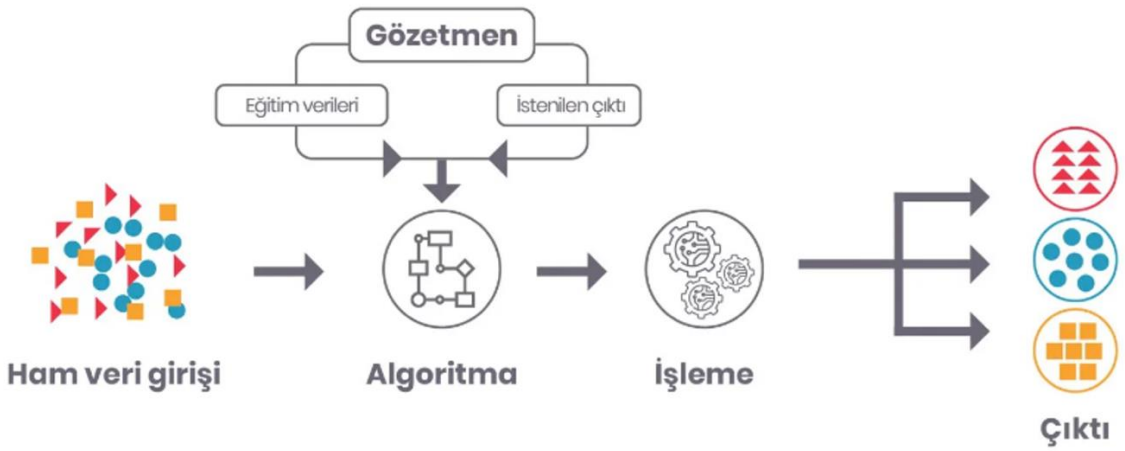
raporu yapay zeka yatırımlarını azaltmış, veri ve teknoloji yetersizliği ise yapay zeka kışının yaşanmasının diğer sebepleri arasında sıralanmıştır (Hendler, 2008).

Geliştirilen uzman sistemlerle 1980 sonrası yapay zekaya olan ilgi tekrar artmıştır (Buchanan & Shortliffe, 1984). Artan veri miktarıyla birlikte 1990 sonrasında makine öğrenmesi kavramı önem kazanmıştır. 2010 yıllarına gelindiğinde ise daha karmaşık veri setleriyle işlemler yapılabilir hale gelmiştir (LeCun, Bengio, & Hinton, 2015). Krizhevsky, Sutskever ve Hinton (2012) görüntü işleme noktasında geliştirdikleri model ile büyük başarı elde etmiştir.

Veri miktarının artmasıyla birlikte verilerin analizi için otomatik sistemlere ihtiyaç duyulmaktadır (Murphy, 2012). Makine öğrenmesi algoritmaları büyük verilerden anlamlı çıkarımlar yapma ve keşif noktasında kullanılmaktadır (Saraswathi, Balu, Sri Ram & Prakash 2024). Makine öğrenmesinde bir problemin çözümü için iki adım bulunmaktadır. 1) veri toplamak 2) veri setine dayalı bir istatistiksel model geliştirmektir (Burkov, 2019). Deisenroth, Faisal ve Ong (2020) ise bu süreci 1) veri, 2) model ve 3) öğrenme olarak tanımlamıştır. Makine öğrenmesi algoritmaları dört kategoriden oluşmaktadır. Burkov (2019)'a göre bunlar, denetimli makine öğrenmesi (supervised algorithms), denetimsiz makine öğrenmesi (unsupervised algorithms), yarı denetimli algoritmalar (semi-supervised algorithms), takviyeli makine öğrenmesi algoritmaları (reinforcement algorithms).

## Denetimli Makine Öğrenmesi (Supervised Algorithms)

Denetimli makine öğrenmesi algoritmaları, geliştiricinin denetimini gerektiren algoritmalardır. Yani etiketlenmiş verilerle eğitilmektedir. Bu algoritmada amaç, bağımsız değişkenler üzerinden hedef değişkeni tahmin etmektir (Burkov, 2019; Mahesh, 2018).

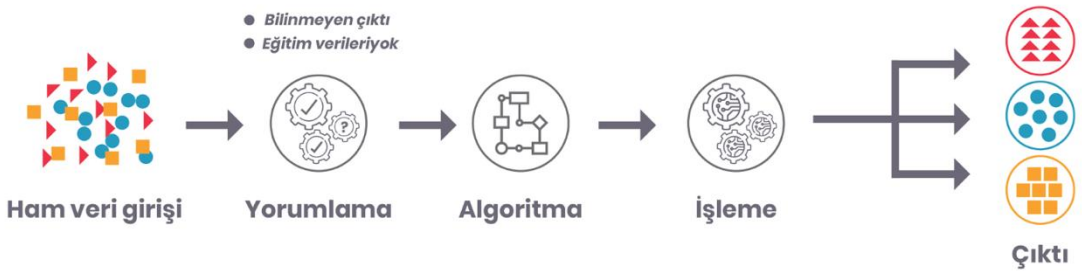



*Şekil 1*. Denetimli Makine Öğrenmesi Sürecine İlişkin Görsel (Turhost,2020).

Şekil 1 incelendiğinde denetimli makine öğrenmesi süreci görülmektedir (Turhost, 2020). Sınıflama ve regresyon denetimli makine öğrenmesi yaklaşımlarındandır (Murphy, 2012). Doğrusal Regresyon (Linear Regression), Lojistik Regresyon (Logistic Regression), Karar Ağaçları (Decision Trees), Destek Vektör Makineleri (Support Vector Machines, SVM), K-En Yakın Komşu (K-Nearest Neighbors, KNN) ve Naive Bayes algoritmaları denetimli makine öğrenmesi algoritmalarına örnektir (Mahesh, 2018).

### Denetimsiz Makine Öğrenmesi (Unsupervised Algorithms)

Denetimsiz Makine Öğrenmesi algoritmaları; eğitim verilerinin sınıflandırılmadığı veya etiketlenmediği hallerde kullanılmaktadır. Denetimli öğrenmede bir hedef varken, denetimsiz öğrenmede bir hedef yoktur. Denetimsiz öğrenmede etiketlenmemiş bir veri setinde bulunan gizli desen ve yapılar keşfedilmektedir (Watson, 2023). Şekil 2'de denetimsiz makine öğrenmesi sürecine ilişkin görselde bu algoritmanın çalışma yapısı gösterilmiştir (Turhost, 2020).




*Şekil 2.* Denetimsiz Makine Öğrenmesi Sürecine İlişkin Görsel (Turhost,2020).

Denetimsiz öğrenme insanların ve hayvanların öğrenme biçimlerine benzer olduğu ifade edilen bir türdür (Murphy, 2012). Analizler sonucunda bir çıkarıma ulaşılmaktadır (Mahesh, 2018). Örneğin; kümeleme (clutering) denetimsiz öğrenme algoritmalarındandır (Dridi, 2021; Watson, 2023). Ayrıca boyut indirgeme (dimensionality reduction) de denetimsiz makine öğrenmesine örnektir (Murphy, 2012).

### Yarı Denetimli Algoritmalar (Semi-Supervised Algorithms)

Yarı denetimli makine öğrenmesinde eğitim için etiketli ve etiketlenmemiş veriler kullanılmaktadır (Chapelle, Schölkopf & Zien, 2006). Etiketlenmemiş veriler etiketli verilere göre daha kolay toplanabilir ancak veri kullanımı noktasında sınırlılıklar mevcuttur. Etiketli verilere ulaşmak genellikle daha zor ve maliyetlidir (Engelen & Hoos, 2020).

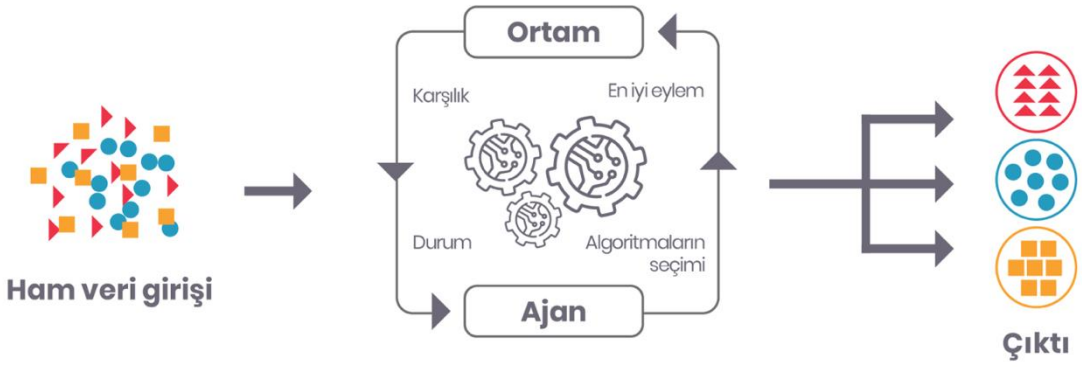

*Şekil 3.* Yarı Denetimli Makine Öğrenmesi Sürecine İlişkin Görsel (Turhost,2020).

Şekil 3'te yarı denetimli makine öğrenmesi sürecine ilişkin görselde bu algoritmanın çalışma yapısı gösterilmiştir (Turhost, 2020). Yarı denetimli makine

öğrenmesi algoritmaları etiketli ve etiketlenmemiş verilerin kullanımına imkan sağlayarak etkin sınıflandırıcılar geliştirmeye imkan sağlayan bir yaklaşımdır (Pise & Kulkarni, 2008). Yarı denetimli öğrenme etiketleme maliyetini azaltmakta, daha iyi genelleme yapmaya imkan sağlamakta ve özellikle doğal dil işleme, görüntü işleme gibi alanlarda etiketleme maliyeti çok yüksek olduğundan avantaj sağlamaktadır (Engelen & Hoos, 2020).

### Takviyeli Makine Öğrenmesi (Reinforcement Machine Learning)

Takviyeli makine öğrenmesi keşfederek öğrenmektedir. Bu Makine Öğrenmesi algoritması çevresi ile etkileşime girerek sonuçları gözlemlemektedir. Daha sonra bir sonraki adımda eylemi gerçekleştirirken bu gözlemlediği sonuçları dikkate almaktadır. Bu durum algoritma gelişip doğru stratejiyi seçene kadar tekrarlanır (Mahesh, 2018). Şekil 4'te bu algoritmanın çalışma yapısı gösterilmiştir (Turhost, 2020).




*Şekil 4.* Takviyeli Makine Öğrenmesi Sürecine İlişkin Görsel (Turhost, 2020).

Açık uçlu maddelerin puanlanmasında derin öğrenme yöntemlerinde insan beyninin işleme yeteneklerini simüle etmek amacıyla yapay sinir ağları kullanılmaktadır. Yapay sinir ağları dışında otomatik puanlamada denetimli makine öğrenmesine dayalı algoritmalardan yararlanılmaktadır (Gierl vd., 2014). Bu tez çalışmasında etiketli veriler ile model eğitimi yapıldığından denetimli makine öğrenmesi algoritmaları (NB, LR, DT, RF, GB ) ve BERT kullanılmıştır.

**Çok Terimli Naif Bayes (Multinomial Naive Bayes)**

Naif Bayes, öğrenme sürecinde, eğitim verileri kullanılarak sınıf olasılıkları ve koşullu olasılıklar hesaplar ve daha sonra bu olasılıkların değerleri yeni gözlemleri sınıflandırmak için kullanır. Önceki bilgilere dayalı olarak sonraki bilgiyi elde etmeye yarayan bir olasılıksal modeldir (Mahesh, 2018). Makine öğrenimi modelleri genellikle büyük miktarda eğitim verisine dayanmaktadır. Ancak Naif Bayes eğitim verisi küçük olduğunda da iyi bir performans gösterebilmektedir. Bu da Naif Bayes algoritmasının metin sınıflandırmada yaygın olarak kullanılmasına sebep olmaktadır (Kilimci & Ganiz, 2015).

Naif Bayes, basit, kullanışlı, yorumlaması kolay bir algoritmadır. Naif Bayes küçük veri setlerinde kullanılabilmektedir. Ayrıca hem kategorik hem sürekli veriye uygulanabilmesi gibi avantajlara sahip olması tercih sebebi olabilmektedir. Ancak kategorik verilerde, sınıf sınırlarının belirsiz olduğu durumlarda gücü kestirilememektedir. Bu da Naif Bayes'in dezavantajı olarak görülmektedir (Hamalainen & Vinni, 2011). Çok Terimli Naif Bayes doğal dil işleme çalışmalarında da sıklıkla kullanılan bir algoritmadır. Bu algoritma belge sınıflandırma, hastalık tahmini, spam filtreleme, duygu analizi gibi pek çok alanda metin sınıflandırma amacıyla kullanılmaktadır (Abbas, Memon, Jamali, Memon, & Ahmed, 2019). Çok Terimli Naif Bayes formülü şu şekildedir (McCallum & Nigam, 1998):

$$P(w_t \mid c_j) = \frac{\sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is} P(c_j \mid d_i) + |V|}{\sum_{i=1}^{|D|} N_{it} P(c_j \mid d_i) + 1}$$

$P(w_t \mid c_j)$: Belirli bir sınıf ($c_j$) verildiğinde, kelime ($w_t$)'nin görülme olasılığı.

|V|: Kelime dağarcığının (vocabulary) büyüklüğü.

|D|: Belge sayısı.

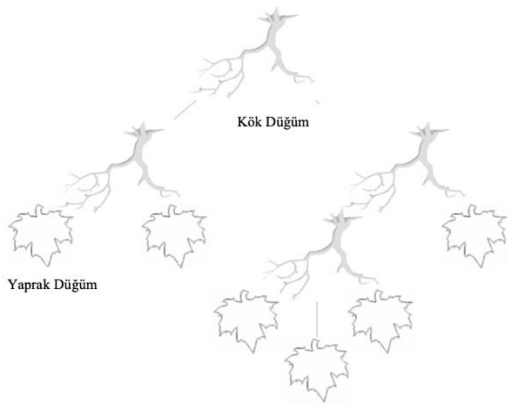
$N_{is}$: $d_i$ belgesindeki ($w_s$) kelimesinin sayısı.

($P(c_j \mid d_i)$): ($d_i$): belgesinde ($c_j$) sayısı

**Karar Ağacı (DecisionTree)**

Karar ağacı, seçimleri ve sonuçlarını bir ağaç şeklinde temsil eden bir grafiktir. Grafikte yer alan düğümler bir seçimi temsil etmektedir. Grafiğin kenarları ise kararı temsil etmektedir. Her bir ağaç düğüm ve dallardan oluşmaktadır. Bu düğümler, sınıflandırılacak gruptaki özellikleri temsil etmektedir (Mahesh, 2018).

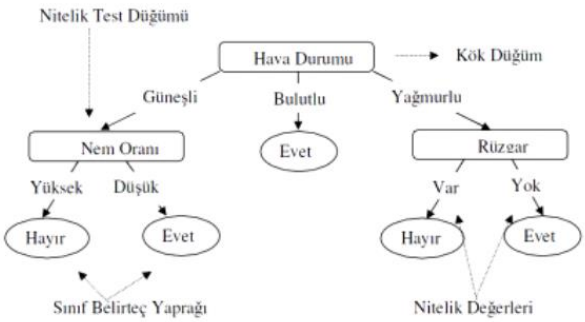
Karar ağacı yapıları 3 temel bileşenden oluşmaktadır. Bunlar: kök düğümler, iç düğüm ve yaprak düğümler. Kök düğümler bağımlı değişkeni gösteren ve sınıflandırma için hangi değişkenin kullanılacağını gösteren kısımdır. İç düğüm, bir gelen dal, iki veya daha fazla çıkan dala sahip olan düğümlerdir. Yaprak düğümler ise gelen bir dala sahip düğümlerdir. Yaprak düğümlerin başka çıkan dalı bulunmamaktadır (Çelik ve Altunaydın, 2018).




*Şekil 5.* Karar Ağacı Yapısının Şematik Diyagramı (Zhang ve Yuan, 2022).

Karar ağacı, ağaç dalları oluşturur ve her dal bir evet-hayır ifadesi gibi düşünülebilir. Kök düğüm başlangıç noktasını ifade etmektedir. Dallar, veri kümesini özelliklere göre alt kümelere bölmektedir. Son sınıflandırma ise, karar ağacının yapraklarında gerçekleşmektedir (Zhang & Yuan, 2022).

Karar ağacı algoritmasında önce bir ağaç oluşturulmakta, bir durdurma kriteri belirlenmekte ve ardından gereksiz dallar budama işlemiyle kaldırılmaktadır (Bishop, 2006).




*Şekil 6.* Karar Ağacı Yapısı (Çelik ve Altunay,2018).

Yu ve Zheng (2023)'e göre karar ağacında en uygun bölme için kullanılan indeks Gini indeksi formülü şu şekildedir:

1. Gini İndeksi:

$$gini(\mathcal{C}) = 1 - \sum_{k=1}^{K} p_k^2$$

gini$(\mathcal{C})$, mevcut düğümdeki Gini indeksini ifade eder.
$p_k$ ise $k$ sınıfına ait olma olasılığıdır.
$K$ ise sınıf sayıdır.

2. Bölme Kriteri:

$$Gix(\mathcal{C}, \mathcal{X}) = gini(\mathcal{C}|\mathcal{X}) - gini(\mathcal{X})$$

$Gix(\mathcal{C}, \mathcal{X})$, niteliğe göre veri setinin ne kadar iyi bölündüğünü ifade eder.
$gini(\mathcal{C}|\mathcal{X})$, bölme sonrası düğümdeki Gini indeksidir.
$gini(\mathcal{X})$, bölme yapılmadan önce mevcut düğümdeki Gini indeksidir.

**Rastgele Orman (Random Forest)**

Çok sayıda tahminciden elde edilen sonuçları bir araya getiren, topluluk yöntemi türüdür. Breiman (2001) tarafından ortaya Rastgele Orman algoritması, karar ağaçlarının

bir birleşimidir. Rasgele orman algoritması için marjin fonksiyonu formülü aşağıda verilmiştir (Breiman, 2001):

$$mg(X,Y) = \frac{1}{K}\sum_{k=1}^{K} I(h_k(X) = Y) - max_{j \neq Y} \ \frac{1}{K}\sum_{k=1}^{K} I(h_k(X) = j$$

$mg(X,Y)$: Margin fonksiyonu.

$I$: Gösterge fonksiyonu (1, eğer ifade doğru; 0, aksi takdirde).

$hk(X)$: $k$ numaralı ağacın tahmin sonucu.

$Y$: Gerçek sınıf etiketi.

$j$: Sınıf endeksi.

Rastgele orman, çoklu karar ağaçlarını kullanarak sınıflandırma ve regresyon yapmaktadır. Her ağaç, rastgele seçilen veri setleri ve özellikler kullanarak eğitilmekte ve bu ağaçlar birlikte tahminler yapmaktadır (Yu & Zheng, 2023).

**Gradyan Artırma (Gradient Boosting)**

Gradyan Artırma sınıflandırıcısı zayıf tahmincilerin birleşmesiyle oluşan güçlü tahminleri ortaya koyan bir makine öğrenmesi algoritmasıdır. Gradyan destekli ağaçlar makine öğrenimindeki en çok kullanılan tekniklerden biridir (Natekin & Knoll, 2013). Gradyan Artırma formülü aşağıda verilmiştir:

$$F_m(x) = F_{m-1}(x) + y \cdot h_m(x)$$

$F_m(x)$: m adımında yer alan toplam tahmin

$F_{m-1}(x)$: (m-1) adımındaki toplam tahmin

$y$: Her yeni tahminin katsayısını düzenleyen sabit

$h_m(x)$: m adımındaki yeni tahmincidir.

Gradyan Artırma sınıflandırıcılar birçok gerçek durum verileriyle denenmiş, doğruluk ve genelleme açısından oldukça başarılı sonuçlar elde edilmesine olanak sağlamıştır (Natekin & Knoll, 2013).

**Lojistik Regresyon (Logistic Regression)**

Lojistik Regresyon diğer analizlere göre daha esnek bir yapıda olması sebebiyle eğitim araştırmalarında sıklıkla kullanılmaktadır. Lojistik Regresyonun amacı, grup üyeliklerinin doğru tahmin edilmesine dayanmaktadır. Yordanan ve yordayıcı değişkenler arasında ilişki oluşturulmaktadır. Değişken setleriyle mümkün olan en az sayıdaki değişkenle bir model oluşturulur. Bu model yardımıyla yeni bireylerin de üyelik tahminleri yapılmış olmaktadır (Tabachnick & Fidell, 2013). Sosyal bilim araştırmalarında genellikle kesikli (süreksiz) veriler kullanılmaktadır. Bu gibi durumlarda lojistik regresyon analizinden yararlanılmaktadır (Deniz, 2021).

Bu model, verilen bir girdinin belirli bir sınıfa ait olma olasılığını tahmin etmektedir. Lojistik regresyon, girdilerin doğrusal kombinasyonunu almakta ve sigmoid (logistik) fonksiyon ile olasılık tahminleri üretmektedir. Lojistik Regresyonun olasılık fonksiyonu şu şekildedir (Bishop, 2006):

$$p(t|w) = \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n}$$

$t_n$, $n$- inci veri noktası için hedef sınıf

ıy$_n$, $= p(C_1|\phi_n)'$dir.

Manning ve Schütze (1999), Lojistik Regresyonun metin sınıflandırmada genellikle uygun bir temel model olarak tercih edildiğini ifade etmiştir.

**BERT (Bidirectional Encoder Representations from Transformers- Çift Yönlü Kodlayıcı Temsilleri)**

Geleneksel dil modelleri tek yönlü bir yaklaşım sunmaktadır. Yani geleneksel dil modelleri sadece sağdan sola veya soldan sağa girişleri inceleyebilmektedir. BERT modeli ise iki yönlü olarak metni anlama özelliğine sahiptir. Model sol ve sağ bağlamı birleştirebilir. Bu birleştirme işlemini yapabilmek için BERT, metinde yer alan bazı kelimeleri maskeler ve kelimelerin özgün kimliklerini tahmin eder. Yani kullanılan dilden bağımsız, bir anlam çıkarımı yaparak işlem yapabilmektedir (Devlin, Chang, Lee & Toutanova, 2018).

BERT modeli bir transformer (dönüştürücü) mimarisine dayanmaktadır. Bu model iki ayrı dildeki kelimelerin birbiriyle olan ilişkisini anlamak için kullanılmaktadır. Bu modelde kodlayıcı ve çözücü (encoder-decoder) yapısı kullanılmaktadır (Vaswani vd., 2017). Örneğin kodlayıcı "The book is on the table." cümlesindeki İngilizce sembolleri anlamlandırıp bir vektöre dönüştürürken, çözücü kodlayıcıdan aldığı vektörü çözümleyerek "Kitap masanın üstünde" cümlesini elde etmektedir.

Tranformer mimarisinde dikkat mekanizması yer almaktadır. Bu dikkat mekanizması cümlede yer alan kelimelerin birbiri ile olan ilişkisine odaklanmaktadır. Bu durum şu şekilde açıklanabilir. Örneğin Türkçe bir metinde geçen bir kelimenin anlamı sorulduğunda cümlenin tamamına bakmak, kelimenin o cümlede ne anlama geldiği anlama konusunda bir fikre sahip olmamızı sağlamaktadır. Benzer şekilde dikkat mekanizması da cümledeki diğer kelimelere de bakarak kelimenin anlamını kavramaya çalışmaktadır. Transformer modeli geliştirilirken İngilizce-Almanca ve İngilizce-Fransızca dilleri üzerinde denemeler yapılmıştır. Sonuç olarak insan çevirisine yakın çeviriler elde edilmiştir (Vaswani vd., 2017).

BERT modeli, büyük metin verileriyle eğitilmiş ve çok dilli bir model olduğu için birçok dili anlama yeteneğine sahiptir. Bert modeli kelime saklama özelliği sayesinde metinde geçen bazı kelimeleri saklar ve tahmin etmeye çalışır. Ayrıca metinler arasındaki bağlantıyı kavramak için sonraki cümle tahmini yaparak metinde geçen iki cümle arasındaki anlam ilişkisini öğrenmektedir. BERT modelinin OpenAI GPT den de temel farkı budur. Yani GPT sadece kelimenin solundaki kelimelere dikkat ederken BERT iki yönlü bir anlama yeteneğine sahiptir (Devlin vd., 2018).

Metinlerin yer aldığı her alanda sıklıkla kullanılmaya başlanan bu teknolojiler eğitim alanında, özellikle de doğal dilin kullanıldığı açık uçlu metinlerde de kullanılması önemli hale gelmiştir. Literatürde yazılı sınavları puanlamak üzere geliştirilmiş bazı sistemler bulunmaktadır. E-rater ve Criterion bunlara örnek olarak verilebilir (Dıklı, 2006).

E-rater, Educational Testing Service (ETS) tarafından geliştirilmiştir. E-rater ve Criterion araçlarında DDİ tekniklerinden yararlanılarak metinlerin dilbilimsel özellikleri tanımlanmakta ve metinler analiz edilmektedir. E-rater sözdizimsel modül, anlatı modülü ve konu analizi modülü olmak üzere üç ana bölümden oluşmaktadır. Sözdizimsel modül cümledeki fiil öbeklerini ve cümlecikleri tanımlamaktadır. Anlatı modülünde deneme yazısına ilişkin anlamsal yapılar yer ve yazı bütünselliği gibi yazının gelen hatları yer almaktadır. Konu analizi modülünde ise konu ile ilgili kullanılan kelimelerin konuyla

ilişkisi değerlendirilmektedir. E-rater sisteminde yazılan yazı ile veri tabanındaki yazının özellikleri karşılaştırılarak yeni yazıya bir puan ataması yapılmaktadır. Criterion ETS'nin internet tabanlı bir deneme puanlama ve değerlendirme sistemidir. Criterion sisteminde E-rater sistemine dayalı olarak anında puanlama ve öğretmenler tarafından öğrencilere verilen değerlendirme yazıları baz alınarak öğrencilere geri bildirim sunulmaktadır (Dıklı, 2006).

Otomatik puanlama ile ilgili tüm bu çalışmaların temelinde DDİ uygulamaları bulunmaktadır. Bilgisayar biliminin dili sayılardır. Dolayısıyla tüm bu metinler önce bilgisayarın anlayabileceği bir dil olan sayılara çevrilmektedir. Daha sonra bilgisayar bu sayılar yardımıyla metinleri işleyebilmekte ve anlamlandırabilmektedir. Bu noktada da makine öğrenmesi algoritmaları devreye girmekte ve metinlerin sınıflandırılması, kümelenmesi ve puanlanması gibi konularda bu algoritmalar kullanılmaktadır.

## Doğal Dil İşleme (DDİ-Natural Language Processing-NLP)

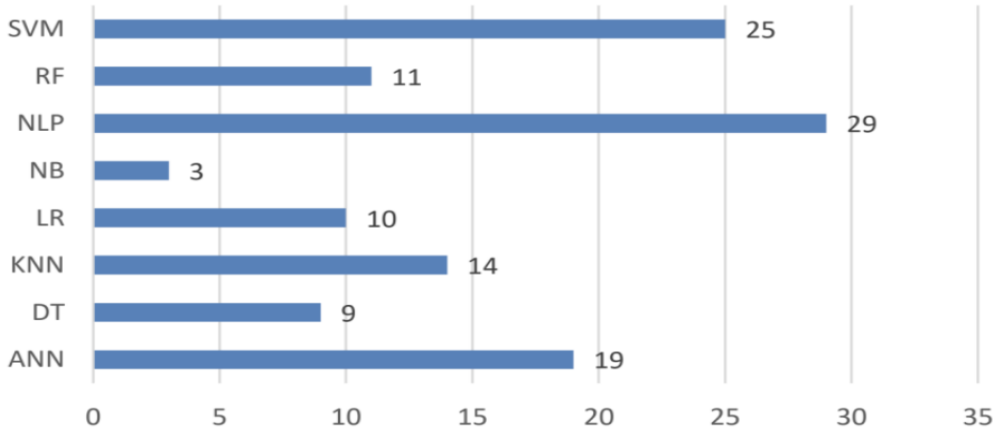
İnsanlık tarihinde dil önemli bir yere sahiptir. İnsanlar dil aracılığıyla birbirleriyle iletişim kurmaktadır. Diller üzerine yapılan çalışmaları, bilişim teknolojilerindeki gelişmeler teşvik etmiştir (Adalı, 2016). Dil, makine dili ve insanlar tarafından kullanılan doğal dil olmak üzere ikiye ayrılmaktadır (Şeker, 2015). DDİ, dillerin bilgisayar yardımıyla işlenmesi üzerinde çalışmaktadır (Adalı, 2016). Doğal dil işleme çalışmaları ile metin madenciliği çalışmaları, çoğu zaman beraber ele alınmaktadır. Metni veri kaynağı olarak kabul eden çalışmalardan oluşmaktadır. Metin madenciliği çalışmalarında; duygu analizi, metinlerden konu çıkarılması, metin özetleme, varlık ilişki modellemesi, metinlerin sınıflandırılması, gibi çalışmalar hedeflenmektedir (Şeker, 2015). Doğal dil işlemenin kullanıldığı çalışma alanlarından biri de eğitim bilimleridir. Bilgisayarların, insanların dillerini anlamlandırması için DDİ süreçlerini kullanması gerekmektedir (Şeker, 2015).

Liddy (2001) Doğal dil işleme süreçlerini aşağıdaki gibi açıklamaktadır.

1. Ses Analizi (Fonoloji): Bu aşama konuşulan dilin hangi harf ve kelimelerden oluşturduğunun çözümlenmesi ile ilgilidir. Böylece konuşmalar metne dönüştürülebilmektedir.

2. Kelime Yapısı (Morfoloji): En küçük anlam birimlerinde kelimelerin analizidir. Bu aşamada ekler, önekler, sonekler ve zaman ifadeleri ayrılmaktadır. Böylece bilgisayar kelimelerden anlam çıkarabilmektedir.
3. Kelime Anlamı (Leksik): Bir kelimenin ne anlama geldiğinin bilgisayar tarafından anlaşılmasıdır. Örneğin, "yaz" kelimesi, "bir şey yazmak" anlamında mı yoksa "yaz mevsimi" anlamında mı kullanılmış bunun belirlenme aşamasıdır.
4. Cümle Yapısı (Sentaks): Bir cümle düzeyinde analiz, dilbilgisi yapılarını anlam çıkarmaktır. Burada amaç bir eylemin ne üzerine yapıldığını, bir cümlenin öznesini ve nesnesini anlamaktır.
5. Anlam Analizi (Semantik analiz): Kelimelerin birbirleriyle oluşturduğu anlam ilişkisinin belirlendiği aşamadır. Bir kelimenin bir cümlede konuşlandırıldığı bağlamı tanımlamaktır. Bilgisayar bir kelimenin cümlede hangi anlamıyla kullanıldığını belirlemektedir.
6. Bağlam Analizi (Söylem analizi): Bir cümleden daha uzun metin birimlerinden anlam çıkarma işlemidir. Örneğin, "Ayşe defterini unuttu. Eve dönmek zorunda kaldı." cümlelerindeki öznenin Ayşe olduğunu anlamaktadır.
7. Bağlam ve Amaç (Pragmatik analiz): Bir metinde yer alan metin açıkça belirtilmeyen bağlamın incelenmesidir. Örneğin; garsona "su alabilir miyim?" diye soran biri aslında izin istemiyor su getirilmesini istiyordur. Bu durumun bilgisayar tarafından anlaşılması pragmatik analiz evresinde gerçekleşmektedir.

Bu süreçler bilgisayarın insan dilini anlamasına ve kullanmasına olanak sağlamaktadır. Metin sınıflandırma ve kategorizasyon, adlandırılmış varlık tanıma, dil çeviri ve yazım denetimi gibi alanlarda doğal dil işleme tekniklerinden yararlanılmaktadır (Adalı, 2016). DDİ günümüzde eğitim alanında da kullanımı yaygınlaşmakta olan bir alandır. Eğitimde DDİ uygulamaları; konuşma ve yazma tabanlı uygulamalar, eğitim için metin madenciliği, otomatik madde üretimi ve otomatik puanlama yaklaşımlarıdır (Flor & Hao, 2021). Şekil 1 incelendiğinde çalışmalarda en fazla kullanılan yöntemin doğal dil işleme olduğu görülmektedir. Bunu destek vektör makineleri ve yapay sinir ağları izlemektedir (Alanoğlu, 2022).




*Şekil 7.* Eğitim Alanındaki Makine Öğrenmesi İle İlişkili Çalışmalarda Kullanılan Yöntemler (Alanoğlu, 2022).

DDİ uygulamalarının genel yapısı Flor ve Hao (2021) tarafından şu şekilde özetlenmiştir: DDİ uygulamalarında ilk aşama ham metinlerdir. Daha sonra metin temizleme işlemi yapılarak metinde saklanması ve atılması gereken bileşenler belirlenmektedir. Durak sözcükler olarak adlandırılan kelimeler, metnin içeriğini belirleme noktasında önemsizdir. Bu nedenle DDİ uygulamalarında durak sözcükler metinden çıkarılmaktadır. Bir kelimenin metin içinde ne sıklıkla geçtiği 'unigram sıklığı' ile ifade edilmektedir. Örneğin; bir metinde 'kitap' kelimesi sıkça geçiyorsa konu 'kitap' ile ilişkilidir denilebilmektedir. Metinler vektörlerle temsil edilmekte, metinler sayılarla ifade edilmekte ve böylece DDİ uygulamaları gerçekleştirilmektedir.

Bu çalışmada, ABİDE sınavında yer alan Türkçe ortak sorularının verileri kullanılarak DDİ yöntemleri ile metin vektörleştirme işlemleri gerçekleştirilmiştir. Bu işlem, DDİ uygulamalarında sıklıkla kullanılır ve metinlerin anlamlarını ve özelliklerini kodlamaya yardımcı olur. Aşağıda metin vektörleştirme yöntemlerinden bazıları şu şekilde listelenebilir:

Kelime Çantası (BOW- Bag of Words): Bu yöntem, metnin kelime dağarcığını kullanır ve her bir kelime için bir vektör oluşturur. Bu vektörler, kelimelerin metinde kaç kere geçtiğini göstermektedir (Jurafsky & Martin, 2023).

Terim ve Ters Terim Frekansı (TF/IDF- Term Frequency/ Inverse Document Frequency): Kelimelerin yer aldıkları belgedeki önemini belirlemek için kullanılan bir yöntemdir (Leskovec vd., 2020). Bu yöntem, BOW yöntemine benzer şekilde metinler için kelime vektörleri oluşturur, ancak bu vektörler kelimelerin metinlerdeki sıklığının yanında kelimelerin tüm metinler içinde ne kadar yaygın olduğunu da hesaba katar. Bu,

kelimelerin önemini belirlemeye yardımcı olur. TF, bir kelimenin belgede görülme sıklığını ifade etmektedir (Ramos, 2003). IDF, kategoriler arasında ayrım yapma yeteneğinin bir ölçüsüdür. Bir kelimenin IDF'i kelimeyi içeren toplam belge sayısının, bölüm logaritmisinden sonra toplam belge sayısına bölünmesiyle elde edilebilir (Chen, Chen, & Liang, 2016). Kelimenin bir belgedeki görülme sayısı arttıkça IDF değeri düşmektedir.

Word2Vec: verilen metinleri vektörlerle temsil etmek için kullanılan bir yöntemdir. Bu yöntem, metinleri kelime öbeklerine (word embeddings) ayırarak, kelime öbeklerini vektörlerle temsil eder. Word2Vec yöntemi, verilen metinlerdeki kelime sıklıklarını değil, kelime öbeklerinin anlamlarını temsil eden bir vektör ölçeği oluşturur. Bu sayede, Word2Vec yöntemi, metinlerdeki kelime sıklıklarından çok, kelime anlamlarının benzerliklerini ölçerek, metinleri vektörlerle temsil eder (Mikolov, Chen, Corrado & Dean, 2013).

FastText: Facebook tarafından geliştirilen bir kelime temsil modelidir. Word2Vec tabanlı bir modeldir. Bu yöntemin kelimelerin n- gramlara ayrılması ile Word2Vec modelinden ayrılmaktadır. Örneğin, bir metin içerisinde 2 adet ardışık kelimeyi ifade etmek için "bigram" kullanılırken, 3 adet ardışık kelimeyi ifade etmek için "trigram" kullanılmaktadır. Bu algoritmalar, n-gramları kullanarak metin içerisindeki düzenli veya anlamlı dilbilgisi veya semantik yapıları tespit edebilmektedir (Bojanowski, Grave, Joulin, & Mikolov, 2017). Metnin vektörleştirilmesi (sayısallaştırılması) araştırmanın önemli bir adımıdır. Çünkü metinlerin hangi yöntemle elde edilen vektörlerle daha iyi temsil edildiğinin ortaya konulması gerekmektedir. Çelik ve Koç (2021)'un çalışmasında FastText, Word2Vec ve TF-IDF yöntemleriyle vektörleştirilen Türkçe metinlerin sınıflama başarısının birbirine çok yakın olduğu ifade edilmiştir. Bu bağlamda bu çalışmada en sık kullanılan geleneksel vektörleştirme yöntemlerinden biri olan TF-IDF, hesaplama ve uygulama kolaylığı sağlaması, bakımından araştırma kapsamında kullanılmak amacıyla tercih edilmiştir (Ramos, 2003). Dönüştürücü tabanlı bir yaklaşım olan BERT yöntemi ise bağlamsal bilgilere odaklanması ve vektörleştirme adımını da kendi içinde gerçekleştirmesi bakımından araştırma kapsamına dahil edilen diğer bir yöntemdir. Öğrenci cevaplarından elde edilen metinler bu iki yöntem ile ayrı ayrı vektörleştirilerek makine öğrenmesi ve BERT yöntemlerinden elde edilen sonuçlar karşılaştırılacaktır. Böylece Türkçe metin hem vektörleştirme hem de uygun otomatik puanlama algoritması belirlenecektir. Otomatik puanlamada kullanılan makine

öğrenmesi algoritmaları genellikle 4 adımdan oluşmaktadır (Powers, Escoffery & Duchnowski, 2015).

1. Bilgisayarı eğitmek için nitelikli bir puanlamayı bilgisayara tanımlamak
2. Eğitim verileri olacak metinlerden çeşitli özellikleri kaldırmak (durak kelime, noktalama işaretleri, satır sonlar vb.)
3. Metnin tüm nitelikleri hakkında bir model geliştirmek
4. Belirlenen modeli kullanarak değerlendirilmemiş metinlere değer atamak

Makine öğrenmesi yöntemlerinin gerçekleştirilmesinde önemli etkenlerden biri elle sınıflandırılmış eğitim kümesidir. Bir diğeri ise en yüksek başarıyı sağlayacak sınıflandırma algoritmasının seçilmesidir. Veriler eğitim kümesi (training set) ve test kümesi (test set) olmak üzere veri iki gruba ayrılmaktadır. Eğitim kümesi, algoritmaların veriden öğrenmesini sağlamak üzere kullanılmaktadır. Test kümesi ise sınıflandırma algoritmalarının başarısının ölçülmesinde kullanılmaktadır. Böylece sınıflandırıcı daha önce görmediği test kümesindeki metinleri sınıflandırabilmektedir. Sınıflandırma sonuçları ve metinlerin önceden elle yapılan sınıflandırma sonuçları karşılaştırılarak modellerin başarı oranı bulunmuş olmaktadır (Tantuğ, 2012).

Bu çalışmada metinlerin vektöre dönüştürülmesinin ardından gerçek puanlayıcıların puanladıkları cevaplar eşleştirilmiştir. Daha sonra vektörleşmiş metinler ve puanlar eşleştirilerek Naif Bayes, Lojistik Regresyon, Karar Ağacı, Rastgele Orman, Gradyan Artırma makine öğrenmesi algoritmaları ile BERT modelinden elde edilen sonuçlar karşılaştırılmıştır. Araştırmanın son aşamasında makine öğrenmesi modelleri ile yapılan puanlama ile gerçek puanlayıcılardan elde edilen puanların uyumu hesaplanmıştır. İki farklı puan uyumunu değerlendirmek için kullanılan en yaygın istatistiksel yöntem ağırlıklandırılmış kappa (QWK) istatistiğidir (Rodriguez, Jafari & Ormerod, 2019). Bu çalışmada insan ve makine puanlayıcı arasındaki puanlayıcı uyumu QWK metriği ile hesaplanmıştır.

Dünya üzerinde oluşan ve kullanılan veriler sayısal verilerle sınırlı değildir. Hatta sayısal veriler sözel verilere oranla günlük yaşamda daha az kullanım alanına sahiptir. İnsanlar farklı diller aracılığıyla iletişim kurmaktadır. Farklı dillerden oluşan yazılı pek çok eser (kitap, makale, mektup vb.), sosyal medya yazıları (twitter, instagram, facebook vb.) çok büyük bir metin verisi oluşturmaktadır. Doğal dil işleme teknikleri bu denli büyük metin verilerini işleyerek anlamlı sonuçlar ortaya çıkarmaktadır. Doğal dil işleme tekniklerinin bu gücü eğitim alanının da ilgisini çekmiş ve doğal dil işlemenin eğitime

entegrasyonu yeni bir çalışma alanı oluşturmuştur. Bu çalışmada Türkçe doğal dil işleme ele alınmaktadır. Türkçe DDİ diğer dillerden farklı bir dil yapısına sahip olması sebebiyle incelenmesi gereken bir alandır.

## Türkçe Doğal Dil İşleme Yaklaşımları ve Karşılaşılan Zorluklar

Türkçe Altay dillerinin (Mogol, Tunguz, Kore ve Japon) Türk dilleri ailesi içinde yer alan bir dildir. Altay dillerinde Türk dillerinin dışında dil aileleri de bulunmaktadır (Oflazer, 2016). Türkçe cümlelerde "Özne- Nesne -Yüklem" öge dizimi bulunmaktadır. Bunlar dışında zaman, yer vb. belirteç ögeler bulunmaktadır.

Sözcükler cümle içinde kullanıldığında bir dizi çekim ve yapım eki almaktadır. Bu duruma bir örnek vermek gerekirse bir sözcük İngilizce'de bir cümleye karşılık gelebilmektedir. Örneğin (Oflazer, 2016, s.2);

yap+abil+ecek+se+k!

if we will be able to do (it)

Morfolojik olarak Türkçe, "ipteki boncuklar" gibi bir kök kelimeye eklenen morfemlere sahip sondan eklemeli bir dildir. Ön ek ve birleştirme yoktur (Oflazer ve Saraçlar, 2018). Özne - Nesne – Yüklem farklı yerlerde de kullanılabilmektedir. Bu değişik öge sıralarını modellemek olanaklı olsa da modelin beklendiği kadar temiz veya basit olmadığını ifade eden Oflazer'den (2016, s.6) uyarlanan aşağıdaki örneklerle açıklamaktadır.

Ali Ahmet'i gördü.

Ahmet'i Ali gördü. (gören Ali başka birisi değil!)

Gördü Ali Ahmet'i. (ama görmemesi gerekiyordu.)

Gördü Ahmet'i Ali. (zaten görmesini bekliyordum!)

Ali gördü Ahmet'i. (başkası da görebilirdi)

Ahmet'i gördü Ali. (başkasını da görebilirdi!)

Yukarıdaki cümlelerin hepsinde eylem Ali'nin Ahmet'i görmesidir. Ögeler arasındaki sıra değişiklikleri cümlenin anlamını etkilemektedir. Oflazer (2016, s.6) "Türkçe'de, bir kelimenin yapısal yapısını genel olarak bir kök kelimeye eklenen ve bilgi içeren işaret dizileriyle kodladıklarını" ifade etmektedir. Oflazer (2016, s.6)'e göre "bu işaret dizilerinden biri olan "DB" yapım eklerinin sınırlarını belirtmektedir. Sözcüğün başından ilk yapım ekine, son yapım ekinden sözcüğün sonuna ve iki yapım eki arasındaki çekim eklerinde oluşan her bir gruba "çekim grubu" veya "inflectional group (IG)" adı verilmektedir.

Örneğin "uzaklaştırılacak" sözcüğünün gösterimi şu şekildedir:

> uzak+Adj DB+Verb+Become DB+Verb+Caus DB+Verb+Pass+Pos DB+Adj+FutPart+Pnon" "(Sembollerin Türkçe anlamları şu şekildedir: Sıfat (Adj), Eylem (Verb), Dönüşüm yapım eki (Become), Ettirgen (Caus), Edilgen (Pass), Olumlu (Pos), Gelecek zaman ortacı (FutPart), İyelik eki (Pnon).)" (Oflazer, 2016, s.7).

> Birinci çekim grubu sıfat kökünü ifade eder. İkinci çekim grubu, sıfat kökü üzerine anlamı o sıfata dönüşmek olan bir eylem türeterek kullanılır (örneğin, "uzaklaşmak"). Üçüncü çekim grubu ise bir önceki eylemden nesne alabilen bir eylem türetildiğini gösterir (örneğin, "uzaklaştırmak"). Dördüncü çekim grubu ise bir önceki eylemin edilgen hali olduğunu ifade eder (örneğin, "uzaklaştırılmak"). Son olarak, bir önceki edilgen eylemden gelecek zaman ortacı türetilir ve bu ortaç, cümlede başka bir ismin nitelendiricisi olarak kullanılır.(Oflazer, 2016, s.7).

Türkçe'nin bu yapısı doğal dil işleme noktasında özel çalışmaların yapılmasına gereksinim oluşturmaktadır. Türkçe dil işleme için pek çok araç ve kaynak, son yirmi yılda temelde sıfırdan başlayarak geliştirilmiştir (Saraçlar & Oflazer, 2018). Türkçe, morfolojik yapısı karmaşık bir dildir. Kök bir kelime veya gövdeye eklenen morfemler, isimden fiile veya zarfa dönüşebilmektedir (Oflazer, 1994).

Oflazer (1994) tarafından Türkçe için Xerox sonlu durum araçları kullanılarak yapılan iki seviyeli bir morfolojik analizci tanımlamıştır. Bu çözümleyici, Türkçe kelimelerin morfolojik yapısını ve olası telaffuzlarını üreten genel bir sistem oluşturmak için kullanılmıştır (Oflazer & Inkelas 2006). Morfolojik analiz üzerine yapılan diğer çalışmalar bulunmaktadır (Hakkani Tür, Oflazer & Tür, 2002; Sak, Güngör & Saraçlar, 2007; Oflazer & Kuruöz, 1994; Oflazer & Tür, 1996; Yuret & Türe, 2006).

Oflazer ve Kuruöz (1994) çalışmalarında Türkçe'nin eklemeli yapısı nedeniyle oluşan morfolojik belirsizliklerin giderilmesi amacıyla kurallar, kısıtlamalar v e istatistiksel bilgiler kullanan bir araç geliştirmişlerdir. Bu araç metinlerin %98-99'unu doğru etiketleyebilmektedir. Bu araçta çoklu kelime yapıları ve deyim gibi ifadeler de

dikkate alınmaktadır. Oflazer ve Tür (1996), Türkçe'nin morfolojik yapısı üzerine bir çalışma daha yapmıştır. Bu çalışmada hem manuel olarak belirlenen dilsel kurallar hem de gözetimsiz öğrenme yöntemleri ile keşfedilen kurallar bir arada kullanılmıştır. Bu çalışmada dilin morfolojik belirsizliklerini ortadan kaldırmak amaçlanmıştır. Çalışma sonucu incelendiğinde, çalışmada kullanılan yöntemle %96-97 doğruluk oranı elde edilmiş olduğu ve belirsizliğin büyük oranda azaltıldığı görülmektedir.

Hakkani Tür, Oflazer ve Tür (2002) çalışmalarında Türkçe dilinin karmaşık yapısını ve kelimelerin aldıkları ekleri anlamlandırmak için bir model önerisinde bulunmuşlardır. Bu modelde bir kelimeye hangi ekler eklenerek kelimenin nasıl türetildiğini anlamlandırmak için gruplara ayrılan özelliklerden yararlanılmaktadır. Çalışmada dört model test edilmiştir. Bu çalışmada test verilerinin %93.95 doğruluk oranına ulaşarak morfolojik ve anlamsal özellikleri doğru bir şekilde belirlenmiştir. Sözdizimsel olarak ilgili özelliklere odaklanarak ve küçük bir anlamsal özellik setini göz ardı edilerek modelin doğruluk oranı %95.07'ye yükselmiştir. Yani, model sadece kelimenin dil bilgisel işlevini (özne mi, nesne mi, hangi ekler kullanılmış gibi) göz önüne alarak çalıştığında daha yüksek bir başarı elde etmektedir.

Yuret ve Türe (2006) de benzer şekilde Türkçe'nin morfolojik yapısının incelenmesi ve gerekli çözümlerin belirlenmesi adına bir çalışma yapmışlardır. Bu çalışmada Türkçe metinlerde morfolojik belirsizlik çözümü için yeni bir kural tabanlı model sunulmaktadır. Çalışma, yaklaşık 1 milyon kelimelik bir Türkçe haber metni veri seti kullanılarak bir karar listesi oluşturulmuştur. Ayrıca veri seyrekliği durumu da çalışmada ele alınmıştır. Sonuç olarak model %96 doğruluk oranına ulaşmıştır. Ayrıca, tam etiketler üzerine eğitilen tek bir karar listesi modeli ise %91 doğruluk oranı sağlamıştır.

Sak, Güngör, & Saraçlar, (2007)'de Türkçe'nin morfolojik yapısı üzerine çalışmış ve Perceptron Algoritması kullanarak modelin doğruluğunu %93.61'den %96.80'e çıkarmıştır. Perceptron, POS etiketleyici olarak %98.27 doğruluk oranı sağlamıştır. Eryiğit ve Oflazer (2006, s.91) ise Türkçe kelimeler için tahminler yapan istatistiksel bir bağımlılık ayrıştırıcı geliştirmişlerdir. Çalışmada bağımlılık ayrıştırıcı mantığı aşağıdaki örnek cümle ile açıklanmıştır.

"Bu eski ev+de+ki gül+ün böyle büyü+me+si herkes+i çok etkile+di"

"This old house-at+that-is rose's such grow+ing everyone very impressed"

"(*Such growing of the rose in this old house impressed everyone very much.*)"

Yukarıdaki cümlede "bu" kelimesi "eski" kelimesi ile bağımlıdır. Bağımlılık ayrıştırıcı, cümleyi sondan başlayarak analiz eden ve kelimeler arasındaki bağımlılık ilişkilerini belirleyen bir algoritma kullanılmaktadır.

Çetinoğlu (2009) doktora tezinde kurallar, sıfat, zarf bileşenlerden türetilmiş yapılara kadar geniş bir alanı kapsamaktadır. Çalışmada yapılan gramer değerlendirmesi sonucunda isim öbeklerinde %85.5 oranında başarı elde edilmiştir. Bu çalışmada Türkçe için geniş kapsamlı bir dil bilgisel gramer kullanılarak çözümleyici geliştirilmiştir. Bu çalışmada gramer sıra farklıkları modellenmiştir (Oflazer, 2016).

Zeyrek ve diğerleri (2009) tarafından Türkçe söylem bağlaçlarının ek açıklamalarını içeren bir Türkçe söylem bankası geliştirilmiştir. Bu banka, Türkçe dilbilim ve doğal dil işleme çalışmalarında kullanılmak üzere tasarlanmıştır. Bu çalışmada kullanılan bazı örnekler şu şekildedir:
Türkçe'de özne düşmesine izin verilir ve karmaşık cümlelerde, paylaşılan özne sonuç cümlesinin fiilindeki kişi uyumu eki ile gösterilmektedir. 'Neriman' öznesi, sonraki cümlenin fiilindeki kişi uyumu ekinin bulunması nedeniyle ikinci cümlede yazılmamıştır. Ancak burada özenin yine 'Neriman' olduğu anlaşılmaktadır.

> Neriman yatak odasında sigara içilmesini istemediği halde şimdilik sigaraya ses çıkarmıyor.
>
> Although Neriman does not want people to smoke in her bedroom, she doesn't say anything for the moment.

Eğer özne paylaşılmıyorsa, özne kesintiye neden olsa bile anotasyona dahil edilir. Aşağıdaki örnekte olduğu gibi, "ARG2", "ARG1"e müdahale ederek onu öznesinden ayırır şekilde gösterilmiştir.

> Rukiye, kendisinden üç yaş ufak olmasına rağmen, erkek kardeşini kendi oğlu sanıyordu.
>
> Rukiye, although he is three years younger than her, thought of her brother as her own son (Zeyrek vd., 2009, s.46).

Bu cümlede de aynı durum söz konusudur. Rukiye'nin özne olduğu, fiil ekiyle anlaşılmakta ve Rukiye öznesi tekrar edilmektedir.

Tufiş, Cristea ve Stamou (2004) tarafından geliştirilen projede yaklaşık 15.000 eş anlamlı kelime grubundan oluşan bir veri tabanı oluşturulmuştur. BalkaNet projesi 2001-2004 yılları arasında AB fonlu bir proje olarak yürütülmüştür. Projenin amacı; Türkçe, Bulgarca, Sırpça, Romence gibi balkan dilleri için 'kelime ağları' geliştirmektir. Araştırma sonucunda 8.516 ortak kavram dilbilimsel analizlerde kullanılabilir hale gelmiştir.

İlgili literatür incelendiğinde Fince, Korece ve Japonca gibi dil yapısı itibariyle Türkçe'ye benzeyen dillerde sınırlı sayıda da olsa otomatik puanlama çalışmalarının yapıldığı görülmektedir. Örneğin; Fin dilindeki metinlerin otomatik puanlanması için bir puanlama sistemi oluşturulmuştur. Bu çalışmada gerçek puanlayıcılar tarafından yazının puanlanmasında, benzer yazılarla karşılaştırma ve ders içeriği ile ilgili materyallerden yararlanma olmak üzere 2 yaklaşımdan yararlanılmıştır. Ayrıca Fin dili sondan eklemeli olduğu için son ekleri ayrıştırma modülü kullanılmıştır. Araştırmada LDA ve PLDA (örtük anlamsal ve olasılıksal örtük anlamsal) benzerliği belirlemeye dayalı vektörler kullanılmıştır. Bu yaklaşım konu modellemesi olarak da kullanılmaktadır. Araştırma sonuçlarına göre örtük anlamsal yöntem daha iyi sonuçlar vermiştir. Bu yöntem için gerçek puanlayıcılar ile otomatik puanlama sistemi arasında 0.54 ile 0.90 aralığında değişen korelasyon değerleri hesaplanmıştır (Kakkonen vd., 2005).

Kore diline dayalı, bir otomatik puanlama sistemi (KASS-Korean Automatic Scoring System for Short-answer Questions) geliştirilmiştir. Araştırma her bir maddeye 3012 ila 3035 öğrencinin yanıt verdiği 7 madde üzerinde yürütülmüştür. Çalışmada DDİ veri ön işleme adımlarının ardından öğrenci cevaplarında yer alan kelimeler frekansa göre sıralanmış ve puanlayıcılar tarafından puanlanmıştır. Anahtar kelime seçeneği kullanıldığında puanlayıcılar sisteme anahtar kelimeleri tanımlamaktadır. Böylece yanlış cevaplar tespit edilebilmektedir. Araştırma sonucunda otomatik puanlama gerçek puanlayıcılarla ortalama %95 uyum yüzdesine erişmiştir. Soru 6 ve 7 de bu oran %99'a ulaşmıştır. Buna göre sonuçlar, sistemin geniş ölçekli testlerde kullanılabileceğini göstermiştir (Jang vd., 2014).

Ishioka ve Kameda (2006) ise; JESS (Uzmanlar Tarafından Yazılmış Makalelere Dayalı Otomatik Japonca Kompozisyon Puanlama Sistemi) adlı bir kompozisyon puanlama sistemi geliştirmişlerdir. Bu otomatik puanlama sisteminde uzman puanlayıcılar yerine referans metinleri kullanmıştır. Bu sistemde metinler üç değerlendirme ölçütüne göre değerlendirilmektedir. Bunlar; retorik: söz dizimi çeşitliliği kelime kullanımı ve cümle yapısı, organizasyon: fikirlerin mantıklı sunumu ve gerekli

bağlantıların kurulması, içerik: konuya uygunluk kelime dağarcığı ve bilgi seviyesidir. Sistemde tam puan 10'dur. Bu puanın 5'i retorik, 2'si organizasyon, 3'ü içerik kriterlerine ayrılmıştır. Puanlama aşamasında bir gazetenin makale ve başyazıları kullanılmıştır. Daha sonra sistemin sınıflar arası korelasyon katsayısı ile performans değerlendirmesi yapılmıştır. Jess ile uzman değerlendiriciler arasındaki korelasyon 0.57'dir. Uzmanların kendi arasındaki korelasyon ise 0.48'dir. Bu durum Jess sisteminin insan değerlendiricilerden daha iyi performans gösterdiğini ifade etmektedir. Araştırmada yapılan diğer bir puanlama denemesinde üniversite öğrencilerinin 'sigara' konulu yazıları için otomatik puanlama sistemi ve gerçek puanlayıcılar arasındaki korelasyon 0.83, gerçek puanlayıcıların kendi aralarındaki korelasyon ise 0.70 bulunmuştur. Araştırmada ayrıca E-rater puanlama sisteminde yer alan 7 farklı kompozisyon Japoncaya çevrilerek sistemler arası karşılaştırma yapılmış Jess sisteminin E-rater sistemi ile uyumlu puanlar verdiği sonucuna ulaşılmıştır.

Literatür incelendiğinde Türkçe metinlerle yürütülen çalışmalar olduğu da görülmektedir. Uysal (2019)'un yaptığı çalışmada, otomatik puanlama sistemleri ve gerçek puanlayıcılardan elde edilen puanların uyumunu incelenmiştir. Buna ek olarak çalışmada otomatik puanlama sistemlerince puanlanan maddelerin test eşitleme hataları üzerindeki etkisi değerlendirilmiştir. Çalışmada otomatik puanlamada destek vektör makinesi (SVM) lojistik regresyon (LR), çok terimli sade bayes (MNB), kısa uzun süreli bellek (LSTM) ve iki yönlü kısa uzun süreli bellek (BLSTM) yöntemleri kullanılmıştır. Çalışmada en iyi sonuçların BLSTM yöntemiyle elde edildiği vurgulanmıştır (Uysal, 2019; Uysal ve Doğan, 2021). Test eşitleme sonuçlarına bakıldığında ise Madde Tepki Kuramı'na dayalı test eşitleme yöntemlerinin Klasik Test Kuramı'na dayalı yöntemlere göre daha düşük hata oranlarıyla eşitleme yaptığı belirlenmiştir.

Türkçe dilinde yapılan başka bir otomatik puanlama çalışmasında öğrencilere açık uçlu Türkçe Fizik soruları yöneltilmiş ve 246 lisans öğrencisi cevabı makine öğrenmesi algoritmalarıyla puanlanmıştır. Çalışmada kullanılan algortimalar, SVM (Destek Vektör Makineleri), Decision Tree (Karar ağacı), KNN (K en yakın komşuluk), Bagging ve Adaboost.M1 algortimalarıdır. En iyi performans sağlayan algoritma AdaBoost.M1 algortimasıdır. Adaboost.M1 algoritmasıyla elde edilen doğruluk (accuracy) değeri 10 iterasyon sonucunda 0.84-0.99 arasında değişmektedir. Çalışmada Türkçe otomatik puanlamanın yapılabildiğine ve geniş ölçekli testlerde açık uçlu soruların kullanılabileceğine dikkat çekilmiştir (Çınar vd., 2020).

Çelik ve Koç (2021) ise çalışmalarında Türkçe haber metinlerinin sınıflandırılması amaçlanmıştır. Veri ön işleme adımlarının ardından metinler Tf/idf vectorizer, Word2Vec ve FastText yöntemleri kullanılarak ayrı ayrı sayısallaştırılmıştır. Daha sonra NB, LR, RF, SVM ve yapay sinir ağı yöntemleri ile sınıflandırılmıştır. Metinler bilim-teknoloji, eğitim, kültür-sanat, sağlık, siyaset ve spor olmak üzere 6 kategoride sınıflandırılmıştır. Araştırma sonucunda metin vektörleştirme yaklaşımlarının sonuçlarının birbirine yakın olduğu ancak Fasttext yönteminin %95.75 oranıyla diğer yöntemlere göre daha iyi performans sergilediği görülmüştür.

Türkçe ve Türkçe'ye benzer yapıdaki çalışmalar incelendiğinde genel olarak farklı vektörleştirme yöntemleri (Tf/idf, Word2Vec ve FastText) makine öğrenmesi algoritmaları, yapay sinir ağları ile çalışmaların yürütüldüğü görülmektedir. Bu tez çalışmasında ise makine öğrenmesi algoritmalarına ek olarak dönüştürücü tabanlı BERT algoritması ele alınmıştır.

2004-2016 yılları arasında otomatik puanlamaya ilişkin geliştirilen sistemler incelendiği bir çalışmada geliştirilen bu sistemlerin avantaj ve dezavantajlarına yer verilmiştir (Rokade, Patil, Rajani, Revandkar & Shedge, 2018). Buna göre 2004 yılında ele alınan açık uçlu cevaplar için akıllı puanlama sisteminin sadece anahtar kelimelerin eşleşmesini ele aldığı ve cevapları bağlamsal olarak değerlendirmediği ifade edilmiştir. 2010 yılında geliştirilen sistemin geleneksel ve dijital sınavları puanlamayı amaçladığı ancak sadece büyük miktarda veri ile çalışılması gerektiği vurgulanmıştır. 2010 yılında geliştirilen diğer bir sistemde Hint dili bağlamında otomatik puanlama sistemi tasarlanmıştır. Sistemde İngilizce dışındaki yerel dil etkisi ele alınmıştır. Ancak sistemin yerel dilde genel bir çözüm sunmadığı ortaya çıkmıştır. 2011 yılına gelindiğinde ise makine öğrenimi kullanılarak otomatik makale puanlama çalışması yapılmıştır. Bu çalışmada öğrenci kısa cevaplarının anlamsal özelliklerine dayalı otomatik puanlama amaçlanmıştır. Ancak sitemin anahtar kelimeler üzerinde yoğunlaştığı görülmüştür (Rokade vd., 2018).

2012 yılında benzer bir yaklaşım sunularak makine öğrenmesi ile makale puanlaması yapılmıştır. Sistemde büyük veri setleri kullanılması amaçlanmış ancak yine içerik ve dil özellikleri bakımından sınırlı kalmış ve sözcüklerin eş anlamlarını dikkate almamıştır. 2012 yılında bir başka sistemde serbest metin cevaplarından bilgi çıkarma temelli bir yaklaşım benimsenmiştir. Ancak sistemin dezavantajı olarak veri toplama standardizasyonu eksikliği olduğu ve sözcüklerin anlam ilişkilerini incelemediği vurgulanmıştır. 2013 yılında büyük bir veri seti kullanılarak makale puanlaması

yapılmıştır. Ancak bu sistemde de yapısal özelliklerin dikkate alınmadığı görülmüştür (Rokade vd., 2018).

2014 yılında DDİ kullanılarak öğrenci cevaplarını değerlendiren çalışmada, öğrenci kısa cevaplarını tam anlamıyla işleyen bir sistem amaçlanmıştır. Bu sistemin de yalnızca anahtar kelimeleri kontrol ettiği görülmüştür. Son olarak 2016 yılında yapılan çalışmada makine öğrenmesi teknikleri ele alınmıştır. Çalışmada gizli anlamsal analizin kullanıldığı sistemde öğrenci cevapları arasındaki anlamsal gizli ilişkiler yakalanmaya çalışılmıştır. Bilgisayar bilimi ile ilgili daha kapsamlı terimlerin kullanılması gerektiği ve konuların daha detaylı incelenmesi gerektiği ifade edilmiştir (Rokade vd., 2018).

Alikaniotis, Yannakoudakis ve Rei (2016) çalışmalarında Uzun-Kısa Süreli Bellek (LSTM) kullanarak bir otomatik puanlama sistemi geliştirmişlerdir. Geleneksel makine öğrenmesi yaklaşımlarında (regresyon, sıralama vb.) manuel çabanın yoğun olduğunu ifade eden araştırmacılar otomatik puanlama sürecinin daha aktif kullanılabilmesi ve iş yükünün azaltılabilmesi için sinir ağlarının kullanımını önermektedir. Araştırmacılar tarafından geliştirilen SSWE (Score-Specific Word Embeddings) adlı kelime temsil yöntemi, kelimelerin bir metnin puanına katkısını öğrenerek kelime temsilcileri oluşturan bir yaklaşımdır. Bu yöntemin Word2Vec gibi kelime temsil yöntemlerinden daha iyi performans sergilediği ifade edilmiştir. Çalışmaya göre en başarılı model SSWE ve iki katmanlı BLSTM yöntemlerinin birlikte kullanıldığı model olmuştur. Bu model için değerlendirme metrikleri şu şekildedir: Spearman Korelasyonu (ρ): 0.91, Pearson Korelasyonu (r): 0.96, RMSE (Root Mean Square Error): 2.4, Cohen's Kappa (κ): 0.96.

Wang, Wei, Zhou ve Huang (2018) Çalışmasında QWK metriği sadece bir değerlendirme ölçütü olarak değil aynı zamanda veri eğitiminin bir parçası olarak ele alınmıştır. QWK optimize edilerek sürece dahil edilmiştir. Ayrıca uzun makalelerde bilgi kaybını önlemek için ‘dilated LSTM’ yöntemi önerilmiştir. Böylece daha adil bir otomatik puanlama sağlanmaya çalışılmıştır.

Krishnamurthy, Gayakwad ve Kailasanathan (2019) tarafından kısa cevaplı maddelerin puanlanmasına ilişkin yürütülen çalışmada derin öğrenme yaklaşımları karşılaştırılmıştır. Çalışmada derin öğrenme modellerinin geleneksel makine modellerine kıyasla daha az özellik mühendisliği gerektirdiği vurgulanmıştır. Çalışmada kısa cevaplar insan puanlayıcılar tarafından puanlanmış ve böylece model eğitilmiştir. Veri kümesi olarak Hawlett Vakfı tarafından düzenlenen yarışmadan elde edilen veriler kullanılmıştır (Kaggle, 2012.). Verilerin 17.000’i eğitim, 5100’ü test verisi olarak kullanılmıştır.

Çalışmada karşılaştırılan yöntemler CharCNN, CNN, LSTM ve BERT modelleridir. Çalışmada uyumu ölçmek için Ağırlıklandırılmış Kappa (QWK) metriğinden yararlanılmıştır. Bu modellerin sırasıyla QWK değerleri 0.60, 0.62, 0.65, 0.71 şeklindedir. BERT modeli diğer modellere göre daha iyi bir performans sergilemiştir. LSTM modeli ise CNN modellerine göre daha iyi performans göstermiştir.

Diğer bir çalışmada öğrencilerin öğrenme durumlarına ilişkin geri bildirim sağlayan bir model tanıtılmıştır. Bu modelde k-means kümeleme analizinden yararlanılmıştır. Çalışmada 3.370 öğrenciden elde edilen 168.000 veriden yararlanılmıştır. CMD model olarak tanıtılan bu modelde öğrencilerin soruyu doğru-yanlış cevaplama durumu, öğrencinin cevabını değiştirme- düzeltme sayısı ve sorunun zorluk seviyesi ele alınmaktadır. Bu özelliklerden yararlanan model, öğrencileri sekiz sınıfa ayırmaktadır. Model, öğretmenlere öğrenciler hakkında bilgi vermekte, öğrencilerin güçlü ve zayıf yönlerini tespit etmektedir. Böylece öğrencilere uygun geri bildirimler verilebilmektedir (Sun & Shao, 2019).

Eğitimde metin madenciliği yaklaşımlarının incelendiği bir çalışmada en sık kullanılan metin madenciliği teknikleri ele alınmıştır. Bunlar; metin sınıflandırma, bilgi getirme, yazılı cevapların değerlendirilmesi, metin kümeleme, metin özetleme gibi tekniklerdir. Araştırma bulguları incelendiğinde metin madenciliği yaklaşımlarının, değerlendirme, öğrenci motivasyonunu destekleme, öğrenci davranışlarını ve ders performansını izleyerek daha iyi rehberlik etme gibi noktalarda yarar sağladığı görülmektedir (Ferreira, André, Pinheiro, Costa & Romero, 2020).

Doğal dil işlemenin açık uçlu maddeleri puanlamada kullanıldığı çalışmaları inceleyen diğer bir çalışmada daha az insan müdahalesi gerektiren yöntem olan denetimsiz makine öğrenmesi yöntemlerine odaklanılmıştır. Çalışmada metin ön işleme (durak kelimeleri, gereksiz işaretleri, boşlukları kaldırma vb.), veri işleme (verilerin sayısallaştırılması ve benzerliklerinin belirlenmesi) ve veri kümeleme (benzer cevapları gruplama) gibi adımlar ele alınmıştır. Çalışma doğal dil işlemenin genel yapısı ve kullanılan tekniklerin sunulması üzerine yapılmış bir tarama çalışmasıdır (Kerkhof, 2020).

Bir başka çalışmada kısa cevaplı metinlerin değerlendirilmesine ilişkin zorluklara ve puanlayıcı kaynaklı hatalara değinilerek, otomatik puanlamanın bu anlamda olumlu katkılar sağlayacağı fikri savunulmuştur. Çalışmada DDİ uygulamalarının kısmen veya tamamen sınavların değerlendirilmesinde kullanılabileceği ifade edilmiştir. Çalışmada ayrıca otomatik puanlamanın sadece sınav sonucu puanlamada değil, aynı zamanda

öğrenciye geri bildirim verme noktasında da kullanılabileceği vurgulanmıştır. İlgili çalışmada hem denetimli hem denetimsiz makine öğrenmesi yaklaşımları ele alınmıştır. Çalışmada model cevabı ile öğrenci cevabı arasındaki uzaklıklardan yararlanılarak bir model geliştirilmiştir. Çalışmada, algoritmanın işlemekte zorlandığı soruların puanlayıcılar arası da uyumsuzluğa sebep olduğu görülmüştür. Çalışma sonucunda iki insan ve bilgisayar puanlayıcı arasındaki korelasyon; 0.81 ve 0.83 olarak hesaplanmıştır. Geliştirilen model ile insan puanlamaları arasındaki hata oranı 0.17 iken insan puanlamalarının hata oranı 0.25'tir. Dolayısıyla modelin insan puanlamasından daha düşük bir hata payı ile çalıştığı görülmüştür. Sonuç olarak otomatik puanlamanın yararlı olduğuna, öğrenciye geri bildirim verme noktasında kullanılabileceğine vurgu yapılmıştır (Süzen, Gorbana, Levesleya & Mirkes, 2020).

Literatür incelendiğinde genel olarak çalışmalarda metin sınıflandırma yaklaşımları yoluyla otomatik puanlama yapıldığı görülmektedir. Ancak Wang (2020) çalışmasında metin sınıflandırma yerine metin kümeleme yaklaşımını benimsemiştir. Denetimsiz makine öğrenme yöntemlerinden biri olan kümeleme yöntemi eğitim süreci gerektirmeden doğrudan veriler üzerinde işlem yapabilmektedir. Verileri benzerlik ölçütlerine göre kümelemektedir. Bu yöntemde metinlerin önceden etiketlenmesi gerekmemektedir. Çalışmada metinlerin bilgisayarın anlayabileceği bir formatta metin temsillerine dönüştürülmesinin gerekliliğini vurgulanmaktadır. Benzerlik hesaplamasının bu metin temsilleri arasındaki öklidyen mesafesi veya vektörler arasındaki açısının kosinüsü ile ölçüldüğü ifade edilmiştir. Bu mesafe yaklaştıkça veya açı küçüldükçe iki metin arasındaki benzerlik artmaktadır. Sonuç olarak metin kümeleme yönteminin benzer metinleri gruplandırarak İngilizce metinlerin otomatik puanlama doğruluğunun arttırılabileceği ifade edilmiştir (Wang, 2020).

Gunawansyah, Rahayu, Nurwathi, Sugiarto ve Gunawan (2020) çalışmalarında tekrar programlanmaya ihtiyaç duymayan, kendi kendine öğrenen bir sistem geliştirmiştir. Bu sistemde öğretmen sınav sorularını sisteme girdiğinde anahtar kelimeler oluşturulmaktadır. Bu anahtar kelimelerin eş anlamlıları program tarafından kullanıcıya sunulmaktadır. Öğretmen isterse eş anlamlı kelimeleri de ekleyebilmektedir. Bu da insanın doğal dilini anlama noktasında yarar sağlamaktadır. Öğrenciler, öğretmen tarafından sorulan soruları yanıtlayıp yanıtı gönderdiklerinde sistem otomatik olarak cevabı gerçek zamanlı olarak değerlendirerek öğretmenin sayfasında öğrencinin aldığı puan görüntülenebilmektedir. Bu araştırmada 3 değerlendirme aşaması bulunmaktadır. Bunlar; insan değerlendirmesi, kelime eş anlamlısı eklenmemiş otomatik sistem ve

kelime eş anlamlısı eklenmiş otomatik sistemdir. Araştırma sonucuna göre; anahtar kelimelerin eş anlamlısı kullanıldığında sistem insan değerlendirici gibi bir yetenek kazanmış olmakta ve daha iyi performans göstermektedir. Sonuç olarak insan puanlayıcının 30-65 arası yaptığı bir puanlama, eş anlamlıların kullanılmadığı sistemde 9-44 arası değer üretirken, eş anlamlıların kullanıldığı sistemde 37.3, 55.55 değerlerini üretmiştir. Dolayısıyla eş anlamlıların kullanıldığı sistemde, insan puanlayıcıya yakın değerler verdiği görülmüştür (Gunawansyah vd., 2020).

Chamidah, Santoni, Irmanda, Astriratma, Tua ve Yuniati (2021) öğrenci cevabı ve referans cevap arasındaki benzerlikten yararlanılan bir makale puanlama sistemini tanıtmışlardır. Makalede Endonezya'da kısa cevaplı maddelerin yanıtlanmasından elde edilen metin verileri kullanılmıştır. Geliştirilen yöntemde önce referans cevaptaki kelimeler eş anlamlılarıyla genişletilmiştir. Daha sonra Kosinüs benzerliği, Jaccard benzerliği ve Dice benzerlik ölçülerine göre karşılaştırılmıştır. Makalede sorular; politika, yaşam tarzı, spor ve teknoloji olmak üzere 4 kategoriye ayrılmıştır. Metinler DDİ metin ön işleme aşamalarından geçirilerek benzerlik ölçülerinin belirlenmesine hazır hale getirilmiştir. Çalışma sonuçları incelendiğinde, benzerlik ölçülerinin performansı kategoriden kategoriye değişiklik gösterdiği görülmektedir.

Başka bir çalışmada öğrencilerin açık uçlu maddelere verdiklerin cevapların manuel olarak değerlendirilmesinin güçlüğüne dikkat çekilmiş ve bu alanda doğal dil işleme ve makine öğrenmesi yaklaşımlarının kullanımının gerekliliğine vurgu yapılmıştır. Metin madenciliği ve doğal dil işleme tekniklerinin kullanıldığı makalede iki aşamalı bir yaklaşım sunulmuştur. İlk aşamada cevaplar benzerlik ölçülerine göre değerlendirilmektedir. İkinci aşamada ise bu sonuçlardan yararlanılarak makine öğrenmesi modelleri eğitilmektedir. Makalede TF-IDF, Bag of Words ve Word2vec metin temsil teknikleri kullanılmıştır. Benzerlik ölçüsü olarak kelime taşıma ve kosinüs benzerliği kullanılmıştır. Sınıflandırma tekniği olarak ise Çoklu Naif Bayes tekniğinden yararlanılmıştır. Araştırma sonucunda kelime taşıma yaklaşımının kosinüs benzerliğine göre daha iyi performans gösterdiği vurgulanmıştır. Ayrıca araştırmacılar tarafında önerilen algoritmaların %88 oranında doğruluk sağladığı görülmektedir (Bashir, Arshad, Javed, Kryvinska & Band, 2021).

Doğal dil işlemenin en yaygın kullanıldığı dil İngilizce dilidir. Bunun ana nedenlerinden biri geniş çapta açık uçlu metin verisi içeriklerinin genelde İngilizce dilinde olmasıdır. Yukarıda bahsi geçen birçok araştırma Kaggle (2012) platformunun kamuya açık olarak sunduğu, öğrenciler tarafından yazılan İngilizce metinler üzerinde

yürütülmüştür. Ancak son yıllarda yapılan çalışmalar incelendiğinde diğer dillerde de doğal dil işleme çalışmalarının yapılmakta olduğu görülmektedir. Bu çalışmalardan biri Brezilya Portekizcesi için yapılan çalışmadır. Bu çalışmada Brezilya lise öğrencileri tarafından çevrimiçi bir ortamda yazılmış makalelerden yararlanılmıştır. Puanlama kriteri olarak ENEM adlı ulusal lise sınavı kriterleri kullanılarak 5 farklı yetenek üzerinden uzmanlar tarafından manuel puanlama yapılmıştır. Makale konuları sağlık, spor, kültürel etkinlik, sahte haber, insan hakları, gibi konulardan oluşmaktadır. Çalışma sonucunda QWK değeri 0.51 düzeyinde elde edilmiştir. Yani makine-insan puanlayıcı arasındaki uyumun orta seviyede olduğu görülmektedir (Marinho vd., 2021).

Mathias, Murthy, Kanojia & Bhattacharyya (2021) çalışmalarında eğitim verisi olmadan sıfırdan otomatik puanlama yapmaya imkan sağlayan bir model geliştirmiştir. Bu modelde önceden eğitim verisi olmaksızın yeni yazılmış konuların otomatik puanlanması hedeflenmektedir. Çalışma ayrıca kişinin göz hareketleri ile bireyin metni nasıl anladığına ilişkin bilgiler sunmaktadır. Bir kelime üzerinde geçirilen süre, gözün daha önce gördüğü kelimeye tekrar dönmesi, bir kelimenin kaç kez okunduğu veya kelimenin tamamen atlanıp atlanmadığı gibi durumlar göz hareketleri parametrelerini oluşturmaktadır. Çalışma sonucunda ortalama QWK değeri göz hareketi olmadan 0.44, göz hareketi ile 0.49 olmuştur. Göz hareketleri eklendiğinde model performansında %5 artış olduğu gözlemlenmiştir (Mathias vd., 2021).

Zhang ve Yuan (2022) tarafından yapılan çalışmada İngilizce için var olan kompozisyon puanlama yaklaşımlarındaki eksiklerin giderilmesi amacıyla yeni bir yaklaşım önerilmiştir. Bu amaç doğrultusunda doğal dil işleme ve makine öğrenmesi tekniklerini kullanarak bir otomatik metin puanlama modeli geliştirmişlerdir. Bu modelde N-gram ve karar ağacı algoritmalarından yararlanılmıştır. Ayrıca modeli eğitmek için de rastgele orman algoritması kullanılmıştır. Model farklı öğrenme hızlarında test edilmiştir. Modelin doğruluk oranı 0.81 olarak hesaplanmıştır. Çalışmada İngilizce makale puanlamada makine öğrenmesi ve doğal dil işleme tekniklerinin uygulanabilir olduğu vurgulanmıştır (Zhang & Yuan, 2022).

Günümüzde doğal dil işlemenin birçok farklı dilde geliştirilmekte olduğu görülmektedir. Bu dillerden biri de Arapça'dır. Arapça, İngilizce gibi dillere göre çok daha karmaşık bir yapıya sahiptir. Arapça, kullanıldığı bölgelere göre bile değişiklik göstermesi, bir kelimenin birden fazla çeşidinin olması, her kelimenin üç harfli bir kökünün olması ve bunun da kök çıkarma (lemmatizasyon) işlemini zorlaştırması gibi güçlükler barındırmaktadır. Çalışmada bu nedenle Arapça dilinde DDİ araştırmalarının

geri kaldığı ifade edilmiştir. Arapça'daki bu zorluklar sebebiyle n-gram ve kelime çantası gibi geleneksel kelime temsil yöntemlerinin de bağlamdan kopmalara sebep olduğu görülmüştür. Çalışmada Arapça kısa cevapların puanlanmasına ilişkin derin öğrenmeye dayalı bir sistem geliştirilmiştir. Tranformer tabanlı modeller olan BERT ve ELECTRA gibi modellerin Arapça dil işlemede oldukça başarılı olduğu vurgulanmıştır. Çalışma sonucunda puanlayıcılar arası uyum ELECTRA ile 0.78, BERT ile 0.77 QWK olarak elde edilmiştir (Nael vd., 2022).

Açık uçlu maddelerin otomatik puanlanmasına ilişkin kullanılan yöntemlerin incelendiği bir başka çalışmada, derin öğrenme yaklaşımları sinir ağı tabanlı, dikkat tabanlı ve dönüştürücü tabanlı yaklaşımlar olarak üç kategoride ele alınmıştır. Metin benzerliğini ölçmek ve cevapları puanlamak amacıyla sinir ağlarından yararlanılmaktadır. Dikkat tabanlı yaklaşımlar ise; bir dikkat mekanizması kullanarak, metinler arası ilişkileri öğrenmeye çalışmaktadır. Örneğin; tranformer (dönüştürücü) modelleri dikkat tabanlı modellerdir. Üçüncü yaklaşım dönüştürücü tabanlı yaklaşımlardır. Bu modeller geniş metin verisi kullanarak çok yönlü bir dil algılama özelliği gösterirler. BERT modeli tranformatör tabanlı bir yaklaşımdır. Çalışmada bu modellerin genel özelliklerine değinilmiştir. Çalışma sonucunda açık uçlu puanlama için transformer tabanlı modellerin, önceki sinir ağı tabanlı modellerinin diğer yöntemlere göre daha iyi performans sergilediği sonucuna ulaşılmıştır. Çalışmada açık uçlu maddelerin puanlanması noktasında gelişen son teknoloji ürünleri olan GPT2, GPT3, T5 ve XLNET gibi son transformer araçlarının da keşfedilmesi gerekliliği vurgulanmıştır (Ahmed vd, 2022).

Kaggle (2012) platformunda düzenlenen otomatik puanlama yarışmasından elde edilen verilerle yürütülen bir çalışmada kelime gömme (word embedding) yolu olarak Word2Vec yöntemi kullanılmıştır. Verilerin %80'i eğitim, %20'si test verisi olarak ayrılmıştır. Bu aşamada çapraz doğrulama (cross validation) kullanılmıştır. Böylece veri setindeki her bir alt küme test verisi olarak kullanılmış ve doğrulaması yapılmıştır. Çalışmada LSTM ve BLSTM yöntemleri kullanılmıştır. LSTM algoritmasında QWK değerinin ilk denemede %73.53 olduğu son denemede ise %95.80'e ulaştığı görülmektedir. Bu da makine ve insan puanlayıcı arasındaki uyumun oldukça yüksek olduğunu göstermektedir (Ibrahim, Elfakharany & Hamed, 2022).

Kaggle (2012) platformundaki verilerle yapılan başka bir çalışmada, öğrenci makalelerinin otomatik değerlendirildiği bir model geliştirilmiştir. Çalışmada veri seti olarak Kaggle (2012) platformunda Hewlett Vakfı tarafından sağlanan 7 ve 10. Sınıf

öğrencilerin yazdığı makaleler kullanılmıştır. Veri setinde sekiz konu başlığında toplam 12.000 makale yer almaktadır. Çalışmadaki sistem öğrenci cevaplarındaki intihal oranı da tespit edilebilecek şekilde geliştirilmiştir. Öğrencilerin makaleleri yazarken herhangi bir internet sitesinden kopya alıp almadığı tespit edilmektedir. Geliştirilen sistemde LSTM ve çapraz doğrulama (K-fold) kullanılmıştır LSTM algoritması ile elde edilen QWK değeri 0.91 olarak hesaplanmıştır. Çalışmada ayrıca twitterdan veri çekilerek duygu analizi ile desteklenmiştir. Duygu analizinde öğrencilerin bir konu hakkındaki görüşlerinin olumlu veya olumsuz oluşu sınıflanmıştır. Çalışmanın duygu analizi bölümünde 10.000 tweetin 5000'i pozitif 5000'i negatif yöndedir. Modeli eğitmek için 7000 tweet, modeli doğrulamak için 3000 tweet kullanılmıştır. Duygu analizinde Naive Bayes kullanılmış ve %99.4 oranında bir doğru sınıflama elde edilmiştir (Sadanand, Guruvyas, Patil, Acharya & Suryakanth, 2022).

Otomatik puanlamanın metin verilerinin de ilerisine giderek grafik ve çizim puanlamaya kadar uzandığı görülmektedir. Bu çalışmada TIMMS 2019 grafik verilerinden yararlanılarak bir yapay sinir ağı yaklaşımıyla otomatik puanlama denemesi yapılmıştır. Çalışmada evrişimsel sinir ağları (CNN) ve beslemeli sinir ağları (FFN) da karşılaştırılarak en iyi performansı gösteren sinir ağının CNN olduğu sonucuna ulaşılmıştır. CNN modelleri görüntü verileri %97.71 oranında doğru sınıflandırdığı ve uygun puanlama kategorisine yerleştirebildiği ifade edilmiştir. Ayrıca CNN modellerinin insan puanlayıcıların yanlış değerlendirdiği görüntü verilerini de doğru değerlendirdiği görülmüştür. Makale sonucunda görüntü yanıtlarının otomatik puanlanmasının mümkün olduğu ve bu yaklaşımın geniş ölçekli testlerde kullanılabileceği öne sürülmüştür. Böylece insan gücü, maliyet gibi durumlar ortadan kaldırılabileceği ifade edilmiştir (Davier vd., 2022).

Metin verilerinin otomatik puanlanmasına ilişkin karşılaşılan en büyük problemlerden biri uygun veri yetersizliğidir. Bir diğeri ise ulaşılan verilerin puan kategorilerine dağılımlarındaki dengesizlik durumudur. Bu bağlamda her iki probleme de çözüm sunabilecek bir çalışma yapılmıştır. Bu çalışmada GPT-4 otomatik puanlama için dengesiz verileri düzenleme amacıyla kullanılmıştır. Çalışmada öğrencilerin iki soruya verdikleri yanıtlar toplanmış ve GPT-4'ün bu yanıtlara benzer yanıtlar üretmesi sağlanmıştır. Özellikle azınlıkta olan cevaplar için DistillBERT modeli eğitilmiştir. Çalışma sonucunda üretilmiş verilerin doğruluk, hassasiyet duyarlılık ve F1 skorlarında artış sağladığı vurgulanmıştır. SMOTE gibi veri dengeleme yaklaşımlarında öğrenci cevaplarının bağlamının tam olarak yansıtılamayacağı ifade edilmiştir. Otomatik

puanlama çalışmalarındaki veri dengesizliği sorunu için GPT-4'ü kullanarak veri arttırma işlemi yapılabileceği sonucuna ulaşılmıştır (Fang vd., 2023).

Tanveer, Adam, Khan, Ali ve Shakoor (2023) çalışmalarında makine öğrenme algoritmasının çeşitli veri setleri üzerindeki performansını incelemişlerdir. Çalışmada metin sınıflandırma, veri seti karmaşıklığı, sınıf dengesizliği ve algoritmaların genelleme yetenekleri üzerinde durulmuştur. Derin öğrenme, karar ağaçları ve destek vektör makineleri çalışmada incelenen makine öğrenmesi algoritmalarıdır. Derin öğrenme: doğruluk (accuracy): %85.2, hassasiyet (precision): 0.87, geri çağırma (recall): 0.84, f1 skoru: 0.85 değerlerini alırken; karar ağaçları: doğruluk (accuracy): %78.6, hassasiyet (precision): 0.75, geri çağırma, (recall): 0.80 değerlerini almıştır. Çalışmada derin öğrenme algoritmalarının DDİ alanında daha etkili sonuçlar verdiğini göstermiştir. Karar ağaçlarının basit görevlerde daha hızlı ve yorumlanabilir sonuçlar verdiği, destek vektör makinelerinin ise denge sağlayan bir seçenek olduğu sonuçlarına ulaşılmıştır.

Otomatik puanlama yaklaşımlarının doğruluk, adalet ve genellenebilirlik bağlamında incelendiği bir çalışmada 25.000'den fazla kompozisyonun yer aldığı bir veri seti kullanılmıştır. Bu çalışmada dokuz farklı AES modeli karşılaştırılarak en iyi performansı gösteren modeller belirlenmiştir. Araştırma sonucunda modellerin QWK değerleri şu şekildedir: BERT: 0.81, CNN-LSTM-ATT: 0.78, R2BERT: 0.77, SVM: 0.74, SKIPFLOW-LSTM: 0.73. Buna göre en yüksek performans BERT modelinde elde edilmiştir (Yang vd., 2024).

Otomatik puanlama yaklaşımlarından bir diğeri ise Ait Khayi ve Rus (2024) tarafından yapılan çalışmada ele alınmıştır. Bu çalışmada söylem dışı bilgiler de otomatik puanlamaya dahil edilerek anlama dayalı otomatik puanlama yaklaşımı benimsenmiştir. Bu araştırmada Otomatik Öğrenci Değerlendirme Ödülü (ASAP) veri kümesi kullanılmıştır. Model değerlendirme metriği olarak Ağırlıklandırılmış kappa (QWK) kullanılmıştır. Geliştirilen modelin QWK değerlerinin 0.81 ve 0.79 olduğu belirtilmiştir.

Sun ve Wang (2024) çalışmalarını 9000 makaleden oluşan ELLIPSE ve gerçek sınav puanlarını içeren IELTS veri setleri üzerinde yürütmüşlerdir. Çalışmada kullanılan modeller RoBERTa ve DistilBERT modelleridir. Çalışmada farklı boyutları değerlendirmek amacıyla çoklu regresyon teknikleri kullanılmıştır. Çalışmada ayrıca bağlamsal anlama ve model performansını artırmak için karşılaştırmalı öğrenme kullanılmıştır. Yani metinler ek bilgilerle (konu, tür, gereksinimler) birlikte modellenmiştir. ELLIPSE veri seti sonuçları incelendiğinde her iki modelin de tutarlı

sonuçlar verdiği görülmektedir: Ortalama kesinlik: 0.85, F1 Skoru: 0.88, QWK: 0.83., IELTS veri seti sonuçlarında ise ortalama QWK: 0.85 olarak elde edilmiştir.

Büyük dil modellerinin bireylere gerçek zamanlı geri bildirim vererek bireysel öğrenmeyi destekleyen bir sistem üzerine çalışma yapılmıştır. Bu çalışmada İngilizceyi yabancı dil olarak alan 33 öğrenci ve 21 İngilizce uzmanının geri bildirimleri analiz edilmiştir. Geri bildirimlerin genellikle olumlu, yüzeysel, geri bildirimin detay ve doğruluğunun yetersiz olması sınırlanmalarına rağmen geleneksel yöntemlere göre daha uygun ve doğru geri bildirim verdiği görülmüştür. Çalışma sonucunda öğrencilerin rubrikleri anlama ve özgüvenlerinde artış gözlenmiştir. Büyük dil modellerinin gerçek zamanlı öğrenme durumlarını tespit etme ve bireysel öğrenmeyi desteklediği sonucuna ulaşılmıştır. Eğitimde önemli bir potansiyele sahip olduğu ileri sürülmüştür (Han vd., 2024).

Günümüzde doğal dil işleme ile ilgili son gelişmelere göre Open AI tarafından tanıtılan ve neredeyse tüm internet metinleri ile eğitilmiş doğal dil işleme modelleri bulunmaktadır. Son yıllarda GPT 3 ve Chat GPT gibi modeller insan dilini çok iyi anlayıp analiz edebilecek seviyeye ulaşmıştır. Yani makine insan eliyle eğitilerek tıpkı insanlar gibi özgün metinler ortaya koymaya başlamıştır. Bu bağlamda otomatik yazılı puanlamada GPT 3,5 modelinin incelendiği bir çalışmada GPT 3,5'in mevcut halinde öğrenci cevaplarını doğru puanlamadığı görülmüştür. Çünkü öğrencilerin kurduğu cümleler Wikipedia gibi internet kaynaklarıyla benzer yapıda cümleler değildir. Çalışmada GPT 3,5 üzerinde ince ayar çalışması yapılarak öğrenci cevapları puanlatılmıştır. Sonuçlar GPT 3,5 modelinin ince ayar yapılmış halinin BERT modeline göre daha iyi performans sergilediği yönündedir (Latif & Zhai, 2024).

Literatürde yer alan çalışmalar incelendiğinde Türk dili ve benzer yapıdaki dillerle ilgili yapılan çalışmaların kısıtlı olduğu görülmektedir. Kullanılan yöntemler bağlamında incelendiğinde makine öğrenmesi modellerinin sıklıkla kullanıldığı yapay sinir ağları modellerinin de bunu takip ettiği görülmektedir. Bu tez çalışması Türk dilinde otomatik puanlama uygulaması yapılan bir çalışmadır. Diğer çalışmalardan farklı olarak bu tez çalışmasında Tf/idf metin vektörleştirme yaklaşımı ile transformatör tabanlı bir yaklaşım olan BERT model Türk dili üzerinde uygulanarak Türk dili için en uygun vektörleştirme yaklaşımı belirlenmeye çalışılacaktır. Çalışmada doğal dil işleme yöntemlerinden yararlanılarak metin vektörleştirme yoluyla makine öğrenmesi yöntemleri (NB, LR, DT, RF, GB) ve BERT modeli karşılaştırılmıştır. Ayrıca makine ve insan puanlayıcı

arasındaki puanlayıcı güvenirliği QWK ile karşılaştırılmıştır. Çalışmada makine öğrenmesi model performanslarını otomatik karşılaştıran bir uygulama da geliştirilmiştir.

# BÖLÜM 3

# YÖNTEM

## Araştırmanın Modeli

Bu çalışma açık uçlu soruların otomatik puanlanmasına ilişkin NB, LR, DT, RF, GB ve BERT yöntemleri ile sınıflama doğruluğunun tespit edilmesine yönelik var olan durumu tanımlaması, puanlayıcılar arası güvenirliğin belirlenmesi ve farklı makine öğrenmesi yaklaşımlarının model performansını karşılaştırması bakımından nicel araştırma yöntemlerinden ilişkisel tarama araştırması olarak görülmektedir (Bryman, 2016).

## Çalışma Grubu

Araştırmanın örneklemi ABİDE 2016 sınavına katılım sağlayan ve Türkçe ortak soruları yanıtlamış olan 8. Sınıf öğrencilerden oluşmaktadır. Türkiye'de ABİDE sınavına katılan öğrencilerden 2000 öğrenci araştırmanın çalışma grubunu oluşturmuştur. Çalışma grubunu oluşturan 2000 öğrenci toplamda 9 açık uçlu maddeyi cevaplandırmıştır. Bu araştırmada öğrenci cevaplarının OCR sistemleri ile tam ve eksiksiz bir şekilde okunup bilgisayar ortamına aktarılamaması sebebiyle veriler manuel olarak girilmiştir.

## Verilerin Toplanması

Bu çalışma Türkiye genelinde yapılan ABİDE 2016 sınavı açık uçlu ortak Türkçe maddeleri ile yürütülmüştür. Çalışma verileri Ölçme Değerlendirme ve Sınav Hizmetleri Genel müdürlüğünden talep edilerek Milli Eğitim Bakanlığı izni dahilinde alınmıştır. Veriler JPEG formatında alınmış olup bilgisayar ortamına manuel olarak işlenmiştir. Zaman kısıtlaması sebebiyle mevcut 9 maddeden 5'i hem iki hem üç kategorili maddeleri örnekleyecek şekilde seçilerek çalışma 5 madde üzerinden yürütülmüştür. Toplamda 2000 öğrencinin 5 maddeye verdikleri yaklaşık 10.000 uzun ve kısa yanıt bilgisayar

ortamına girilmiştir. Bu maddeler için hesaplanan madde güçlük (p) değerleri şu şekildedir: Madde 2 için 0.95, Madde 7 için 0.68, Madde 8 için 0.65, Madde 10 için 0.57 ve Madde 11 için 0.66 olarak hesaplanmıştır. Madde güçlük indeksleri 0 ile 1 arasında değişmektedir. Değer 0'a yaklaştığında madde zor, 1'e yaklaştığında maddenin kolaylaştığı bilinmektedir (Crocker & Algina, 1986). Maddeler incelendiğinde madde 2'nin çok kolay, madde 7, 8 ve 11'in kolay, madde 10'un ise orta güçlükte olduğu görülmektedir.

## Verilerin Çözümlenmesi

Bu çalışmada verilerin çözümlenmesinde Python (3.11.6) programlama dilinden yararlanılmıştır. Çalışma ortamı olarak PyCharm (2023.2.2-Professional Edition) ortamı kullanılmıştır. Çalışmanın BERT algoritması kısmında kullanılan modelin gereksinimleri gereği Google Colaboratory (Google, 2023) alanı kullanılmıştır. İlk etapta cevaplar orijinal haliyle ele alınarak ve Türkçe DDİ teknikleriyle veri ön işleme süreçlerinden yararlanılmaktadır. Otomatik puanlama modeli insan puanlayıcıların verdiği puanlar ile eğitilmiştir. Böylece insan puanlayıcıların nasıl puanlama yaptığı modele öğretilmiştir. Daha sonra, eğitiminde kullanılmayan test verileri otomatik olarak puanlanmıştır. Çalışmada Python kütüphanelerinden Pandas (McKinney, 2010), Scikit-Learn (Pedregosa vd., 2011) kütüphanesinden ve Türkçe paketlere ek olarak çalışma için gerekli diğer paketlerden yararlanılmıştır.

Makine öğrenmesi algoritmalarını karşılaştırabilmek için önce kapsamlı bir metin ön işleme yürütülmüştür. Bu işlemi gerçekleştirebilmek için MintLemon (Sarıgil ve Koç., 2023) Türkçe paketi kullanılmıştır. Metinlerdeki kelimeler ek ve köklerine ayrılmış, metin vektörleştirildiğinde vektörler arası benzerliğin yanıltıcı olmaması için durak sözcükler kaldırılmıştır. Bu kelimeler çok sık tekrar ettiği için metinlerin ortak noktalarının belirlenmesi, benzerliklerin bulunması noktasında doğru sonuçlara ulaşmak için önemlidir. Gereksiz boşluk ve noktalama işaretleri kaldırılmıştır. Bu noktada Türkçe dili için geliştirilen MintLemon kütüphanesi 'durak sözcükler' dizini kullanılmıştır.

Veriler çapraz doğrulama (K-Fold Cross Validation) yoluyla sırayla hem test verisi hem eğitim verisi olarak kullanılmıştır. Makine öğrenmesi algoritmaları için 10 kat çapraz doğrulama yapılmıştır. BERT modelinde ise eğitim sürecinin uzun sürmesi sebebiyle 5 kat çapraz doğrulama yapılmıştır. Daha sonra makine öğrenmesi

yöntemlerinden Naif Bayes, Lojistik Regresyon, Karar Ağacı, Rastgele Orman, Gradyan Artırma ve BERT algoritmalarından yararlanarak sınıflama doğruluğu sonuçlarına ulaşılarak hangi algoritmanın daha yüksek başarıyla tahminde bulunduğu ortaya konulmuştur. Araştırmanın son aşamasında bilgisayar ve insan puanlayıcı arasındaki puanlayıcılar arası uyuşma durumu Ağırlıklandırılmış kappa ile ölçülmüştür. Ağırlıklandırılmış kappa, Python'da scikit-learn kütüphanesinde yer alan "metrics" modülünden yararlanılarak hesaplanmıştır. Formülin uyarlanmış hali şu şekildedir (Cohen, 1968):

$$QWK = 1 - \frac{\sum_{i=1}^{K}\sum_{j=1}^{K}\omega_{ij}o_{ij}}{\sum_{i=1}^{K}\sum_{j=1}^{K}\omega_{ij}e_{ij}}$$

$K$, kategori sayısını ifade eder.

$o_{ij}$, gerçek ve tahmin edilen puanlar arasındaki gözlemlenen karışıklık matrisidir.

$e_{ij}$, tahmin edilen ve gerçek puanların dağılımlarından rastgele seçildiğinde beklenen matrisi temsil eder.

$\omega_{ij}$ ağırlık matrisi ise şu şekilde verilmektedir:

$$\omega_{ij} = \frac{(i-j)^2}{(K-1)^2}$$

QWK metriği insan puanlayıcı ile bilgisayarın verdiği puan arasındaki uyumu ölçmek için sadece kaç hata yapıldığına değil hataların ne kadar ciddi olduğuna odaklanmaktadır. Rastgele tahmin etkisini azaltarak daha doğru bir uyum ölçmektedir (Vanbelle, 2014). Landis ve Koch (1977) tarafından kappa katsayısının yorumlanmasına ilişkin belirlenen ölçütler Tablo 1'de verilmiştir.

Tablo 1

*Kappa Uyum Değer Aralığı*

| QWK Değeri | Anlamı |
|---|---|
| 0.00-0.20 | Çok Zayıf |
| 0.21–0.40 | Zayıf |
| 0.41–0.60 | Orta |
| 0.61–0.80 | Güçlü |
| 0.81–1 | Mükemmel uyum |

Makine ve insan puanlayıcıların puanları arasındaki uyumu gösteren QWK uyum değerleri ve yorumları Tablo 1'de görülmektedir. Klasik makine öğrenmesi modellerinin her biri için her madde de 10 kat çapraz doğrulama (cross-validation/k-fold) uygulanmıştır. Çapraz doğrulama makine öğrenmesi modellerinin performansını değerlendirmek için kullanılan bir tekniktir. Veri K eşit parçaya ayrılır kalan K-1 parça modelin eğitimi için kullanılmaktadır. Verinin tamamı eğitim ve test verisi olarak kullanılarak modelin performansı test edilmektedir. Böylece modelin genelleme yeteneği daha iyi anlaşılmaktadır (Dietterich, 1998; Kohavi, 1995). Makine öğrenmesi ve BERT model için incelenecek diğer metrikler; doğruluk, kesinlik, hassasiyet/duyarlılık ve F1-skoru değerleridir.

Doğruluk (Accuracy) değeri, modelin doğru sınıflandırdığı durumların toplam durumlara oranıdır.

Doğru pozitif tahminler: TP (True Positive)

Doğru negatif tahminler: TN (True Negative)

Yanlış pozitif tahminler: FP (False Positive)

Yanlış negatif tahminler: FN (False Negative) anlamına gelmektedir. Model performansı değerlendirme metriklerinin formülleri şu şekildedir (Scikit-learn Developers, n.d.):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Kesinlik (Precision) değeri modelin pozitif tahminlerinin ne kadarının gerçekten pozitif olduğunu gösteren ölçüttür.

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

Hassasiyet/Duyarlılık (Recall) değeri modelin bütün gerçek pozitifleri ne kadar doğru tespit edebildiğini ölçer.

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

F1-Skoru (F1-Score) kesinlik ve duyarlılığın dengeli bir şekilde değerlendirilmesini sağlar.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Bu araştırmada öğrencilerden alınan açık uçlu cevaplar sayısal verilere dönüştürülerek gerçek puanlayıcılardan alınan puanlarla eşleştirilmiştir. Bu eşleştirme sayesinde bilgisayar hangi kelimelerden oluşan cevapların doğru hangilerinin yanlış olduğu konusunda eğitilmiştir. Çapraz doğrulama yoluyla veriler sırayla hem test hem eğitim verisi olarak kullanılacak böylece sağlaması yapılmıştır.

DDİ süreçlerinin ilk basamağı veri ön işleme işlemlerinin yapılmasıdır. Sonraki aşamalarda daha fazla ayırt edilebilir ve yorumlanabilir konulara ulaşabilmek için bu adım çok önemlidir. DDİ projelerinde verilerin modellemeye hazırlanması için gerekli olan adımlar şu şekildedir: Çalışmada MintLemon Türkçe DDİ kütüphanesinden yararlanılmıştır. Her sözcük küçük harfe çevrilir (1). Noktalama işaretleri kaldırılır (2). Rakamlar kaldırılır (3). Satır sonları \n kaldırılır (4). Gereksiz kelimeler çıkarılır (Stopwords) (5). Tokenize (Jetonlaştırma) (6). Kelimeler ek ve köklerine ayrılır (Lemma ve Stemma) (7). Metinler vektöre dönüştürülerek metinler bilgisayarın anlayabileceği bir forma dönüştürülür (8). Böylece Türkçe metinlerin veri ön işleme adımları tamamlanmıştır.

Metin vektörleştirme adımında metinler sayısal değerlerle ifade edilmektedir. Metinler önceden işlendikten sonra, vektorizasyon yaklaşımı kullanılarak sayılara dönüştürülmektedir (Shin, 2021). Bu araştırmada metinler TF-IDF ve BERT yöntemiyle vektörleştirilmiştir. Metin vektörleştirme adımlarının ardından metinler artık sayılarla ifade edilebildiği için cevaba verilen puanlarla metinler eşleştirilmiştir. Daha sonra bu eşleştirmelerden yola çıkarak makine öğrenmesi algoritmaları yardımıyla sınıflama doğruluğu için Naif Bayes, Lojistik Regresyon, Karar Ağacı, Rastgele Orman, Gradyan Artırma ve BERT algoritmalarıyla sınıflama doğruluğu ele alınmıştır. Çalışmada ayrıca klasik makine öğrenmesi modellerine ilişkin bir uygulama geliştirilmiştir. Uygulama

Streamlit kütüphanesi kullanılarak geliştirilmiştir. Uygulama excel formatında veri yükleme imkanı sağlamaktadır. Ayrıca maddelerin otomatik puanlanmasına ilişkin en iyi performansı sağlayan algoritmanın seçilebilmesi için model performanslarını tablo halinde sunmaktadır. Uygulama bilgileri Ek 3'te yer almaktadır.

# BÖLÜM 4

## BULGULAR VE YORUMLAR

Bu bölümde önce maddelere ilişkin bilgilere yer verilecek daha sonra araştırma problemine ve alt amaçlara ilişkin bulgulara yer verilecektir.

### Birinci Alt Amaca İlişkin Bulgular ve Yorumlar: TF-IDF ve BERT Vektörleştirme Yöntemleri Kullanılarak Açık Uçlu Maddeler Puanlandığında, Farklı Makine Öğrenmesi Modellerinin Doğruluk Oranları

Klasik makine öğrenmesi modellerinden olan Naif Bayes, Lojistik Regresyon, Karar Ağacı Rastgele Orman, Gradyan Artırma algoritmalarının ABİDE Türkçe sınavına öğrencilerin verdikleri cevapları hangi doğruluk düzeyinde sınıfladığı her madde için ayrı ayrı verilmiştir. Bu bölümde ikinci maddeye ilişkin bulgulara yer verilmiştir.

Tablo 2

*İkinci Maddeye İlişkin Makine Öğrenmesi Algoritma Performansları Bulguları*

| Algoritma | Doğruluk (Accuracy) | Kesinlik (Precision) | Hassasiyet veya Duyarlılık (Recall) | F1-Skoru | Ağırlıklandırılmış Kappa Değeri (QWK Değeri) |
|---|---|---|---|---|---|
| Naif Bayes | 0.82 | 0.96 | 0.85 | 0.90 | 0 |
| Karar Ağacı | 0.93 | 0.95 | 0.98 | 0.96 | 0.23 |
| Rastgele Orman | 0.94 | 0.94 | 0.94 | 0.91 | 0 |
| Gradyan Artırma | 0.95 | 0.95 | 0.99 | 0.97 | 0 |
| Lojistik Regresyon | 0.94 | 0.94 | 1 | 0.97 | 0 |
| BERT | 0.92 | 0.8 | 0.78 | 0.76 | 0.72 |

Tablo 2'de yer alan ikinci maddeye ilişkin bulgulara bakıldığında Naif Bayes algoritmasının doğruluk değerinin 0.82 kesinlik 0.96 ve hassasiyet ve duyarlılık değerlerinin 0.85 ve F1 skorunun 0.90 olduğu görülmektedir. Karar Ağacı algoritmasının doğruluk değerinin 0.93 kesinlik 0.95 ve hassasiyet ve duyarlılık değerlerinin 0.98 ve F1 skorunun 0.96 olduğu görülmektedir. Rastgele Orman algoritmasının doğruluk değerinin

0.94 kesinlik 0.94 ve hassasiyet ve duyarlılık değerlerinin 0.94 ve F1 skorunun 0.91 olduğu görülmektedir.

Gradyan Artırma algoritmasının doğruluk değerinin 0.95 kesinlik 0.95 ve hassasiyet ve duyarlılık değerlerinin 0.99 ve F1 skorunun 0.97 olduğu görülmektedir. Lojistik Regresyon algoritmasının doğruluk değerinin 0.94 kesinlik 0.94 ve hassasiyet ve duyarlılık değerlerinin 1 ve F1 skorunun 0.97 olduğu görülmektedir.

İkinci madde diğer maddelerden farklı olarak QWK değeri ile dikkat çekmektedir. Makine insan arasında uyum görünmemektedir. Bu duruma sınıf dengesizliği sebep olabilmektedir. Bu durumun sebebi veriler incelendiğinde ortaya çıkmaktadır. İkinci madde, öğrenciler tarafından çoğunlukla doğru cevaplanmıştır. Tahmin edilen değerlerle gerçek değerlerin aynı olması sebebiyle QWK skoru hesaplanırken test ve tahmin değerlerinin tamamına yakınının aynı olması durumunda, QWK skoru sıfır olabilmektedir. Bu durumun birkaç çözümü bulunmaktadır. Rastgele veri artırma (azınlık veya çoğunluk sınıfı baz alarak), sınıf dengesini sağlama (SMOTE-Synthetic Minority Oversampling Technique- sentetik azınlık örnekleme tekniği), sınıf ağırlıklandırmasıdır (class weighting) (He & Garcia, 2009).

Fang ve diğerleri (2023) SMOTE gibi veri dengeleme çalışmalarının öğrenci cevabını tam olarak yansıtamadığını ifade etmektedir. Bu çalışmada klasik makine öğrenmesi modelleri ile BERT gibi transformatör tabanlı modeller karşılaştırılmaktadır. Bu nedenle veri dengesizliğine çözüm olarak da BERT modeli ile 5 kat çapraz doğrulama yapılarak QWK uyum değerleri hesaplanmıştır.

BERT modelin iyi bir performans (0.72) gösterdiği görülmektedir. Klasik makine öğrenmesi algoritmaları ile elde edilen sonuçlarla karşılaştırıldığında Tablo 3'te QWK değerlerinin 0 olarak hesaplandığı ve makine insan arasında uyumun olmadığı yönünde olduğu dikkat çekmektedir. Ancak bu değer veri dengesizliği sebebiyle bu şekilde hesaplanmıştır. Burada veri yapısının uyum değerlerinin hesaplanmasında ne derece önemli olduğu görülmektedir.

İkinci madde incelendiğinde öğrencilerin büyük çoğunluğunun maddeye doğru yanıt verdiği görülmektedir. Bu da geleneksel makine öğrenmesi yöntemlerinin eğitimi açısından yeterli olmadığı ve bu modellerin yanlış cevaplarla da eğitilmesi gerektiği anlamına gelmektedir. BERT modelde 5 kat çapraz doğrulama yapılarak veri sırayla hem eğitim hem test verisi olmuştur. Böylece verideki dengesizlikler tolere edilebilmiştir. Veri yapısı dengesiz de olsa BERT modelin, yeterli model performansı gösterdiği görülmektedir.

Tablo 3'te yedinci maddeye ilişkin geleneksel makine öğrenmesi yöntemlerine ilişkin bulgulara yer verilmiştir.

Tablo 3

*Yedinci Maddeye İlişkin Bulgular (3 Kategorili)*

| Algoritma | Doğruluk (Accuracy) | Kesinlik (Precision) | Hassasiyet veya Duyarlılık (Recall) | F1-Skor | Ağırlıklandırılmış Kappa Değeri (QWK Değeri) |
|---|---|---|---|---|---|
| Naif Bayes | 0.79 | 0.70 | 0.79 | 0.73 | 0.73 |
| Karar Ağacı | 0.87 | 0.88 | 0.87 | 0.87 | 0.85 |
| Rastgele Orman | 0.89 | 0.85 | 0.83 | 0.83 | 0.89 |
| Gradyan Artırma | 0.89 | 0.89 | 0.89 | 0.89 | 0.90 |
| LojistikRegresyon | 0.85 | 0.81 | 0.74 | 0.74 | 0.84 |
| BERT | 0.87 | 0.81 | 0.81 | 0.80 | 0.91 |

Yedinci maddeye ilişkin bulgular incelendiğinde Naif Bayes algoritmasının doğruluk değerinin 0.79 kesinlik 0.70 ve hassasiyet ve duyarlılık değerlerinin 0.79 ve F1 skorunun 0.73 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.73 olduğu yani makine insan arası puan uyumunun orta seviyede olduğu görülmektedir.

Karar Ağacı algoritmasının doğruluk değerinin 0.87 kesinlik 0.88 ve hassasiyet ve duyarlılık değerlerinin 0.87 ve F1 skorunun 0.87 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.85 olduğu ve bu değerin güçlü seviyede olduğu görülmektedir. Rastgele Orman algoritmasının doğruluk değerinin 0.89 kesinlik 0.85 ve hassasiyet ve duyarlılık değerlerinin 0.83 ve F1 skorunun 0.83 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.89 olduğu ve bu değerin mükemmel seviyede olduğu görülmektedir.

Gradyan Artırma algoritmasının doğruluk değerinin 0.89 kesinlik 0.89 ve hassasiyet ve duyarlılık değerlerinin 0.89 ve F1 skorunun 0. 89 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.90 olduğu ve bu değerin güçlü seviyede olduğu görülmektedir. Çoklu Lojistik Regresyon algoritmasının doğruluk değerinin 0.85 kesinlik 0.79 ve hassasiyet ve duyarlılık değerlerinin 0.75 ve F1 skorunun 0.76 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.84 olduğu ve bu değerin mükemmel seviyede olduğu görülmektedir.

BERT model sonuçları incelendiğinde modelin iyi bir performans gösterdiği görülmektedir. Yedinci madde 3 kategorili bir maddedir. Yani doğru. yanlış ve kısmi doğru cevaplardan oluşmaktadır. Diğer maddelerle kıyaslandığında 3 kategorili olan bu maddenin makine-insan arasındaki uyumunun yüksek olduğu görülmektedir. Klasik makine öğrenmesi modelleriyle karşılaştırıldığında ise QWK değerlerinin BERT modelinde artış gösterdiği görülmektedir.

Tablo 4

*Sekizinci Maddeye İlişkin Bulgular*

| Algoritma | Doğruluk (Accuracy) | Kesinlik (Precision) | Hassasiyet veya Duyarlılık (Recall) | F1-Skoru | Ağırlıklandırılmış Kappa Değeri (QWK Değeri) |
|---|---|---|---|---|---|
| Naif Bayes | 0.86 | 0.88 | 0.86 | 0.86 | 0.78 |
| Karar Ağacı | 0.91 | 0.91 | 0.91 | 0.91 | 0.84 |
| Rastgele Orman | 0.93 | 0.93 | 0.93 | 0.93 | 0.88 |
| Gradyan Artırma | 0.94 | 0.94 | 0.94 | 0.94 | 0.90 |
| Lojistik Regresyon | 0.94 | 0.94 | 0.94 | 0.94 | 0.88 |
| BERT | 0.95 | 0.94 | 0.94 | 0.94 | 0.93 |

Tablo 4'te Naif Bayes algoritmasının doğruluk değerinin 0.86 kesinlik 0.88 ve hassasiyet veya duyarlılık değerlerinin 0.86 ve F1 skorunun 0.86 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.78 olduğu yani makine insan arası puan uyumunun güçlü seviyede olduğu görülmektedir. Karar Ağacı algoritmasının doğruluk değerinin 0.91 kesinlik 0.91, hassasiyet ve duyarlılık değerlerinin 0.91 ve F1 skorunun 0.91 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.84 olduğu ve bu değerin güçlü seviyede olduğu görülmektedir.

Rastgele Orman algoritmasının doğruluk değerinin 0.93 kesinlik 0.93 ve hassasiyet ve duyarlılık değerlerinin 0.93 ve F1 skorunun 0.93 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.88 olduğu ve bu değerin mükemmel seviyede olduğu görülmektedir. Gradyan Artırma algoritmasının doğruluk değerinin 0.94 kesinlik 0.94 ve hassasiyet ve duyarlılık değerlerinin 0.94 ve F1 skorunun 0.94 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.90 olduğu ve bu değerin mükemmel seviyede olduğu görülmektedir. Lojistik Regresyon algoritmasının doğruluk değerinin 0.94 kesinlik 0.94 ve hassasiyet ve

duyarlılık değerlerinin 0.94 ve F1 skorunun 0.94 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.88 olduğu ve bu değerin mükemmel uyuma işaret ettiği görülmektedir. Ayrıca algoritmalarının birbirine yakın değerler verdiği görülmektedir.

BERT model için kesinlik, duyarlılık ve F1-skorları sırasıyla 0.93, 0.93 ve 0.93 olarak ölçülmüştür. Bu da sınıflar arası tutarlılığın yüksek düzeyde olduğunu ve modelin oldukça başarılı sonuçlar verdiğini göstermektedir.

Tablo 5

*Onuncu Maddeye İlişkin Makine Öğrenmesi Algoritmaları Bulgular*

| Algoritma | Doğruluk (Accuracy) | Kesinlik (Precision) | Hassasiyet veya Duyarlılık (Recall) | F1-Skoru | Ağırlıklandırılmış Kappa Değeri (QWK Değeri) |
|---|---|---|---|---|---|
| Naif Bayes | 0.82 | 0.80 | 0.91 | 0.85 | 0.69 |
| Karar Ağacı | 0.77 | 0.79 | 0.81 | 0.80 | 0.55 |
| Rastgele Orman | 0.81 | 0.80 | 0.89 | 0.84 | 0.68 |
| Gradyan Artırma | 0.81 | 0.79 | 0.90 | 0.84 | 0.67 |
| Lojistik Regresyon | 0.82 | 0.80 | 0.90 | 0.85 | 0.68 |
| BERT | 0.84 | 0.87 | 0.85 | 0.85 | 0.71 |

Naif Bayes algoritmasının doğruluk değerinin 0.82 kesinlik 0.80 ve hassasiyet ve duyarlılık değerlerinin 0.91 ve F1 skorunun 0.85 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.69 olduğu yani makine insan arası puan uyumunun orta seviyede olduğu görülmektedir. Karar Ağacı algoritmasının doğruluk değerinin 0.77 kesinlik 0.79 ve hassasiyet ve duyarlılık değerlerinin 0.81 ve F1 skorunun 0.80 olduğu görülmektedir. QWK değerinin 0.55 olduğu ve bu değerin orta seviyede olduğu görülmektedir.

Rastgele Orman algoritmasının doğruluk değerinin 0.81 kesinlik 0.80 ve hassasiyet ve duyarlılık değerlerinin 0.89 ve F1 skorunun 0.84 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.68 olduğu ve bu değerin güçlü seviyede olduğu görülmektedir. Gradyan Artırma algoritmasının doğruluk değerinin 0.81 kesinlik 0.79 ve hassasiyet ve duyarlılık değerlerinin 0.90 ve F1 skorunun 0. 84 olduğu görülmektedir. QWK uyum değerleri baz alındığında QWK değerinin 0.67 olduğu ve bu değerin güçlü seviyede olduğu görülmektedir. Lojistik Regresyon algoritmasının doğruluk değerinin 0.82 kesinlik 0.80 ve hassasiyet ve duyarlılık

değerlerinin 0.90 ve F1 skorunun 0.85 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.68 olduğu ve bu değerin güçlü seviyede olduğu görülmektedir. Algoritmalarının birbirine yakın değerler verdiği görülmektedir.

BERT modelin QWK ortalaması 0.71 olarak hesaplanmıştır. Bu da sınıflar arası tutarlılığın güçlü düzeyde olduğunu ve modelin genel olarak başarılı sonuçlar verdiğini göstermektedir.

Tablo 6

*On Birinci Maddeye İlişkin Bulgular*

| Algoritma | Doğruluk (Accuracy) | Kesinlik (Precision) | Hassasiyet veya Duyarlılık (Recall) | F1-Skoru | Ağırlıklandırılmış Kappa Değeri (QWK Değeri) |
|---|---|---|---|---|---|
| Naif Bayes | 0.81 | 0.78 | 0.98 | 0.87 | 0.44 |
| Karar Ağacı | 0.83 | 0.88 | 0.86 | 0.87 | 0.65 |
| Rastgele Orman | 0.86 | 0.85 | 0.96 | 0.90 | 0.68 |
| Gradyan Artırma | 0.86 | 0.86 | 0.93 | 0.90 | 0.68 |
| Lojistik Regresyon | 0.83 | 0.97 | 0.90 | 0.85 | 0.60 |
| BERT | 0.88 | 0.89 | 0.88 | 0.87 | 0.76 |

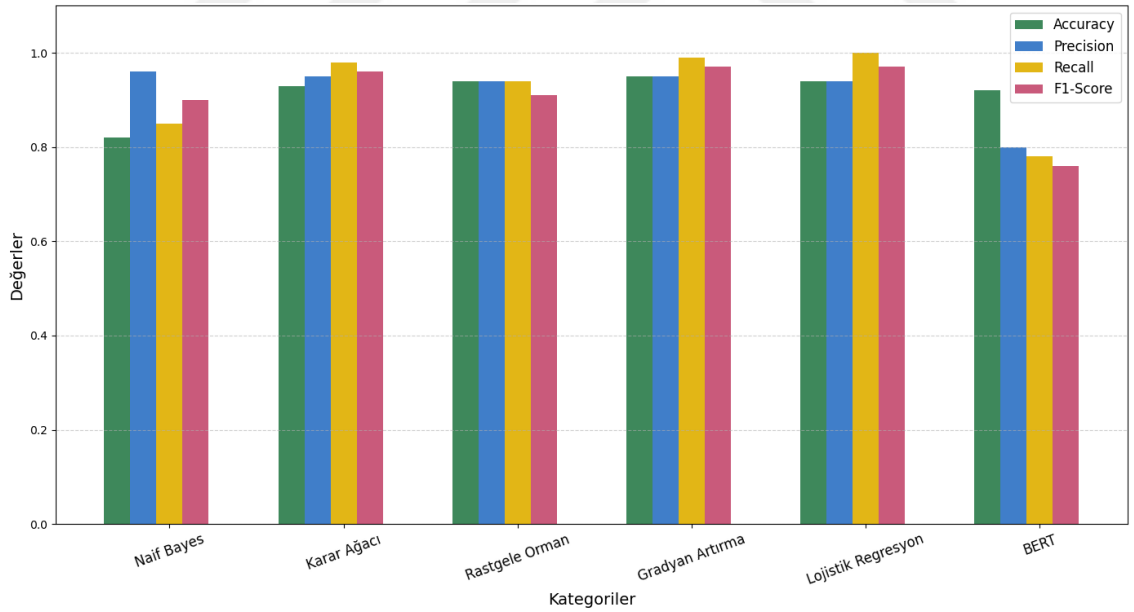
Tablo 6'ya göre Naif Bayes algoritmasının doğruluk değerinin 0.81 kesinlik 0.78 ve hassasiyet ve duyarlılık değerlerinin 0.98 ve F1 skorunun 0.87 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.44 olduğu yani makine insan arası puan uyumunun orta seviyede olduğu görülmektedir. Karar Ağacı algoritmasının doğruluk değerinin 0.83 kesinlik 0.88 ve hassasiyet ve duyarlılık değerlerinin 0.86 ve F1 skorunun 0.90 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.65 olduğu ve bu değerin güçlü seviyede olduğu görülmektedir.

Rastgele Orman algoritmasının doğruluk değerinin 0.86 kesinlik 0.85 ve hassasiyet ve duyarlılık değerlerinin 0.96 ve F1 skorunun 0.87 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.68 olduğu ve bu değerin güçlü seviyede olduğu görülmektedir. Gradyan Artıma algoritmasının doğruluk değerinin 0.86 kesinlik 0.86 ve hassasiyet ve duyarlılık değerlerinin 0.93 ve F1 skorunun 0.90 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.68 olduğu ve bu değerin güçlü seviyede olduğu görülmektedir. Lojistik Regresyon algoritmasının doğruluk değerinin 0.83 kesinlik 0.97 ve hassasiyet ve

duyarlılık değerlerinin 0.90 ve F1 skorunun 0.85 olduğu görülmektedir. Tablo 1'deki kappa uyum değerleri baz alındığında QWK değerinin 0.60 olduğu ve bu değerin orta seviyede olduğu görülmektedir. BERT model sonuçları incelendiğinde doğruluk 0.88, kesinlik 0.89, hassasiyet 0.88, F1skor 0.87 ve QWK 0.76 olarak hesaplanmıştır. BERT modelin hem sınıflandırma metrikleri açısından hem de QWK uyum değerleri açısından diğer modellerden daha iyi performans sergilediği görülmüştür.

## İkinci Alt Amaca İlişkin Bulgular ve Yorumlar: Klasik Makine Öğrenmesi Modelleri İle BERT Modeli, Sınıflandırma Doğruluk Metrikleri Açısından Karşılaştırılması

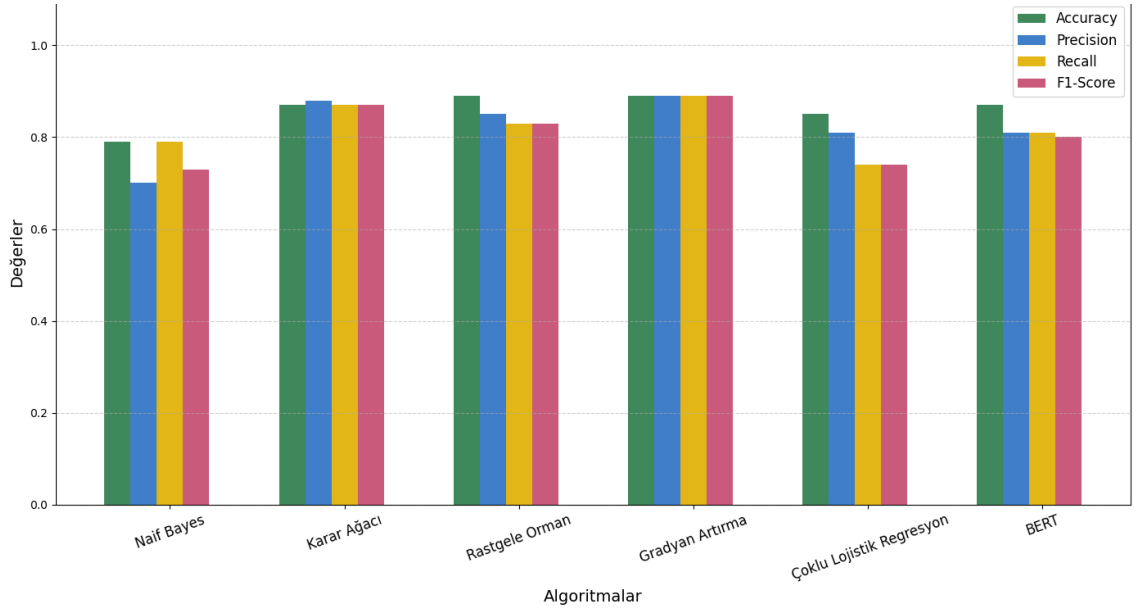
Aşağıdaki şekillerde açık uçlu maddelerin sınıflandırılmasında, klasik makine öğrenmesi modelleri ile BERT model sınıflandırma performans metrikleri karşılaştırması gösterilmektedir. Şekil 8'de ikinci maddeye ilişkin makine öğrenmesi ve BERT model metrikleri verilmiştir.




*Şekil 8.* İkinci Maddeye İlişkin Makine Öğrenmesi Modelleri ve BERT Model Karşılaştırması

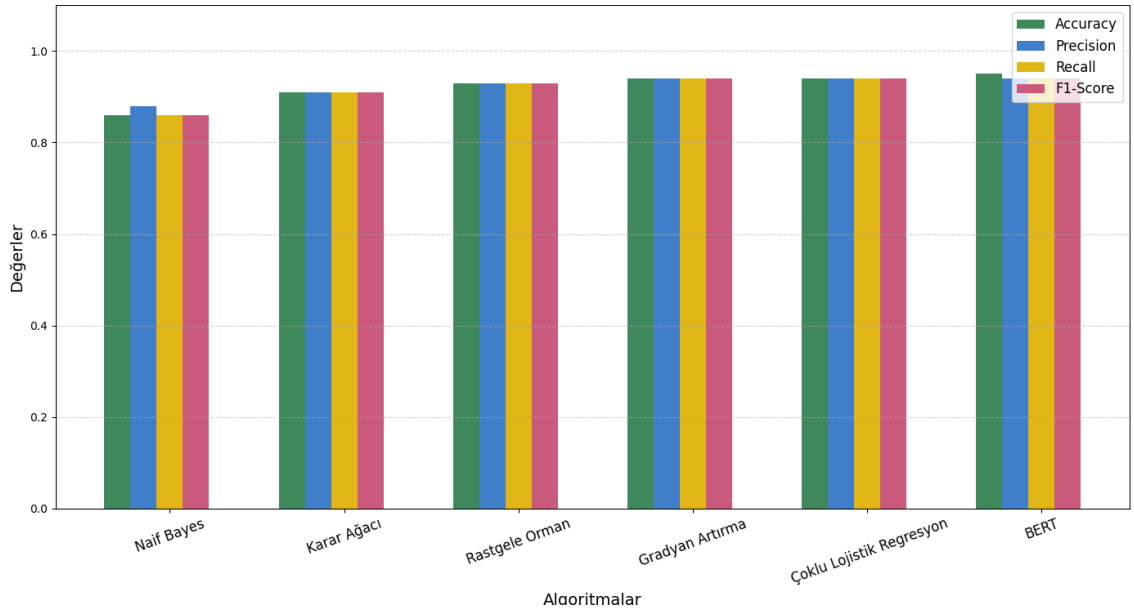
Şekil 8'de ikinci maddeye ilişkin makine öğrenmesi ve BERT model metrikleri karşılaştırıldığında Gradyan Artırma (0.95) ve Lojistik Regresyon algoritmalarıyla elde edilen doğruluk, Gradyan Artırma (0.95), Karar Ağacı (0.95) ve Naif Bayes (0.96)

kesinlik, Lojistik Regresyon (1), Gradyan Artırma (0.99) ve Karar Ağacı (0.98) duyarlılık, Gradyan Artırma ve Lojistik Regresyon (0.97) F1 skor ile oldukça yüksek değerler elde edildiği görülmektedir. Bu durum makine öğrenmesi modellerinin aşırı öğrenme (overfitting) yapmış olabileceğini göstermektedir. BERT modelde ise bu değerler sırasıyla 0.92, 0.8, 0.78, 0.76, 0.72 şeklindedir.



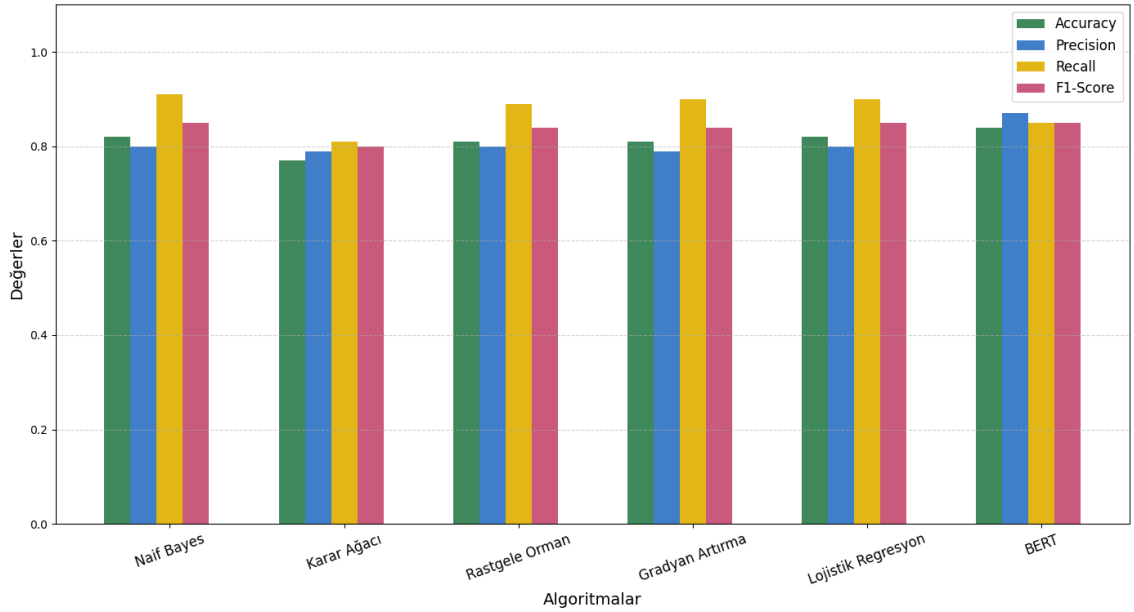

*Şekil 9.* Yedinci Maddeye İlişkin Makine Öğrenmesi Modelleri Ve BERT Model Karşılaştırması

Şekil 9'da yedinci maddeye ilişkin makine öğrenmesi modelleri ve Bert performansı verilmiştir. Buna göre modeller karşılaştırıldığında özellikle Gradyan Artırma algoritmasının performansı dikkat çekmektedir.



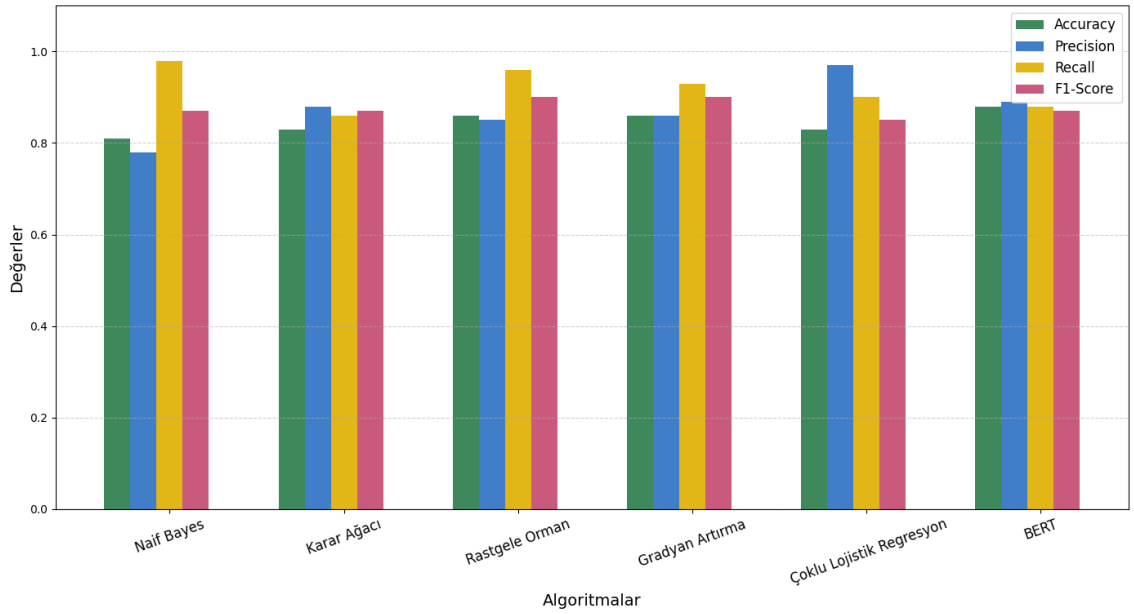

*Şekil 10.* Sekizinci Maddeye İlişkin Makine Öğrenmesi Modelleri ve BERT Model Karşılaştırması

Şekil 10'da sekizinci maddeye ilişkin model performansları yer almaktadır. Genel olarak tüm algoritmaların yüksek performans gösterdiği görülmektedir. Gradyan Artırma, Lojistik Regresyon ve BERT sınıflandırma metrikleri açısından aynı değerleri almıştır (0.94). Rastgele Orman algoritması da yakın değerler alarak bu değerleri takip etmiştir (0.93). Sekizinci madde için genel olarak tüm algoritmalar iyi bir performans sergilemiş birbirini destekler niteliktedir.




*Şekil 11.* Onuncu Maddeye İlişkin Makine Öğrenmesi Modelleri ve BERT Model Karşılaştırması

Şekil 11'de onuncu maddenin makine öğrenmesi modelleri ve BERT modelinin performans karşılaştırmasına ilişkin grafik yer almaktadır. Grafik incelendiğinde algoritmaların tüm performans metriklerinde yakın değerler aldığı görülmektedir. Ancak BERT modelin doğruluk, kesinlik, duyarlılık, F1 ve QWK (0.84, 0.87, 0.85, 0.85, 0.71) değerlerinin genel olarak makine öğrenmesi modellerinden yüksek olduğu görülmektedir. On birinci maddeye ilişkin model performansları Şekil 12'de görülmektedir.

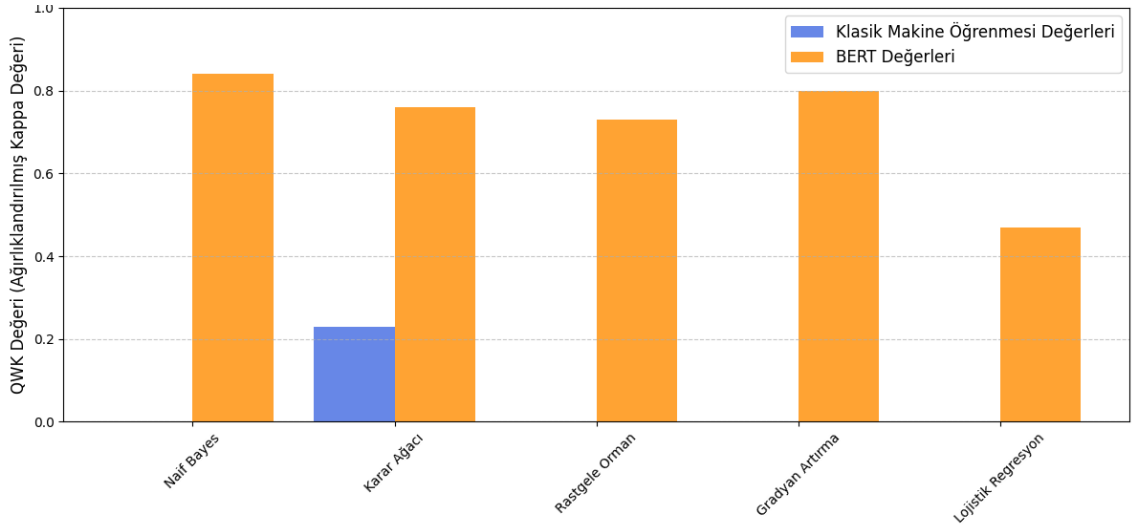



*Şekil 12.* On Birinci Maddeye İlişkin Makine Öğrenmesi Modelleri ve BERT Model Karşılaştırması

Şekil 12'deki model performansları incelendiğinde Gradyan Artırma algoritmasının sınıflama metrikleri açısından oldukça iyi performans sergilediği görülmektedir. Naif Bayes'in ise diğer modellere oranla daha düşük performans sergilediği görülmektedir.

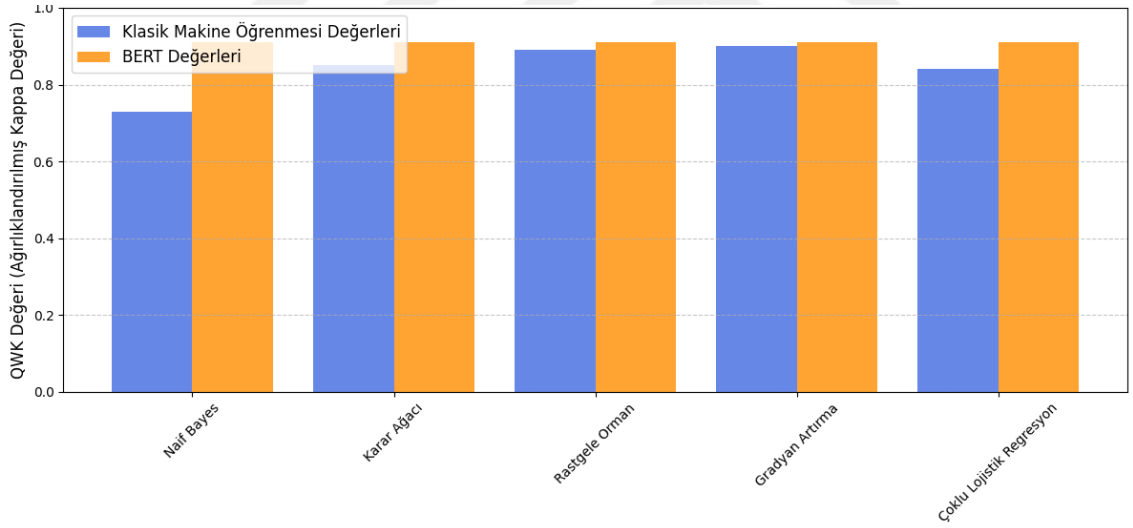
## Üçüncü Alt Amaca İlişkin Bulgular ve Yorumlar: TF-IDF Ve BERT Yöntemleriyle Vektörleştirilen Açık Uçlu Maddelerde İnsan ve Bilgisayar Puanlayıcıları Arasındaki QWK Uyumu

Aşağıdaki şekillerde açık uçlu maddelerde, klasik makine öğrenmesi modelleri ile BERT modeli arasında, insan ve bilgisayar puanlayıcıları arasındaki güvenirliği belirlemede kullanılan Ağırlıklandırılmış Kappa katsayısı (QWK) açısından model performansları gösterilmektedir.




*Şekil 13.* İkinci Maddeye İlişkin QWK Model Performansları

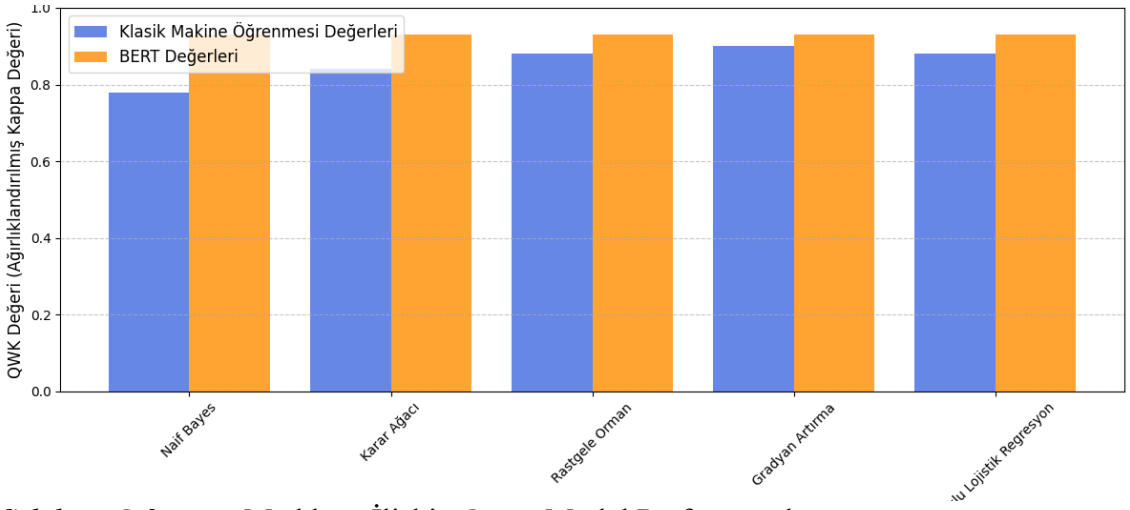
Şekil 13'te ikinci maddeye ilişkin QWK değerleri incelendiğinde makine-insan arası uyumun BERT model dışında sadece karar ağacı algoritmasında hesaplandığı onun da oldukça düşük (0.23) bir değere sahip olduğu görülmektedir.




*Şekil 14.* Yedinci maddeye ilişkin QWK model performansları

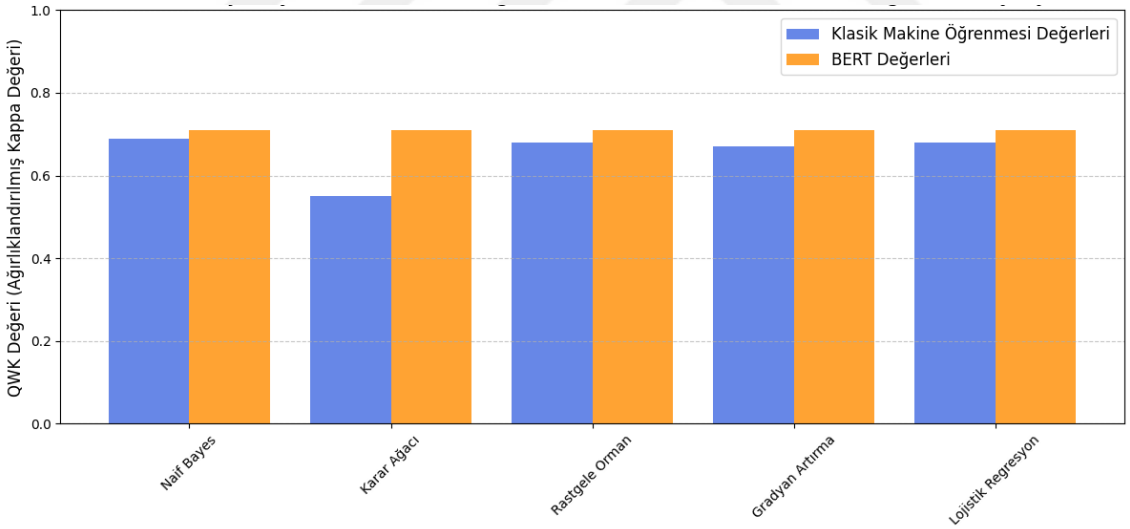
Şekil 14 incelendiğinde yedinci maddeye ilişkin QWK model performansları görülmektedir. Buna göre BERT modelin ortalama QWK değeri 0.91 iken diğer modellerin (sırasıyla 0.73, 0.85, 0.89, 0.90, 0.84) QWK değerleri BERT modelden daha düşüktür.

BERT model sınıflandırma metriklerinde Gradyan Artırma makine öğrenmesinin gerisinde kalsa da QWK uyum iyiliği değerlerine göre en iyi performansı gösteren model olmuştur. BERT modele en yakın değer Gradyan Artırma algoritmasıyla elde edilmiştir.



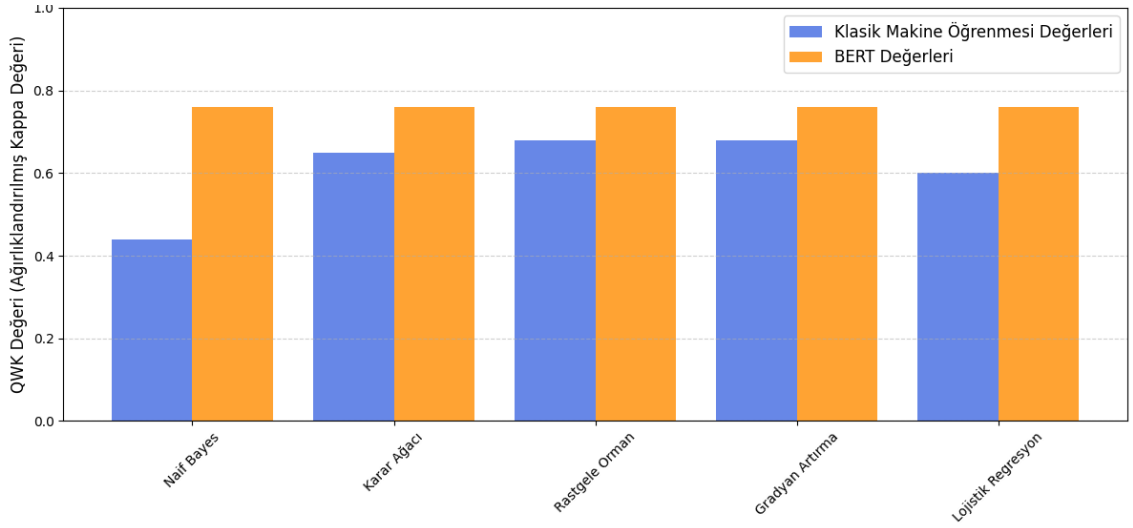

*Şekil 15.* Sekizinci Maddeye İlişkin QWK Model Performansları

Şekil 15'te sekizinci maddeye ilişkin QWK değerlerinin karşılaştırılması görülmektedir. Buna göre sırasıyla makine öğrenmesi modelleri 0.78, 0.84, 0.88, 0.90 ve 0.88 değerlerini alırken BERT model 0.93 değerini almıştır. En yüksek makine insan puanlayıcı uyumu Bert model ile sağlanmıştır.




*Şekil 16.* Onuncu Maddeye İlişkin QWK Model Performansları

Şekil 16 incelendiğinde onuncu maddeye ilişkin QWK değerleri görülmektedir. Buna göre BERT model makine öğrenmesi modellerine göre daha yüksek bir performans sergilemiştir. Bert modeli Naif Bayes, Lojistik Regresyon ve Rastgele Orman algoritmaları izlemiştir.




*Şekil 17.* On Birinci Maddeye İlişkin QWK Model Performansları

Son olarak Şekil 17'de on birinci maddeye ilişkin QWK değerleri incelendiğinde yine en yüksek makine insan uyumu BERT model ile elde edilmiştir (0.71). En düşük uyum Naif Bayes algoritmasına aittir (0.44). Rastgele Orman ve Gradyan Artırma ise aynı değeri almıştır (0.68).

Model performansları genel olarak incelendiğinde tüm maddeler için BERT model en iyi performansı gösteren model olmuştur. Ancak Gradyan Artırma algoritmasının BERT modele yakın değerler verdiği görülmektedir. Özellikle ikinci madde bazında incelendiğinde BERT modelin veri dengesizliğinden etkilenmemesi ve diğer makine öğrenmesi modelleri gibi aşırı öğrenme (overfitting) sorununun olmaması göz önüne alındığında BERT modelin diğer modellerden daha başarılı sonuçlar verdiği görülmektedir.

# BÖLÜM 5

# SONUÇLAR VE ÖNERİLER

Bu bölümde araştırmaya ilişkin sonuçlara ve bu bağlamda uygulama ve araştırmacılara önerilere yer verilmiştir.

## Sonuçlar

Modelin genel olarak iyi bir performans gösterdiği görülmektedir. Dengesiz veriler için önerilen yöntemlerden biri olan dönüştürücü temelli yaklaşımlardan BERT model çalışma kapsamında ele alınan diğer bir model olduğu için veri dengesizliği BERT model ile giderilmiştir.

Yedinci maddeye ilişkin bulgular incelendiğinde Gradyan Artırma algoritması**,** doğruluk, kesinlik, duyarlılık, F1-Skoru ve QWK değerlerinde en iyi performansı göstermiştir. Bu modelin insan puanlayıcılarla uyumlu sonuçlar üretebildiğini göstermektedir. Naif Bayes algoritmasının özellikle karmaşık metin verilerinde sınırlı performans gösterdiği vurgulanmıştır (McCallum & Nigam, 1998). Naif Bayes algoritması, performans açısından diğer algoritmaların gerisinde kalmıştır. Rastgele Orman algoritmasının iyi bir performans sergilediği görülmektedir. Breiman (2001), Rastgele Orman algoritmasının yüksek doğruluk ve düşük varyans sağladığını belirtmiştir. Karar Ağacı algoritması da aynı şekilde yüksek bir performans sergilemiştir. Ancak BERT modeli QWK değeriyle en yüksek performansı sergileyen algoritma olmuştur. Elde edilen bu sonucun literatürdeki diğer çalışmalarla (Krishnamurthy, vd., 2019; Nael vd., 2022; Ahmed vd., 2022) tutarlılık gösterdiği görülmektedir.

Sekizinci maddeye ilişkin modelin genel olarak iyi bir performans sergilediği görülmektedir. Modelin genelleme yeteneği oldukça yüksektir. Ortalama kesinlik değerleri modelin doğru pozitif sınıflarda ne kadar doğru tahmin yaptığını göstermektedir. Modelin yanlış pozitif oranının düşük olduğunu doğru sınıf tanımlama da başarılı olduğu görülmektedir. Duyarlılık değerleri modelin pozitif örnekleri doğru bir şekilde tespit ettiğini göstermektedir.

F1 skorlarına bakıldığında kesinlik ve duyarlılık arasında bir denge olduğunu göstermektedir. Ortalama F1 skoru modelin sınıflandırma görevinde dengeli bir performans sergilediğini göstermektedir. Son olarak QWK değerleri incelendiğinde etiketlerde model tahmin değerlerinin uyumlu olduğunun göstergesidir. Yani makine ve insan puanları arasında yüksek bir uyum olduğuna işaret etmektedir.

Bu tez çalışmasının sonucunda elde edilen tüm uyum değerleri incelendiğinde BERT modelin geleneksel makine öğrenmesi yaklaşımlarına göre tüm maddelerde daha iyi bir performans sergilediği görülmektedir. BERT gibi dönüştürücü modellerin sinir ağlarından daha iyi performans gösterdiğini belirtilmektedir (Ahmed vd., 2022). Bu sonuçlar literatürde incelenen diğer çalışmalarla tutarlılık göstermektedir (Kakkonen vd., 2005; Marinho vd., 2021; Nael vd., 2022; Ahmed vd., 2022; Sun & Wang, 2024; Yang vd., 2024).

Bu çalışmada elde edilen makine-insan puanlayıcı arasındaki uyum literatürde incelenen Nael ve diğerleri (2022) Marinho ve diğerleri (2021), Yang ve diğerleri (2024) tarafından yapılan çalışmalardan daha yüksek bir uyum oranına ulaşmıştır. Çapraz doğrulama katmanları artırıldıkça modellerin performansında artış gözlenmiştir. Verilerin dağılımının dengeli olması model performansı üzerinde etkilidir. Çünkü model yeterince doğru ve yanlış cevap üzerinde eğitildiğinde model puanlamayı öğrenmekte ve insan benzeri puanlar verme eğilimi artmaktadır. Bu tez çalışmasında öğrenci verilerinin tek tek manuel biçimde girilmesi ve veri dengesizliği gibi durumlarla karşılaşılmıştır.

Literatür incelendiğinde Mathias ve diğerleri (2021) eğitim verisi olmadan yaptığı otomatik puanlama denemesinde çalışmaya göz hareketleri modülünün eklenmesi ve eğitim verisi olmadan da sistemin geliştirilmiş olması önemlidir. Fakat QWK değerleri incelendiğinde uyumun düşük olduğu bu nedenle de eğitim verisinin model performansı üzerinde etkin bir role sahip olduğu görülmektedir. Eğitim verisi az veya dengesiz ise GPT modelleriyle veri artırma teknikleri uygulanarak model geliştirme çalışmaları yapılabilir (Fang vd., 2023).

Timms verileriyle grafikler üzerinden görüntü işleme teknikleriyle geliştirilen otomatik puanlama yaklaşımları da kullanılmaktadır (Davier vd., 2022). CNN teknikleri ile geliştirilen bu modellerin doğruluk oranlarının da yüksek olduğu görülmektedir. Türk dili için de benzer çalışmalar yapılabilir.

Türk diline benzer yapıdaki Japonca, Korece ve Fin dilinde yapılan çalışmalar incelendiğinde bu dillerin kendi özelliklerine özgü sistemlerinde geliştirildiği görülmektedir (Kakkonen vd., 2005; Ishioka & Kameda, 2006; Jang vd., 2014). Bu tez

çalışmasında da veri ön işleme adımlarında Türk diline özgü geliştirilen MintLemon (Sarıgil ve Koç, 2023) kütüphanesinden yararlanılmıştır.

Literatürdeki çalışmalar incelendiğinde model performansının karşılaştırıldığı ancak kullanıcıya deneme imkanı veren sitemlerin olmadığı görülmektedir. Bu tez çalışmasında algoritma performanslarını karşılaştırmak için bir örnek uygulama geliştirilmiştir. Uygulama araştırmacı tarafından Streamlit ile gerçekleştirilmiştir. Streamlit Python programlama dilinde kullanılabilen kullanıcılara uygulama yazma imkanı sağlayan bir araçtır (Streamlit, Inc., n.d.). Ek'ler bölümünde Ek 3'te uygulamaya dair bilgiler yer almaktadır. Tüm bu sonuçlar değerlendirildiğinde Türk dilinde otomatik puanlamanın uygun olduğu, geliştirilecek daha kapsamlı sistemlerle insan benzeri puanlamaların elde edilebileceği görülmektedir.

## Öneriler

### Uygulamaya Yönelik Öneriler

1. Milli Eğitim Bakanlığı'nın yaptığı ortak sınavlarda otomatik puanlama sistemleri kullanılabilir.
2. Üç kategorili yani doğru, yanlış ve kısmi doğru puanlamadan oluşan maddelerde hem klasik makine öğrenmesi algoritmaları hem BERT yüksek uyum sağlamıştır. Klasik makine öğrenmesi algoritmalarından Grandyan Artırma ve dönüştürücü tabanlı BERT model bu maddelerin otomatik puanlanmasında kullanılabilir.
3. İkinci maddede olduğu gibi veri dengesizliği olan maddeler için klasik makine öğrenmesi yerine otomatik puanlama için BERT model tercih edilebilir.

### Araştırmacılara Yönelik Öneriler

1. Bu tez çalışmasında klasik makine öğrenmesi modelleri ve Bert model sonuçları karşılaştırılarak Türkçe otomatik puanlama için en iyi performans gösteren yöntemler belirlenmeye çalışılmıştır. Araştırmacılar bu tez çalışmasında ele alınmayan diğer yöntemleri ele alabilir ve bu tez çalışmasından elde edilen bulgularla karşılaştırmalar yapabilir.

2. Çalışmada klasik makine öğrenmesi algoritmaları için TF-IDF sayısallaştırma yaklaşımı kullanılmıştır. Sonraki çalışmalarda farklı sayısallaştırma çalışmalarına yer verilebilir ve sonuçlar karşılaştırılabilir.
3. Bu tez çalışmasında öğrenci cevapları tek tek elle bilgisayar ortamına girilmiştir. Sonraki çalışmalarda yeni teknolojiler ışığında kağıt üzerindeki öğrenci cevapları bu teknolojik gelişmeler doğrultusunda bilgisayar ortamına geçirilebilir.
4. GPT ve Gemini gibi üretken yapay zeka araçları kullanılarak ve ince ayarlar yapılarak daha yüksek performans sağlayan otomatik puanlama sistemleri geliştirilebilir.
5. Sınıf dengesizliklerine yönelik farklı teknikler kullanılarak model performansları artırılabilir.
6. BERT model için 10 kat çapraz doğrulama ile model performansı gözlenebilir.
7. Gelecekteki çalışmalar, hiperparametre optimizasyonu için “Grid Search” veya “Random Search” gibi yöntemler kullanarak model performansını daha da iyileştirebilir.

## KAYNAKLAR


Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). MultinomialNaive Bayes classification model for sentiment analysis. *IJCSNS International Journal of Computer Science and Network Security, 19*(7), 62-66. Retrieved from https://www.researchgate.net/publication/334451164_Multinomial_Naive_Bayes_Classification_Model_for_Sentiment_Analysis

Adalı, E. (2016). Doğal dil işleme. *Türkiye Bilişim Vakfi Bilgisayar Bilimleri ve Mühendisliği Dergisi, 5*(2). Retrieved from https://dergipark.org.tr/tr/pub/tbbmd/issue/22245/238797

Ahmed, A., Joorabchi, A., & Hayes, M. J. (2022). On deep learning approaches to automated assessment: Strategies for short answer grading. In *14th International Conference on Computer Supported Education (CSEDU)* (pp. 85-94). Retrieved from https://www.scitepress.org/Papers/2022/110821/110821.pdf

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 715-725). Retrieved from https://typeset.io/pdf/automatic-text-scoring-using-neural-networks-1gx5bp3fh3.pdf

Akar, A. (2017). *Türk dili tarihi: Dönem-eser-bibliyografya* (12th ed.). İstanbul, Türkiye: Ötüken. Retrieved from https://www.otuken.com.tr/u/otuken/docs/turk_dili_tarihi.pdf

Alanoğlu, M. (2022). *Education and science.* Ankara, Türkiye: Efe Akademi.

Ait Khayi, N., & Rus, V. (2024). Automated essay scoring using discourse external knowledge. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24), Special Track on AI for Good* (pp. 7154-7160).

Bashir, M. F., Arshad, H., Javed, A. R., Kryvinska, N., & Band, S. S. (2021). Subjective answers evaluation using machine learning and natural language processing. *IEEE Access, 9,* 158972-158983. doi:10.1109/ACCESS.2021.3130902

Bezirhan, U., & Davier, M. (2023). Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence, 5,* 100161. doi:10.1016/j.caeai.2023.100161

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York, NY: Springer.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5-32. doi:10.1023/A:1010933404324

Bryman, A. (2016). *Social research methods* (5th ed.). Oxford University Press.

Bulut, O. (2022). Special issue on artificial intelligence, machine learning, and natural language processing applications in education. Journal of Applied Testing Technology, 23(SI1), 1–3.

Buchanan, B. G., & Shortliffe, E. H. (1984). *Rule-based expert systems*. Addison-Wesley.

Burkov, A. (2019). *The hundred-page machine learning book*. Andriy Burkov.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135-146. doi.org/10.1162/tacl_a_00051

Chamidah, N., Santoni, M. M., Irmanda, H. N., Astriratma, R., Tua, L. M., & Yuniati, T. (2021). Word expansion using synonyms in Indonesian short essay auto scoring. In *International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*. doi:10.1109/ICIMCIS53775.2021.9699374

Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning*. MIT Press. Retrieved from http://www.acad.bg/ebook/ml/MITPress-%20SemiSupervised%20Learning.pdf

Chaudhry, M. A., & Kazim, E. (2021). Artificial Intelligence in Education (AIEd): A high-level academic and industry note 2021. *AI and Ethics*, 2, 157-165. doi: 0.1007/s43681-021-00074-z

Chen, J., Chen, C., & Liang, Z. (2016). Optimized TF-IDF algorithm with the adaptive weight of position of word. In *2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016) Advances in Intelligent Systems Research* (Vol. 133).

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston. Retrieved from https://archive.org/details/introductiontoclassicalandmoderntesttheory/mode/2up

Coelho, A. M. L., da Silva, H. F., da Silva, L. A. C., Andrade, M. E., & Rodrigues, R. G. da S. (2023). Inteligência artificial: Suas vantagens e limites em cursos à distância. *Revista Ilustração*, 4(2), 23-27. doi:10.46550/ilustracao.v4i2.150

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220. doi: 10.1037/h0026256

Çelik, Ö., & Altunaydın, S. S. (2018). A research on machine learning methods and its applications. *Journal of Educational Technology & Online Learning*, 1(3), 25-40. doi: 10.31681/jetol.457046

Çelik, Ö., & Koç, B. C. (2021). TF-IDF, Word2Vec ve FastText Vektör Model Yöntemleri ile Türkçe Haber Metinlerinin Sınıflandırılması. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 23(67), 121-127. doi:10.21205/deufmd.2021236710

Çınar, A., İnce, E., Gezer, M., & Yılmaz, Ö. (2020). Machine learning algorithm for grading open-ended physics questions in Turkish. *Education and Information Technologies*. doi:10.1007/s10639-020-10128-0

Davier, M., Khorramdel, L., & Tyack, L. (2022). Scoring graphical responses in TIMSS 2019 using artificial neural networks. *Educational and Psychological Measurement, 83*(3). Retrieved from https://ilsa-gateway.org/papers/scoring-graphical-responses-timss-2019-using-artificial-neural-networks

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press. https://mml-book.com

Deniz, K. Z. (Ed.). (2021). *İstatistikolay 2 - Çok değişkenli istatistik*. Nobel Akademik Yayıncılık.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. L. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. Retrieved from https://arxiv.org/abs/1810.04805

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), 1895-1923. doi: 10.1162/089976698300017197

Dıklı, S. (2006). Automated essay scoring. *Turkish Online Journal of Distance Education*, 7(1), 49-62.

Dridi, S. (2021). Unsupervised learning: A systematic literature review. *Preprint*.

Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, 109(3), 373-440. doi:10.1007/s10994-019-05855-6

Eryiğit, G., & Oflazer, K. (2006). Statistical dependency parsing of Turkish. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy. Retrieved from https://research.sabanciuniv.edu/id/eprint/1211/

Fang, L., Lee, G.-G., & Zhai, X. (2023). Using GPT-4 to augment unbalanced data for automatic scoring. Retrieved from https://arxiv.org/abs/2310.18365

Ferreira, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2020). Text mining in education: 2006-2018. In V. E. Balas, R. Kumar, & R. Srivastava (Eds.), *Recent trends and advances in artificial intelligence and internet of things* (pp. 1-18). Springer. doi:10.1007/978-3-030-32644-9

Flor, M., & Hao, J. (2021). Text mining and automated scoring. In A. A. Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational psychometrics: New methodologies for a new generation of digital learning and assessment*. doi:10.1007/978-3-030-74394-9_14

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning*, 37(5), 1207-1216. doi:10.1111/jcal.12577

Gierl, M.J., Zhou, J., Alves, C. (2008). Developing a Taxonomy of Item Model Types to Promote Assessment Engineering. Journal of Technology, Learning, and Assessment, 7(2). Retrieved [date] from http://www.jtla.org.

Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & de Champlain, A. F. (2014). Automated essay scoring and the future of educational assessment in medical education. *Medical Education*, 48(10), 950–962. doi:10.1111/medu.12517

Gierl, M. J., & Lai, H. (2016). A process for reviewing and evaluating generated test items. *Educational Measurement: Issues and Practice, 35*(1), 6–20. doi: 10.1111/emip.12129

Gunawansyah, R. R., Rahayu, Nurwathi, B., Sugiarto, & Gunawan. (2020). Automated essay scoring using natural language processing and text mining method. In *Proceedings of the 14th International Conference on Telecommunication Systems, Services, and Applications (TSSA)* (pp. 1–4). IEEE. doi: 10.1109/TSSA51342.2020.9310845

Google. (2023). *Google Colaboratory*. https://colab.research.google.com/

Göloğlu Demir, C., & Yılmaz, H. (2018). Sınıf Dışı Eğitim Faaliyetlerinin Öğrencilerin Bilim ve Teknolojiye Yönelik Tutumlarına Etkisi ve Duygu Analizi. *İnsan ve Toplum Bilimleri Araştırmaları Dergisi, 7*(5), 101-116. doi:10.15869/itobiad.483404

Hakkani Tür, D. Z., Oflazer, K., & Tür, G. (2002). Statistical morphological disambiguation for agglutinative languages. *Computers and the Humanities, 36*(4), 381-410. doi: 10.1023/A:1020271707826

Han, J., Yoo, H., Myung, J., Kim, M., Lim, H., Kim, Y., Lee, T. Y., Hong, H., Kim, J., Ahn, S. Y., & Oh, A. (2024). LLM-as-a-tutor in EFL writing education: Focusing on evaluation of student-LLM interaction. *arXiv preprint arXiv:2310.05191*. https://arxiv.org/abs/2310.05191

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. doi: 10.1109/tkde.2008.239

Hendler, J. (2008). Avoiding another AI winter. *IEEE Intelligent Systems, 23*(2), 2-4.

Ibrahim, S. S., Elfakharany, E. F., & Hamed, E. (2022). Improved automated essay Grading system via natural language processing and deep learning. *In Proceedings of the 8th International Conference on Engineering and Emerging Technologies (ICEET).* Kuala Lumpur, Malaysia. doi: 10.1109/ICEET56468.2022.10007407

Ishioka, T., & Kameda, M. (2006). Automated Japanese essay scoring system based on articles written by experts. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL),* 233–240. https://aclanthology.org/P06-1030

Jang, E.-S., Kang, S.-S., Noh, E.-H., Kim, M.-H., Sung, K.-H., & Seong, T.-J. (2014). KASS: Korean automatic scoring system for short-answer questions. *Proceedings of the 6th International Conference on Computer Supported Education (CSEDU-2014)*, 2, 226–230. doi:5220/0004864302260230

Jurafsky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed.). Pearson.

Kaggle. (2012). *ASAP-AES: Automated student assessment prize - Automated essay scoring* [Makine öğrenimi yarışması]. Kaggle. Retrieved from https://www.kaggle.com/competitions/asap-aes

Kakkonen, T., Myller, N., Timonen, J., & Sutinen, E. (2005). Automatic essay grading with probabilistic latent semantic analysis. *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, 29–36. Association for Computational Linguistics. https://aclanthology.org/W05-0206/

Kerkhof, R. (2020). Natural Language Processing for scoring open-ended questions: A systematic review. Unıversıty of twente student theses. Retrieved from: http://essay.utwente.nl/82090/

Kış, A. (2019). Eğitimde yapay zeka. In Y. Kondakçı, S. Emil, & K. Beycioğlu (Eds.), *14. Uluslararası Eğitim Yönetimi Kongresi Tam Metin Bildiri Kitabı* (ss. 197-202). Pegem Akademi. Erişildiği adres: https://depo.pegem.net/9786050370126.pdf

Krishnamurthy, S., Gayakwad, E., & Kailasanathan, N. (2019). Deep learning for short answer scoring. *International Journal of Recent Technology and Engineering* (IJRTE), 7(6). Retrieved from: https://www.researchgate.net/publication/333447428_Deep_learning_for_short_answer_scoring

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1137-1145. Morgan Kaufmann.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence, 6*, 100210. doi:10.1016/j.caeai.2024.100210

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. doi:10.1038/nature14539

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). *Mining of massive datasets* (3rd ed.). Cambridge University Press. doi:10.1017/9781108684163

Liddy, E. D. (2001). Natural language processing. In *Encyclopedia of library and information science* (2nd ed.). Marcel Decker, Inc. Retrieved from https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub

Lighthill, J. (1973). Artificial intelligence: A general survey. In *Artificial intelligence: A paper symposium* (pp. 1-77). Science Research Council. Retrieved from https://www.aiai.ed.ac.uk/events/lighthill1973/lighthill.pdf

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical Natural Language Processing*. MIT Press.

Marinho, J. C., Anchieta, R. T., & Moura, R. S. (2021). Essay-BR: a Brazilian Corpus of Essays. *arXiv*. Retrieved from: https://arxiv.org/abs/2105.09081

Mathias, S., Murthy, R., Kanojia, D., & Bhattacharyya, P. (2021). Cognitively aided zero-shot automatic essay grading. *arXiv preprint arXiv:2102.11258*. https://arxiv.org/abs/2102.11258

McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. *In AAAI-98 workshop on learning for text categorization* (pp. 41-48). Wisconsin, USA:AAAI Press.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A proposal for the Dartmouth summer research project on artificial intelligence. Retrieved from http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

McCarthy, J. (2007). What is artificial intelligence? *Computer Science Department, Stanford University.* Retrieved from http://www-formal.stanford.edu/jmc/whatisai.pdf

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*(4), 115–133. doi:10.1007/BF02478259

McKinney, W. (2010). Data analysis in Python. *Proceedings of the 9th Python in Science Conference* (pp. 51-56).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. https://arxiv.org/abs/1301.3781

Mulianingsih, F., Anwar, K., Shintasiwi, F. A., & Rahma, A. J. (2020). Artificial ıntelligence dengan pembentukan nilai dan karakter di bidang pendidikan. ıjtimaiya: *Journal of Social Science Teaching,* 4(2), 148-154. Retrieved from https://typeset.io/papers/artificial-intellegence-dengan-pembentukan-nilai-dan-3p9wey9zzb

Munzenmaier, C., & Rubin, N. (2013). *Bloom's taxonomy: What's old is new again*. The eLearning Guild. Retrieved from https://publicservicesalliance.org/wp-content/uploads/2013/04/guildresearch_blooms2013.pdf

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Nael, O., Elmanyalawy, Y., & Sharaf, N. (2022). AraScore: A deep learning-based system for Arabic short answer scoring. *Array*, 10, 100109. doi:10.1016/j.array.2021.100109

Newell, A., & Simon, H. A. (1956). The logic theory machine. *IRE Transactions on Information Theory, 2*(3), 61-79.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7, 21. doi:10.3389/fnbot.2013.00021

Nilsson, N. J. (1984). Shakey the robot. *SRI International.* Retrieved from https://www.sri.com/publication/artificial-intelligence-pubs/shakey-the-robot-pub/

Oflazer, K. (1994). Two-level description of Turkish morphology. *Literary and Linguistic Computing, 9*(2), 137-148. doi:10.1093/llc/9.2.137

Oflazer, K., & Kuruöz, İ. (1994). Tagging and morphological disambiguation of Turkish text. *Literary and Linguistic Computing*, *9*(2), 137–150. doi:10.3115/974358.974391

Oflazer, K. (2016). Türkçe ve Doğal Dil İşleme. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi,* 5(2). Erişim adresi: https://dergipark.org.tr/tr/pub/tbbmd/issue/22245/238795

Oflazer, K., & Tür, G. (1996). Combining hand-crafted rules and unsupervised learning in constraint-based morphological disambiguation. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Retrieved from: https://aclanthology.org/W96-0207/

Oflazer, K., & Saraçlar, M. (2018). *Turkish natural language processing.* Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. Retrieved from: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pise, N., & Kulkarni, P. (2008). A survey of semi-supervised learning methods. *IEEE International Conference on Computational Intelligence and Security,* 13-17 December 2008. doi:10.1109/CIS.2008.204

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard". *Applied Measurement in Education*, 28(2), 130-142. doi:10.1080/08957347.2014.1002920

Ramachandran, D., & Rana, R. S. (2024). AI-driven jurisprudence: Navigating legal landscapes in the digital age. *Journal of Digital Jurisprudence*, 4(1), Part B. doi: 10.22271/2790-0673.2024.v4.i1b.103

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. *Department of Computer Science, Rutgers University*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b3bf6373ff41a115197cb5b30e57830c16130c2c

Rodriguez, P. U., Jafari, A., & Ormerod, C. M. (2019). *Language models and automated essay scoring*. arXiv preprint arXiv:1909.09482. Retrieved from https://arxiv.org/abs/1909.09482

Rokade, A., Patil, B., Rajani, S., Revandkar, S., & Shedge, R. (2018). Automated grading system using natural language processing. *In Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018).* IEEE. doi: 10.1109/ICICCT.2018.8473170

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386-408. doi: 10.1037/h0042519

Sadanand, V. S., Guruvyas, K. R., Patil, P. P., Acharya, J. J., & Suryakanth, S. G. (2022). An automated essay evaluation system using natural language processing and sentiment analysis. *International Journal of Electrical and Computer Engineering (IJECE),* 12(6), 6585-6593. doi: 10.11591/ijece.v12i6.pp6585-6593

Sadiku, M. N. O., Ashaolu, T. J., Ajayi-Majebi, A., & Musa, S. M. (2021). Artificial intelligence in education. *International Journal of Scientific Advances, 2*(1), 1-11. https://typeset.io/pdf/artificial-intelligence-in-education-5ggabmq2kf.pdf

Sak, H., Güngör, T., & Saraçlar, M. (2007). Morphological disambiguation of Turkish text with perceptron algorithm. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 107-118). Springer. doi:10.1007/978-3-540-70939-8_10

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210-229. doi:10.1147/rd.33.0210

Sarıgil, Ş., & Koç, T. K. (2023). Mintlemon-turkish-nlp: Türkçe doğal dil işleme kütüphanesi. *PyPI*. Retrieved January 17, 2025, from https://pypi.org/project/mintlemon-turkish-nlp/

Saraswathi, M., Balu, V., Sri Ram, P. V., & Prakash, L. S. (2024). A novel study on machine learning algorithms for big data. *International Journal of Scientific Research in Engineering and Management, 8*(3), 1-4. doi:10.55041/IJSREM29690

Scikit-learn Developers. (n.d.). *Model selection*. Scikit-learn. Retrieved March 3, 2025, https://scikit-learn.org/stable/model_selection.html

Sheikh, H., Prins, C., & Schrijvers, E. (2023). *Mission AI: The new system technology*. Springer Nature Switzerland AG. Retrieved from: doi: 10.1007/978-3-031-21448-6

Shermis, M. D. (2010). Automated essay scoring in a high stakes testing environment. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st century* (pp. 167-185). New York: Springer.

Shin, E. (2021). Automated item generation by combining the non-template and template-based approaches to generate reading inference test items (Doctoral dissertation, University of Alberta). Department of Educational Psychology.

Sun, H., & Shao, Z. (2019). Research on the application of students' answered record analyze model and question automatic classify based on k-means clustering algorithm. *In 10th International Conference on Information Technology in Medicine and Education (ITME).* doi: 10.1109/ITME.2019.0011

Sun, K., & Wang, R. (2024). Automatic essay multi-dimensional scoring with fine-tuning and multiple regression. *arXiv preprint arXiv:2406.01198*. https://arxiv.org/abs/2406.01198

Süzen, N., Gorbana, A. N., Levesleya, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science,* 169, 726-743. doi: 10.1016/j.procs.2020.02.171

Streamlit, Inc. (n.d.). *Streamlit: An open-source app framework for Machine Learning and Data Science teams*. Streamlit. Retrieved January 17, 2025, from https://streamlit.io

Sytnyk, L., & Podlinyayeva, O. (2024). AI in education: Main possibilities and challenges. In *Proceedings of the 8th International Scientific and Practical Conference International Scientific Discussion: Problems, Tasks and Prospects* (pp. 569-579). Brighton, United Kingdom. doi:10.51582/interconf.19-20.05.2024.058.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson. Retrieved from https://hisp.htmi.ch/pluginfile.php/77114/mod_resource/content/0/Using%20Multivariate%20Statistics%20%28Tabachnick%20and%20Fidell%29.pdf

Tanveer, H., Adam, M. A., Khan, M. A., Ali, M. A., & Shakoor, A. (2023). Analyzing the performance and efficiency of machine learning algorithms, such as deep learning, decision trees, or support vector machines, on various datasets and applications. *Asian Bulletin of Big Data Management, 3*(2), 126-136. doi: 10.62019/abbdm.v3i2.83

Turhost. (2020). Makine öğrenmesi (Machine Learning) nedir? *Turhost Blog.* https://www.turhost.com/blog/makine-ogrenmesi-machine-learning-nedir/

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, s2-42*(1), 230–265. doi: 10.1112/plms/s2-42.1.230

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. doi:10.1093/mind/LIX.236.433

Vanbelle, S. (2014). A new interpretation of the weighted kappa coefficients. *Psychometrika*. doi:10.1007/s11336-014-9439-4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*. Retrieved from https://arxiv.org/abs/1706.03762

Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., & Chen, G. (2024). Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*.

Yuret, D., & Türe, F. (2006). Learning morphological disambiguation rules for Turkish. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (pp. 328-334). doi:10.3115/1220835.1220877

Yu, B., & Zheng, Y. (2023). Research on algorithms of machine learning. *Proceedings of the 2023 International Conference on Machine Learning and Automation*, 277–281. doi:10.54254/2755-2721/39/20230614

Wang, J. (2020). Application of text clustering in automatic scoring of college English composition. In *International Conference on Information Technology and Computer Application (ITCA)* (pp. 598–603). doi:10.1109/ITCA52113.2020.00131

Wang, Y., Wei, Z., Zhou, Y., & Huang, X. (2018). Automatic essay scoring incorporating rating schema via reinforcement learning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 791–797. Association for Computational Linguistics. doi:10.18653/v1/D18-1090

Watson, D. S. (2023). On the philosophy of unsupervised learning. *Philosophy & Technology, 36*(28). doi:10.1007/s13347-023-00635-6

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36-45.

Wilson, L. O. (2016). *Understanding the new version of Bloom's taxonomy*. Retrieved from:https://quincycollege.edu/wp-content/uploads/Anderson-and-Krathwohl_Revised-Blooms-Taxonomy.pdf

Yılmaz, K., & Deniz, K. Z. (2024). Natural language processing and machine learning applications for assessment and evaluation in education: Opportunities and new approaches. *Journal of Measurement and Evaluation in Education and Psychology, 15*(4), 421–445. doi:10.21031/epod.1551568

Zeyrek, D., Turan, Ü. D., Bozsahin, C., Çakıcı, R., Sevdik-Çallı, A., Demirşahin, I., Aktaş, B., Yalçınkaya, İ., & Ögel, H. (2009). Annotating subordinators in the Turkish Discourse Bank. *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009*, 44–47. Suntec, Singapore. doi: 10.3115/1698381.1698387

Zhang, D., & Yuan, X. (2022). Intelligent scoring of English composition by machine learning from the perspective of natural language processing. *Hindawi Mathematical Problems in Engineering.* doi:10.1155/2022/9070272

# EKLER

## EK 1. Etik Kurul Onayı

**ANKARA ÜNİVERSİTESİ**
**ALT ETİK KURULU**
**KARAR ÖRNEĞİ**

| | |
|---|---|
| **Karar Tarihi** | **: 17/07/2023** |
| **Toplantı Sayısı** | **: 16** |
| **Karar Sayısı** | **: 148** |

**148-** Üniversitemiz Eğitim Bilimleri Anabilim Dalı, **Öğr. Üyesi Prof. Dr. Kaan Zülfikar DENİZ**'in danışmanlığını yaptığı, doktora öğrencisi **Kübra YILMAZ**'ın "Açık Uçlu Maddelerin Otomatik Planlamasında Doğal Dil İşleme Yönteminden Yararlanılarak Makine' Öğrenmesi Algoritmanın Karşılaştırılması " başlıklı tezi ile ilgili 03/07/2023 tarihli "İnsan Üzerinde Yapılan Klinik Dışı Araştırmalar Başvuru Formu" Etik Kurulumuzca incelendi.

Üniversitemiz Eğitim Bilimleri Anabilim Dalı, **Öğr. Üyesi Prof. Dr. Kaan Zülfikar DENİZ**'in danışmanlığını yaptığı, doktora öğrencisi **Kübra YILMAZ**'ın "Açık Uçlu Maddelerin Otomatik Planlamasında Doğal Dil İşleme Yönteminden Yararlanılarak Makine Öğrenmesi Algoritmanın Karşılaştırılması " başlıklı tezi ile ilgili araştırma protokolüne uyulması ve etik onay tarihinden itibaren geçerli olması koşuluyla uygulanmasının etik açıdan uygun olduğuna oy birliği ile karar verildi.

**ASLININ AYNIDIR**
**17/07/2023**

# EK 2. Millî Eğitim Bakanlığı Araştırma İzni

T.C.
MİLLÎ EĞİTİM BAKANLIĞI
Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü

Sayı : E-57750415-622.03-79057583 23.06.2023
Konu : ABİDE 2016 Veri Talebi (Kübra YILMAZ)

ANKARA ÜNIVERSITESI REKTÖRLÜGÜNE
(Eğitim Bilimleri Enstitüsü Müdürlüğü Enstitü Sekreterliği)

İlgi : 31.05.2023 tarihli ve E-98761816-302.08.01-942536 sayılı yazınız.

Üniversiteniz Eğitim Bilimleri Enstitüsü Eğitim Bilimleri Anabilim Dalı Eğitimde Ölçme ve Değerlendirme Doktora Programı öğrencisi Kübra YILMAZ'ın, "Açık Uçlu Maddelerin Otomatik Puanlamasında Doğal Dil İşleme Yönteminden Yararlanılarak Makine Öğrenmesi Algoritmalarının Karşılaştırılması" konulu çalışması için ABİDE 2016 Türkçe testinde açık uçlu soruların puanlaması sonucu oluşan nihai puanlara ilişkin 5000 kişilik öğrenci verisinin paylaşılması talebi Genel Müdürlüğümüzce uygun görülmüktür. Talep edilen veriler yazımız ekinde yer alan "Teşekkür ve Tez Teslim Taahhütnamesi"nin imzalı bir kopyasının Genel Müdürlüğümüze iletilmesine müteakiben kubrayilmaz.edu@gmail.com elektronik posta adresine gönderilecektir.

Bilgilerini ve gereğini rica ederim.

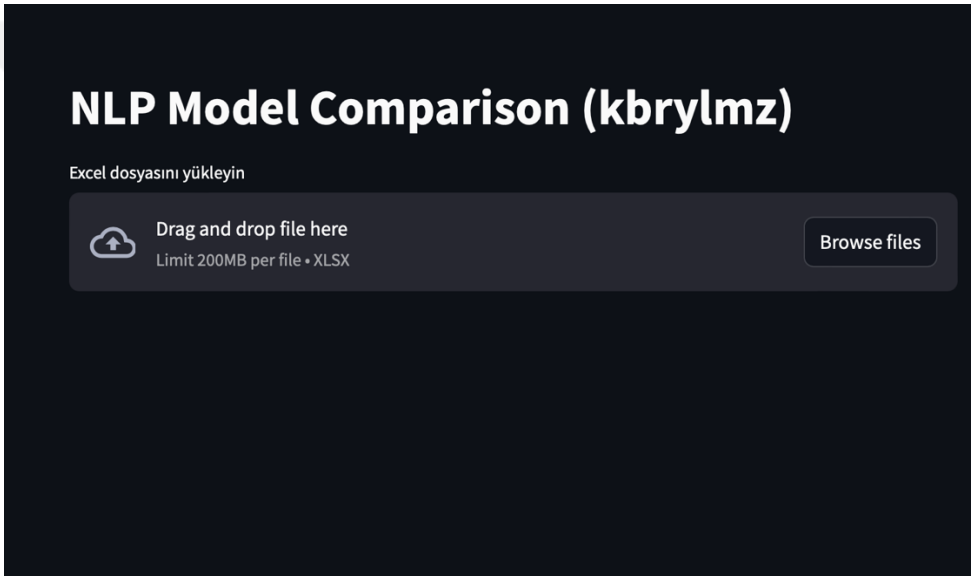
Kemal BÜLBÜL
Bakan a.
Genel Müdür

EK- Teşekkür ve Tez Teslim Taahhütnamesi (1 sayfa)
(Not: "Teşekkür ve Tez Teslim Taahhütnamesi"nin imzalı kopyası baris.ozgurluk@meb.gov.tr mail adresine iletilebilir.)

Bu belge güvenli elektronik imza ile imzalanmıştır.
Adres : Emniyet Mah. Abant 2 Sokak ÖDSGM Ek Hizmet Binası 13/A Yenimahalle/ANKARA
Telefon No : 0 (312) 413 32 40
E-Posta: baris.ozgurluk@meb.gov.tr
Kep Adresi : meb@hs01.kep.tr
Belge Doğrulama Adresi : https://www.turkiye.gov.tr/meb-ebys
Bilgi için: Dr. Barış ÖZGÜRLÜK
Unvan : Milli Eğitim Uzmanı
İnternet Adresi: Faks:

## EK 3. Uygulama ile İlgili Bilgiler

Aşağıda verilen örneklerde kullanıcı analiz etmek istediği maddeyi uygulamaya yüklemekte ve algoritma sonuçlarını anında görebilmektedir. Bu da kod yazma kısmı ile meşgul olmadan kullanıcılara model performasını görme şansı sağlamaktadır. Bu uygulama sonuçları ilk yapılan uygulama ile benzer sonuçlar vermiştir. Aralarındaki fark ise ilk uygulamaların MintLemon Türkçe kütüphanesi kullanılarak veri ön işleme adımlarının yürütülmüş olmasıdır. Uygulamadaki metin ön işleme adımlarında ise NLTK (Natural Language Toolkit) doğal dil işleme kütüphanesinden yararlanılmıştır. Uygulamada daha fazla klasik makine öğrenmesi modeline yer verilmiştir.




| | Accuracy | Precision | Recall | F1 Score | QWK |
|---|---|---|---|---|---|
| Logistic Regression | 0.8479 | 0.843 | 0.8352 | 0.8386 | 0.6774 |
| Naive Bayes | 0.8592 | 0.8584 | 0.8432 | 0.8492 | 0.699 |
| SVM | 0.8479 | 0.8461 | 0.8314 | 0.8372 | 0.6749 |
| Random Forest | 0.8254 | 0.8204 | 0.809 | 0.8136 | 0.6277 |
| Gradient Boosting | 0.8366 | 0.8337 | 0.8196 | 0.8251 | 0.6508 |
| Decision Tree | 0.7859 | 0.7753 | 0.7792 | 0.777 | 0.5541 |


| | Accuracy | Precision | Recall | F1 Score | QWK |
|---|---|---|---|---|---|
| Logistic Regression | 0.9462 | 0.9372 | 0.9361 | 0.9361 | 0.9155 |
| Naive Bayes | 0.8669 | 0.905 | 0.8284 | 0.8559 | 0.7704 |
| SVM | 0.9292 | 0.9086 | 0.921 | 0.9131 | 0.9026 |
| Random Forest | 0.9377 | 0.942 | 0.9188 | 0.929 | 0.8837 |
| Gradient Boosting | 0.9632 | 0.959 | 0.9564 | 0.9577 | 0.9519 |
| Decision Tree | 0.9235 | 0.9205 | 0.9038 | 0.9113 | 0.8377 |

## EK 4. Kod Örnekleri

```
#gerekli kütüphanelerin kurulması

import pandas as pd
from mintlemon import Normalizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import cross_validate, train_test_split
from sklearn.metrics import accuracy_score, precision_score, f1_score,
recall_score, ConfusionMatrixDisplay, confusion_matrix
from sklearn.ensemble import RandomForestClassifier
import warnings
warnings.simplefilter(action='ignore', category=UserWarning)

#NLP metin ön işleme adımları

df['Cevap'] = df['Cevap'].apply(Normalizer.remove_stopwords)
df['Cevap'] = df['Cevap'].apply(Normalizer.remove_numbers)
df['Cevap'] = df['Cevap'].apply(Normalizer.remove_punctuations)
df['Cevap'] = df['Cevap'].apply(Normalizer.lower_case)

#metinlerin sayısal verilerle temsil edilmesi (vektörleştirme)

def vectorize(data, tfidf_vect_fit):
    X_tfidf = tfidf_vect_fit.transform(data)
    words = tfidf_vect_fit.get_feature_names_out()
    X_tfidf_df = pd.DataFrame(X_tfidf.toarray())
    X_tfidf_df.columns = words
    return X_tfidf_df


#metriklerin hesaplanması

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1 Score:", f1)
```

## EK 5. İnsan Puanlayıcı ve Makine Puanlayıcı Puanlama Örnekleri

Aşağıda 2 ve 3 kategorili maddelere insan ve makine tarafından verilen puan örnekleri verilmiştir.

| İnsan_Puanlayıcı | Makine_Tahmini |
|---|---|
| 1 | 1 |
| 0 | 1 |
| 1 | 1 |
| 1 | 0 |
| 1 | 0 |
| 1 | 1 |
| 0 | 1 |
| 0 | 0 |
| 0 | 0 |
| 1 | 1 |
| 1 | 1 |
| 1 | 1 |
| 0 | 0 |
| 1 | 1 |
| 0 | 1 |
| 0 | 0 |
| 1 | 1 |
| 0 | 0 |
| 1 | 1 |
| 0 | 1 |
| 1 | 1 |
| 0 | 1 |
| 0 | 0 |
| 0 | 0 |

| İnsan_Puanlayıcı | Makine_Tahmini |
|---|---|
| 2 | 2 |
| 2 | 2 |
| 2 | 2 |
| 2 | 2 |
| 1 | 2 |
| 1 | 1 |
| 2 | 2 |
| 2 | 1 |
| 0 | 0 |
| 0 | 0 |
| 2 | 2 |
| 2 | 2 |
| 0 | 2 |
| 2 | 2 |
| 2 | 2 |
| 2 | 2 |
| 1 | 1 |
| 0 | 0 |
| 2 | 2 |
| 1 | 1 |
| 2 | 2 |
| 2 | 2 |
| 2 | 2 |
| 2 | 2 |

## BENZERLİK BİLDİRİMİ

"Açık Uçlu Maddelerin Otomatik Puanlanmasında Doğal Dil İşleme Yönteminden Yararlanılarak Makine Öğrenmesi Algoritmalarının Karşılaştırılması" başlıklı tezimin ana bölümü (ön bölüm, kaynaklar ve ekler hariç) Turnitin İntihal Tespit Programı aracılığıyla incelenmiş ve ilgili rapor danışmanım tarafından da kontrol edilmiştir. Kontrol sırasında (1) "Yedi sözcükten daha az olan benzeşmeler" (2) "Kaynaklar" (3) "Doğrudan alıntılar" ve (4) "Lisansüstü tezim ile ilgili yapmış olduğum kendi yayınlarım ve yararlandığım mevzuat metinleri gibi kaynaklar" dışarıda tutulmuştur. Benzerlik kontrolüne ilişkin rapordan elde edilen sonuçlar aşağıda sunulmuştur.

| | |
|---|---|
| **Rapor Tarihi** | : 11.03.2025 |
| **Gönderim Numarası** | : 2581139979 |
| **Sayfa Sayısı** | : 73 |
| **Sözcük Sayısı** | : 17,784 |
| **Karakter Sayısı** | : 128,823 |
| **Benzerlik Oranı** | : % 5 |
| **Savunma Tarihi** | :25.02.2025 |

Yukarıda belirtilen sonuçları gösteren Turnitin İntihal Tespit Programı'na ilişkin orijinal raporu, sonuçlarda herhangi bir değişiklik yapmaksızın bu beyanım ekinde Enstitüye teslim ettiğimi, tezimin %10'dan fazla benzerlik oranı içerdiğinin, tek bir kaynakla eşleşme oranının ise %2'den fazla olduğunun belirlenmesi durumunda, bundan doğabilecek tüm yasal sorumluluğu kabul ettiğimi bildirir, saygılarımı sunarım.

**Öğrencinin Adı Soyadı:** Kübra Yılmaz

**Tarih:** 11.03.2025

**İmza:**

# ÖZGEÇMİŞ

**Kişisel Bilgiler**

**Adı ve Soyadı : Kübra Yılmaz**

**E-Posta Adresi :**

**İş Deneyimi :**

| Unvan | Görev Yeri | Yıl |
|---|---|---|
| Öğretmen | M.Sadık Eratik Ortaokulu | 2017 |
| Öğretmen | H.Emine Oba Ortaokulu | 2017 |

**Akademik Bilgiler**

**Öğrenim Durumu:**

| Derece | Bölüm/Program | Üniversite | Yıl |
|---|---|---|---|
| Lisans | Bilgisayar ve Öğretim Teknolojileri Eğitimi | Sakarya Üniversitesi | 2011-2015 |
| Y. Lisans | Eğitimde Ölçme ve Değerlendirme | Abant İzzet Baysal Üniversitesi | 2016-2019 |
| Doktora | Eğitimde Ölçme ve Değerlendirme | Ankara Üniversitesi | 2019-2025 |

**Yayınlar:**

Yılmaz, K., & Deniz, K. Z. (2024). Natural Language Processing and Machine Learning Applications For Assessment and Evaluation in Education: Opportunities and New Approaches. Journal of Measurement and Evaluation in Education and Psychology, 15(4), 421-445. https://doi.org/10.21031/epod.1551568

Bilicioğlu, A., Yılmaz, K. (2017). Öğrencilerin Sınav Kaygısı, Fene Yönelik İlgi ve Ebeveyn Desteği Değişkenleri Üzerine Uluslararası Bir Karşılaştırma: Türkiye – Singapur. Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 17 (3), 1201-1220.

Bilicioğlu A., Yılmaz K., "PISA 2015 Sonuçlarına göre Öğrencilerin Sınav Kaygısı, Fene Yönelik İlgi ve Ebeveyn Desteği Değişkenleri Üzerine Uluslararası Bir Karşılaştırma: Türkiye- Singapur", 26. Uluslararası Eğitim Bilimleri Kongresi (UEBK), Antalya, Türkiye 20-23 Nisan 2017.