

**ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

DOKTORA TEZİ

**SANAL PARA PAZARLARINDAKİ MANİPÜLATÖRLERİN MAKİNE
ÖĞRENİMİ YÖNTEMLERİ KULLANILARAK TESPİTİ**

Fırat AKBA

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

**ANKARA
2021**

Her hakkı saklıdır

ÖZET

Doktora Tezi

SANAL PARA PAZARLARINDAKİ MANİPÜLATÖRLERİN MAKİNE ÖĞRENİMİ YÖNTEMLERİ KULLANILARAK TESPİTİ

Fırat AKBA

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. İman ASKERBEYLİ

Eş Danışman: Doç. Dr. İhsan Tolga MEDENİ

Günümüzde hisse senedi piyasalarında menkul kıymetler açısından sürekli değişimler yaşanmaktadır. Yatırımcılar temel analiz araçlarını ve göstergelerini çok yaygın olarak kullanarak bu değişimleri anlamlandırmaya çalışırlar. Bu sayede piyasalarda yaşanan durumlar hakkında biraz bilgi sahibi olmak ve kazanç elde etmek mümkün hale gelmektedir. Bu çalışmada yatırımcılar açısından çok popüler yatırım araçlarından biri olan Bitcoin üzerinde yapılan manipülasyonlar ele alınmıştır. Manipülasyonları tespit etmek için popüler makine ve istatistiksel tahmin yöntemleri kullanılmış ve en başarılı şekilde tespit edilebilmesi için izlenmesi gereken yol haritaları paylaşılmıştır. Çalışmalar sırasında manipülasyonlara etkisi olduğu düşünülen sosyal medya duyuları, duygu analizi yöntemleriyle sınıflandırılmıştır. Bunun yanında en başarılı fiyat tahminleme yöntemleri kullanılarak aralarındaki ilişkiler incelenmiştir. Değerlendirmeler aşamasında, kriz öncesi gerçekleştirilen tahmin yöntemlerinin daha başarılı olduğu iddiaları araştırılmıştır. Bu amaçla küresel çapta yaşanan Covid-19 salgını, tüm dünyayı ekonomik bir krize sürüklediği için bu süreçte yaşanan fiyat değişimlerinin etkileri tahminleme yöntemleriyle araştırılmıştır. Tüm süreç dahilinde borsalarda büyük kazanç sağlayanların aslında borsanın yönünü belirleyenler olduğunu söylemek yanlış bir ifade olmaz. Bunu sağlayabilmek için de piyasalarda manipülasyon yapmaktadırlar. Bu çalışmalarda manipülasyon dönemlerinde gerçekleştirilen deneylerle, çeşitli yöntemler kullanılarak doğrulanmaya çalışılmıştır. Deney sonuçları kriz öncesinde periyotlarda fiyat tahminleme başarımlarının %93 F_1 skor başarısına kadar ulaşabildiğini göstermiştir. Bu yüksek başarımların iddialarda bahsedildiği üzere, kriz öncesi periyotlar için başarımların düşmesi beklenmiştir. Bu çalışmada yaşanan başarımların düşüşlerin aslında manipülasyonlardan kaynaklanabileceği ortaya çıkartılmıştır. Sonrasında fiyat tahminleme yöntemlerinin yanlışlıkları noktalar anomali olarak etiketlenmiştir ve bu anomali tespit edilen dönemlerde en yüksek işlem gerçekleştirmiş Bitcoin hesapları potansiyel manipulatörler olarak tespit edilebilmiştir.

Eylül 2021, 77 sayfa

Anahtar Kelimeler: Anomali Tespiti, Makine Öğrenimi, Zaman Serisi Analizi, Duygu Analizi, ISSFS

ABSTRACT

Ph.D. Thesis

DETECTION OF MANIPULATORS IN VIRTUAL MONEY MARKETS USING MACHINE LEARNING METHODS

Firat AKBA

Ankara University

Graduate School of Natural and Applied Sciences

Department of Computer Engineering

Supervisor: Prof. Dr. Iman ASKERBEYLI

Co-Supervisor: Assoc. Prof. Dr. Ihsan Tolga MEDENI

Today, there are constant changes in terms of securities in stock markets. Investors try to make sense of these changes by using fundamental analysis tools and indicators very widely. In this way, it becomes possible to have some information about the situations in the markets and to gain profit. In this study, manipulations on Bitcoin, which is one of the most popular investment tools for investors, are discussed. Popular machine and statistical estimation methods were used to detect manipulations, and roadmaps to be followed for the most successful detection were shared. Social media emotions, which were thought to have an effect on manipulations during the studies, were classified by sentiment analysis methods. In addition, the relationships between them were examined by using the most successful price estimation methods. In the evaluation phase, the claims that the estimation methods performed before the crisis were more successful were investigated. For this purpose, the effects of price changes experienced in this process were investigated by estimation methods, since the global Covid-19 epidemic dragged the whole world into an economic crisis. It would not be wrong to say that those who make big gains in the stock markets during the whole process are actually the ones who determine the direction of the stock market. In order to achieve this, they manipulate the markets. In these studies, manipulation periods are tried to be determined by using various methods. Experiments showed that price prediction performances in the pre-crisis periods could reach up to 93% F1 score success. As mentioned in the claims, these high-performance results were expected to decrease for the pre-crisis periods. It has been revealed that the performance drops in this study may actually be caused by manipulations. Afterwards, the points where the price estimation methods were wrong were labeled as anomalies. Finally, Bitcoin accounts with the highest transactions in the periods when this anomaly was detected could be identified as potential manipulators.

September 2021, 77 pages

Keywords: Anomaly Detection, Machine Learning, Time Series Analysis, Sentiment Analysis, ISSFS

TEŞEKKÜR

Bu çalışmanın yürütülmesi sırasında desteğini esirgemeyen danışmanım Prof. Dr. İman ASKERBEYLİ (Ankara Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı) ve eş-danışman hocam Doç. Dr. İhsan Tolga MEDENİ' ye (Yıldırım Beyazıt Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı),

Bu süreçte destek ve katkılarından dolayı değerli jüri üyelerim Prof. Dr. Ali Aydın SELÇUK (TOBB Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı) ve Doç. Dr. Mehmet Serdar GÜZEL (Ankara Üniversitesi Bilgisayar Mühendisliği Anabilim Dalı) hocalarıma,

Çalışma arkadaşım Sn. Ahmet Kürşat ESİM'e,

Sabrından dolayı sevgili eşim Zeliha AKBA' ya,

Ve çalışmam sırasında yardımını esirgemeyen tüm herkese teşekkürlerimi sunarım.

Fırat AKBA

Ankara, Eylül 2021

İÇİNDEKİLER

TEZ ONAY SAYFASI

ETİK	i
ÖZET	ii
ABSTRACT	iii
TEŞEKKÜR	iv
ÇİZELGELER DİZİNİ	vii
ŞEKİLLER DİZİNİ	viii
KISALTMALAR DİZİNİ	ix
1. GİRİŞ	1
1.1 Problem Tanımı	1
1.2 Motivasyon ve Özgün Değer	2
1.3 Literatür Katkısı	3
1.4 Tez Çalışmasının Ana Hatları	4
2. LİTERATÜR TARAMASI	6
2.1 Piyasa Fiyatlamaları ve Manipülasyonlarla İlgili Literatür Çalışmaları	6
2.2 Doğal Dil İşleme ve Duygu Analizi Yöntemleriyle İlgili Literatür Çalışmaları	10
3. DENEYLERDE KULLANILAN YÖNTEMLER	12
3.1 Başarım Metrikleri	12
3.1.1 F_1 skor hesaplama	12
3.1.2 Matthews korelasyon katsayısı (MCC) hesaplama	13
3.1.3 Doğruluk değeri hesaplama	13
3.2 Borsa Teknik Analizinde Kullanılan Temel Araçlar	14
3.2.1 Trend çizgileri	14
3.2.2 Göreceli güç endeksi (RSI)	15
3.2.3 Hareketli ortalamaların yakınsama ayrımı (MACD)	17
3.2.4 Üssel hareketli ortalama (EMA)	18
3.3 Çalışma Konusu ve Kapsamında Değerlendirilen Yöntemler	18
3.3.1 Duygu analizi yöntemleri	19
3.3.2 Otomatik regresif entegre hareketli ortalama (ARIMA-SARIMAX) yöntemi	19
3.3.3 Derin öğrenme yöntemleri	20
3.3.3.1 Evrişimli sinir ağları (CNN)	21
3.3.3.2 Derin sinirsel karar ormanı (DNDF) yöntemi	22
3.3.3.3 Uzun kısa-süreli bellek (LSTM) yöntemi	23

3.3.4 Diğer makine öğrenimi yöntemleri.....	23
3.3.4.1 Destek vektör makineleri (SVM).....	23
3.3.4.2 Rastgele orman (Random Forest) algoritması	25
3.3.5 Öznitelik seçme metrikleri (Feature selection metrics)	25
3.3.6 Yinelemeli öğrenme kontrol teknikleri	26
4. DENEY ÇALIŞMALARI VE SÜREÇLERİ.....	28
4.1 Deney Ön Hazırlık Çalışmaları	28
4.1.1 Ana fikir.....	28
4.1.2 Teknik olarak imkânsız olan durumlar	29
4.1.3 Metodoloji.....	30
4.1.3.1 İteratif yarı gözetimli öznitelik seçme yöntemi	31
4.1.3.2 Manipülasyon tespiti.....	33
4.1.3.3 Manipülatör tespiti	34
4.1.3.4 Önerilen sistem tasarımı.....	35
4.2 Veri Hazırlama.....	37
4.2.1 Manipülasyon verisi.....	37
4.2.2 Bitcoin fiyat verileri	39
4.2.3 Tüm bitcoin işlem verileri	39
4.2.4 Sosyal medyada yer alan spekülasyon verisi	40
4.3 Tahminleme yöntemlerinin gerçek başarımlarının tespiti.....	42
4.4 Sosyal Medya Spekülasyonlarının Duygu Analizi Teknikleriyle Değerlendirilmesi	44
4.4.1 Verilerin etiketlenmesi	44
4.4.2 ISSFS yönteminin geliştirilmesi ve verilerin çoğaltılması	45
4.5 Manipülasyonların Fiyat Tahmin Modellemeleriyle Değerlendirilmesi.....	52
4.6 Manipülasyon Verisi Üzerinden Manipülatör Tespitinin Gerçekleştirilmesi	55
5. ÖZET, TARTIŞMA, DEĞERLENDİRME VE GELECEK ÇALIŞMALAR	57
5.1 Çalışma Özeti	57
5.2 Tartışma ve Değerlendirme.....	59
5.3 Gelecekte Yapılması Planlanan Çalışmalar.....	60
KAYNAKLAR	61
EK 1 Bitcoin Borsaları için En Belirleyici 375 Adet Kelime	65
EK 2 Manipülatör Tespiti İçin Veritabanı Sorgusu	74
ÖZGEÇMİŞ.....	76

ÇİZELGELER DİZİNİ

Çizelge 2.1 Küresel Pazar Hacimleri	6
Çizelge 4.1 Kriz Öncesi Fiyat Tahmini Başarım Değerleri	43
Çizelge 4.2 Öznitelik Seçme Uygulanmadan Elde Edilen Sınıflama Sonuçları	45
Çizelge 4.3 Öznitelik Seçme Yöntemi Kullanılarak Elde Edilen Sınıflama Sonuçları ..	46
Çizelge 4.4 ISSFS Yöntemi Aşama Karşılaştırmalı Sınıflama Sonuçları	47
Çizelge 4.5 DNDF Değişen Öznitelik Sayısına Göre Başarım Değişimi	48
Çizelge 4.6 DNDF Öznitelik Sayısına Göre Başarım Değişimi (Düşük Döngü Sayısı)	49
Çizelge 4.7 Bütün Yöntemlerin Öznitelik Sayılarına Göre Başarım Değişimi	50
Çizelge 4.8 ISSFS Yöntemi Büyük Verilerdeki Sistem Dengesi Çalışmaları	51
Çizelge 4.9 Manipülasyon Dönemi için Gerçekleştirilen Fiyat Tahmini Sonuçları	53
Çizelge 4.10 Manipülasyon Dönemi Anomalileri Tarih Aralıkları	55

ŞEKİLLER DİZİNİ

Şekil 1.1	Çalışmaların Deneysel Ana Hatları	4
Şekil 3.1	Trend Çizgileri	15
Şekil 3.2	RSI İndikatörü.....	16
Şekil 3.3	MACD İndikatörü.....	17
Şekil 3.4	Basit CNN Yapısı	21
Şekil 3.5	Örnek DNDF Yapısı (Kontschieder 2015)	22
Şekil 3.6	SVM Çalışma Mantığı (Akca 2020)	24
Şekil 3.7	Yinelemeli Öğrenme Kontrolü	26
Şekil 4.1	ISSFS İş Akış Diyagramı.....	31
Şekil 4.2	Köpüğün Aşamaları (Rodrigue, 2020).....	34
Şekil 4.3	Manipülatör Tespiti için Kullanılan İlişkisel Cebir İfadesi	35
Şekil 4.4	Önerilen Sistem Tasarımı	36
Şekil 4.5	Köpüğün Aşamaları Grafiği ve Tarih Belirteçleri	38
Şekil 4.6	Bitcoin Gerçek Fiyat Hareketi	39
Şekil 4.7	Örümcek Yazılım İş Akış Diyagramı	41
Şekil 4.8	Manipülasyon Bölgeleri için Tespit Edilmiş Anomali Periyotları	54

KISALTMALAR DİZİNİ

AHMMAS	Anomali Durumlu Uyarlanabilir Saklı Markov Modeli
ANN	Yapay Sinir Ağı
ARIMA	Otoregresif Entegre Hareketli Ortalama
BTC	Bitcoin
CAD	Konformal Anomali Tespiti
CART	Sınıflandırma ve Regresyon Ağacı
CNN	Evrişimli Sinir Ağları
DA	Duygu Analizi
DL	Derin Öğrenme
DNDF	Derin Sinirsel Karar Ormanı
EMA	Üssel Hareketli Ortalama
ETH	Ethereum
GARCH	Otoregresif Koşullu Değişken Varyans
GMM	Gauss Karışım Modeli
ILC	Yinelemeli Öğrenme Kontrolü
ISSFS	Yinelemeli Yarı Gözetimli Öznitelik Seçme
k-NN	k-En Yakın Komşuluk
k-NN-DTW	Dinamik Zaman Bükümlü k-En Yakın Komşu
LSTM	Uzun Kısa-Süreli Bellek
MACD	Hareketli Ortalamaların Yakınsama Ayrımı
ML	Makine Öğrenimi
MSSQL	Microsoft Yapılandırılmış Sorgu Dili
MCC	Matthews Korelasyon Katsayısı
NLP	Doğal Dil İşleme
OCSVM	TEK Sınıf Destek Vektör Makinesi
RF	Rastgele Orman
RNN	Özyinelemeli Sinir Ağları
RSES	Kaba Set Keşif Sistemi
RSI	Göreceli Güç Endeksi
SARIMAX	Sezonsal Otoregresif Entegre Hareketli Ortalama
SVD	Tekil Değer Ayrışımı
SVM	Destek Vektör Makinesi
TF-IDF	Terim Frekansı-Ters Belge Frekansı
XRP	Ripple

1. GİRİŞ

1.1 Problem Tanımı

Manipülasyonlar, borsalardaki fiyatlamaların yukarı ve aşağı yönde yaptığı hareketlerde en önemli etmenlerin başında gelmektedir. Öyle ki, bu yukarı ve aşağı fiyatlama hareketleri neticesinde trend adı verilen dönemsel periyotlar oluşmaktadır. Bu periyotlar içerisinde fiyatlar fark edilir bir biçimde yükseliyorsa, buna boğa piyasası denir. Öte yandan, bu periyotlarda fiyatlar düşüyorsa da buna ayı piyasası denir. Fark edilir bir değişim yoksa da buna durağan piyasa adı verilir. Bu tanımlara göre; piyasada aniden aşırı hisse satımı yaparak fiyatlarda düşüşe neden olan yatırımcılara “ayı”, aniden aşırı alım yaparak piyasada yüksek talep izlenimi oluşturan yatırımcılara ise “boğa” denir (Fontanills 2002). İşin özünde, “ayı” ve “boğa” yatırımcıları aynı kişiler veya kurumlardır. Bu yatırımcılar ellerinde büyük miktarda nakit ve sanal para enstrümanlarını bir arada bulundurdıkları için de piyasalarda balina olarak adlandırılırlar. Yapılan bu çalışmada manipülasyonların en fazla gerçekleştirildiği sanal para borsaları incelenmiştir. Günümüzde güvenlik teknolojileri açısından zirvede gösterilen blok zincir teknolojisini barındıran bu paralara, merkezi bir otorite tarafından kontrolü sağlanmadığı için çoğu resmi otorite tarafından kullanılmasına pek sıcak bakılmamıştır. Bunun en büyük nedeni de bu hesaplara sahip olunabilmesi için herhangi bir kişisel veya tüzel bilgiye ihtiyaç duyulmamasıdır. Bu şekilde anonim bir kimlik ile bir hesap adresinden başka bir hesap adresine para transferi gerçekleştirilebilmektedir. Bu noktada birçok ülke tarafından sanal paraların transferleri kontrol altına alınana kadar yasaklanmış veya geçici düzenlemeler yapılarak kullanılmasına müsaade edilmiştir. Fakat buradaki tek sorun elbette anonim bir şekilde para transfer edilmesi değildir. Aynı zamanda çok oynak fiyat hareketlenmeleri gözlenmesinden ötürü, yatırım yapan yatırımcıların büyük paralar kazanmasına veya sermayelerinin tamamına yakını kaybedebildikleri de gözlemlenmiştir. Bu aşamada sosyal medyada yapılan spekülasyonlar aracılığıyla da manipülasyonların yapıldığına dair çalışmalar (Mirtaheri 2021), (Chahine 2018) da yer almaktadır. Fakat literatür taramalarında da manipülasyon yapmanın birçok yolu olduğundan ve tek bir yöntemle ifade edilemeyeceğinden de

(Gerace 2014) bahsedilmektedir. Bitcoin, hem ilk sanal para birimi olması hem de sanal para piyasasında en büyük parasal hacme sahip olması nedeniyle incelemede tercih sebebi olmuştur. Bu noktada uluslararası piyasa düzenleyiciler için ortaya konan problemlerin çözümü noktasında faydası olabileceğini düşündüğümüz bir tez çalışması gerçekleştirilmiştir.

1.2 Motivasyon ve Özgün Değer

Bu çalışmada tahminleme yöntemlerinin kullanılmasına iten temel motivasyon, manipülasyonun genel bir tanımının olmayışdır. Bu nedenle, öncelikle yapay zekâ alanında kullanılan makine öğrenmesi ve diğer istatistiksel tahminleme yöntemlerin en asgari oranda manipülatif müdahalelerin yaşandığı bir periyotta başarımları tespit edilmeye çalışılmıştır. Sonrasında, tüm bu olayların arkasındaki manipülatörlerin tespiti aşamasında ne kadar etkili olabileceği ortaya konulmuştur. Bu etkililiği ölçmek içinse ekonomik kriz ve öncesi dönemlerinde meydana gelen piyasalardaki fiyat değişimleri deneyler yapılarak incelenmiştir. Gerçekleştirilen deneyler sadece günlük, aylık, haftalık veya yıllık fiyat verilerinin tahminleme yöntemlerinin analizine dayanmamakta, aynı zamanda sosyal medyadaki duygu analizi gerçekleştirilerek spekülasyonların da manipülasyonlara olan etkisini de göz önünde bulundurmuştur. Yapılan çalışmalar içerisinde, önce metin verilerini daha başarılı bir şekilde analiz edebilmek için bu tez çalışmasında (Akba 2014) yöntemine kıyasen daha başarılı bir yöntem olarak geliştirilen Iterative Semi-Supervised Feature Selection (ISSFS) yöntemi ortaya konmuştur. Bu yöntem derin öğrenme yöntemlerinin uzun zaman alan hesaplamalarının aksine daha başarılı ve çok kısa sürelerde hesaplama yapabilen yeni bir yöntemdir. ISSFS yönteminin tüm araştırma geliştirme süreçleri ve sonuçları ilgili kısımlarda detaylarıyla paylaşılmıştır. Yapılan çalışmaların bütününde; başarılı tahmin yöntemleri manipülasyon olduğu belirtilen periyotlar için araştırılmış, anomali tespiti yardımıyla en başarılı tahminleme yöntemleri ve manipülasyon periyotları belirlenmiştir. Bu aşamada, manipülasyonlar ve manipülatörler ile ilgili paylaşılan resmi ve doğrulanabilir kaynaklı kesin bir veri olmadığı için, bu sorunun üstesinden gelinmesi amacıyla anomali tespiti yöntemi kullanılmıştır. Özellikle (Dunning 2014) çalışmasında bilinmeyen

karakteristiğe sahip verilerde anomali tespiti yöntemlerinin kullanılmasının önerilmesi üzerine tercih edilmiştir. Bu nedenle, elde edilen sonuçlar potansiyel manipölasyon bölgeleri olarak paylaşılmıştır. Öte yandan, çalışmanın verilerinin özgünlüğünün kanıtlanabilir olması açısından; gerçek Bitcoin fiyat verileri, işlem hareketleri ve sosyal medyada yer alan gerçek kişilere ait yorumlar bu çalışmaların bütününde kullanılmıştır. Literatürde yer alan ciddi iddiaların doğrulanmaları aşamalarında deney senaryoları oluşturulmuş ve bu iddialar değerlendirilmiştir. Böylelikle asıl çalışmaların gerçekleştirileceği tüm deneyler ve uygulandığı zaman periyotları neden-sonuç ilişkisine dayandırılarak, ortada şüpheyeye yer bırakmayacak şekilde açıklanmıştır.

1.3 Literatür Katkısı

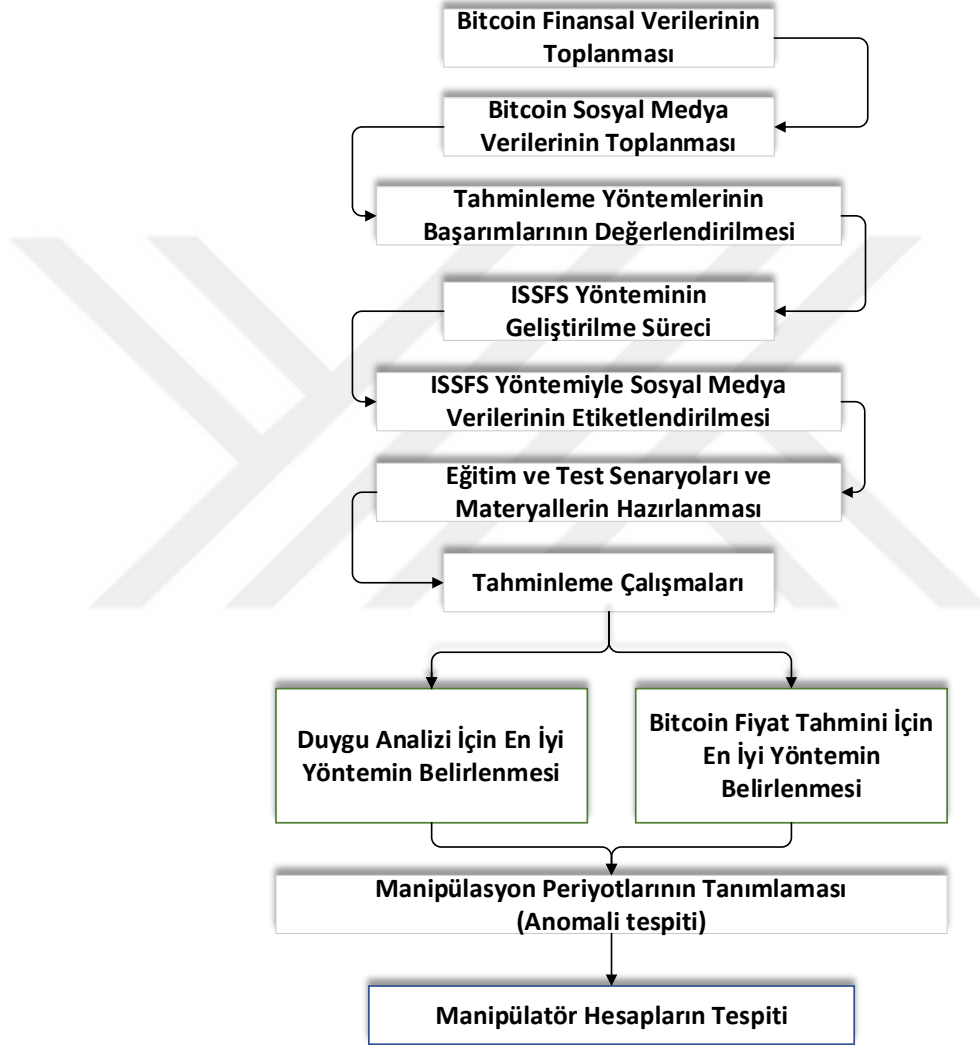
Gerçekleştirilen tez çalışmasının literatüre olan önemli katkıları aşağıda maddeler halinde gösterilmektedir:

- Kripto para piyasalarındaki potansiyel manipölatörleri mümkün olduğunca doğru ve hızlı bir şekilde belirleyebilmek
- Manipölasyon dönemleri için sosyal medyada yer alan spekülasyonların etkilerinin araştırılması
- Sanal para sektöründeki fiyat anomalilerini göz önünde bulundurarak manipölatörleri inceleyen ilk çalışma olması
- Finans sektöründeki potansiyel borsa manipölatörlerini tespit etmeye yönelik uygulanabilir iş modelini paylaşan ilk çalışma olması
- Metin madenciliğinde daha hızlı ve başarılı sonuçlar veren ISSFS yönteminin geliştirilmesi
- Hisse senedi piyasalarında kullanılan başarılı tahmin yöntemlerinin başarımlarını etkileyebilecek unsurların detaylı olarak incelenmesi
- Bitcoin borsaları için İngilizce dili için en önemli özniteliklerin paylaşılması

Bu noktada sağlanan katkıların tümü aşamalı olarak gerçekleştirilen başarılı deney süreçlerinin neticesinde sağlanmıştır. Bu nedenle, başarımlar sonuçları başka yöntemlere kıyasen daha düşük ya da o kısım için asgari olarak belirlenmiş başarımlar ölçütünü

sağlayamayan yöntemlerin sonuçları, deneylerin ilgili kısımlarında paylaşılarak bir sonraki yapılan deney aşamalarında kullanılmamıştır.

1.4 Tez Çalışmasının Ana Hatları



Şekil 1.1 Çalışmaların Deneysel Ana Hatları

Tezin çalışma aşamaları; deneylerde kullanılacak verilerin toplanması, anlamlandırılması, en başarılı yöntemlerin değerlendirilmesi, yeni yöntemlerin geliştirilmesi, senaryoların oluşturulması ve analiz kısımlarından oluşmaktadır. Gerçekleştirilen çalışmaların ana hatlarını Şekil 1.1 üzerinde gösterilmektedir. Bu

süreçlerin deneysel olan aşamaları tamamen en iyi yöntem veya yöntemlerin birleşimi sonucu elde edilen veriler ışığında ilerletilmiştir. Tezin ilerleme süreçleri de sırasıyla; literatür araştırmaları, verilerin elde edilip hazırlanması, eğitim ve test senaryolarının oluşturulması, en iyi yöntemlerin belirlenmesi, yeni yöntemlerin keşfedilmesi, etiketsiz spekülasyon verilerinin duygu analizi ile çoğaltılması, spekülasyonların fiyat tahmini başarımlarına olan etkileri, manipülasyon periyotlarının tespiti ve en son olarak da manipölatörlerin tespit edilmesi şeklinde takip etmiştir.

Bir sonraki bölümde anlatılan kısımda tez konusuyla ilgili en iyi çalışmalar anlatılmaktadır.



2. LİTERATÜR TARAMASI

Bu bölümde literatürde özellikle uluslararası benzer alanlarda gerçekleştirilmiş ve tez konusuyla ilgisi olan başarılı çalışmalar inceleme altına alınmıştır. Konu başlıklarına göre literatürde incelenmiş çalışmalar bu bölümde gruplandırılmıştır.

2.1 Piyasa Fiyatlamaları ve Manipülasyonlarla İlgili Literatür Çalışmaları

Çizelge 2.1 Küresel Pazar Hacimleri

Kategori	Değeri (\$ Milyar, USD)	Kaynak
Gümüş	\$44	World Silver Survey 2019
Kripto Paralar	\$244	CoinMarketCap
Küresel Askeri Harcamalar	\$1.782	World Bank
U.S. Federal Açık (FY 2020)	\$3.800	U.S. CBO (Nisan 2020 itibariyle öngörülen)
Madeni Paralar ve Banknotlar	\$6.662	BIS
Fed Bilançosu	\$7.037	U.S. Federal Reserve
Dünyanın Milyarderleri	\$8.000	Forbes
Altın	\$10.891	World Gold Council (2020)
The Fortune 500	\$22.600	Fortune 500 (2019 listesi)
Menkul Kıymetler Borsaları	\$89.475	WFE (Nisan 2020)
Kısa Dönem Para Arzı	\$35.183	CIA Factbook
Uzun Dönem Para Arzı	\$95.698	CIA Factbook
Küresel Borç	\$252.600	IIF Debt Monitor
Küresel Emlak	\$280.600	Savills Global Research (2018 yaklaşık)
Küresel Zenginlik	\$360.603	Credit Suisse
Türevleri (Piyasa Değeri)	\$11.600	BIS (Aralık 2019)
Türevleri (Nominal Değer)	\$558.500	BIS (Aralık 2019)
Türevleri (Nominal Değer-En Yüksek)	\$1.000.000	Çeşitli Kaynaklar (Resmi Olmayan)

(Desjardins 2020) çalışmasında 2020'nin ilk dönemine kadar, sanal para birimlerinin toplam hacminin küresel borsaların hacminin yaklaşık %0,1'i kadar olduğunu

paylaşmıştır. Yayınlamış olduğu tablo ile küresel pazar hacimlerini Çizelge 2.1’de gösterildiği şekilde paylaşmıştır. Bu noktada sanal para birimlerinin dünya ekonomisine kıyasla ne kadar küçük bir paya sahip olduğuna dikkat çekmiştir. Küçük yatırımcılar MACD, RSI, Trend çizgileri gibi teknik analiz araçlarını beraber kullanarak aşırı alım ve satım alanlarını tespit ederler. Böylece fiyat dalgalanmalarından kazanç sağlamaya çalışırlar. Büyük çaptaki yatırımcılar ise nispeten sığ sanal para borsalarında büyük dalgalılar yaratarak trend çizgilerine müdahale ederler. Bu trend hareketlerindeki ani değişimler yatırımcıları panikleterek piyasaları istedikleri gibi yönetme imkânı sağlar. Bu tür hareket manipülasyonun şekillerinden sadece biridir. (Keidar 2018) çalışmasında sanal para borsalarındaki borsa manipülasyonuna dikkat çekiyor ve tüm küresel piyasalar için yüksek hacimli ve değişken fiyatlandırma için bir düzenlemenin gerekli olduğunu vurgulamışlardır. (Babayan 2018) çalışmasında Bitcoin'deki anlık %85’lik değer kaybına dikkat çekilmiştir. Bu sert değer kaybının arkasındaki itici faktörün, kripto para piyasalarındaki düşük parasal hacminden kaynaklandığını belirtmiştir. İlk deneysel çalışmalarında (Gerace 2014) Hong Kong borsalarında manipülasyonun sadece tek bir tanımı olmadığını ifade etmişlerdir. (Aggarwal 2003) çalışmasında, hisse senedi manipülasyonları hakkında bazı önemli bilgiler paylaşmıştır. Bu bilgilerin bazılarında manipülasyonların potansiyel olarak kötüleşen piyasaların büyümesinde faydalı olduğunu belirtmişlerdir. Ayrıca, manipülatörlerin ise manipülasyon için kullandıkları en büyük aracın piyasalara likidite sağlamak olduğuna dair ciddi argümanlara sahip olduklarını ifade etmişlerdir. (Golmohammadi 2014) çalışmasında gözetimli makine öğrenimi algoritmalarını kullanarak borsa manipülasyonu tahmin sonuçlarıyla ilgili bazı sonuçları paylaşmıştır. Naive Bayesian algoritması kullanarak 2014 yılında %53 F_1 puanı ile en iyi sonuçları elde etmiştir. Devamında (Golmohammadi 2017), S&P 500’de yer alan petrol ve yakıt stoklarında gerçekleştirilen manipülasyonunu tespit etmek için Conformal Anomaly Detection (CAD) kullanarak performansı artırmak için de resmi bir yöntem önermişlerdir. Deneylerinde, makine öğrenmesinin yanı sıra öznitelik seçme teknikleri de uygulamışlardır. Hisse senetleriyle ilgili tweet’lerin duygu analizi yoluyla borsanın beklenen davranışını yakalayarak CAD algoritmasını geliştirmeyi amaçlamışlardır. Sonuçlara istinaden önerdikleri yöntemin yanlış pozitiflerin (FN) %28’ini ortadan kaldırdığını paylaşmışlardır. (Sridhar 2020) tarafından gerçekleştirilen çalışmada Ensemble Neural Networks kullanarak borsa

manipülasyonları tespit edilmeye çalışılmıştır. Ensemble Neural Network kullanımının Hindistan borsalarında gerçekleştirilen yöntemler için en yüksek başarıyı verdiğini, diğer yöntemlerle karşılaştırarak ifade etmişlerdir. (Chen 2019a) yatırımcılar ve düzenleyiciler için manipülasyon modellerinin varlığını araştırmıştır. Bunu kanıtlamak için de borsaların ticaret ağlarını denetlemenin gerekliliğini vurgulamışlardır. Singular Value Decomposition (SVD) kullanarak birbirleriyle ilişkili fiyat dalgalanmalarını tespit etmeye çalışmışlardır. Piyasalarda uygulanan birçok manipülasyon yöntemi olduğunu anlatıp, ayrıca kripto para piyasalarının denetiminin de güçlendirilmesi gerektiğini ima etmişlerdir. Diğer bir çalışmada (Chen 2019b) kullanıcıları algılamak için bir önyükleme algoritması önermiştir. "Pump & Dump" (Pompala & Boşalt) yönteminin bir manipülasyon yöntemi olduğunu ifade etmişlerdir. Bu şekilde manipülasyon yaparak anormal ticaret davranışı sergileyen kullanıcı gruplarını incelemişlerdir. Böylece kripto para piyasasında yatırımcılar ve anlaşıma yapıcılar için faydalı bilgiler sağlamaya çalışmışlardır. (La Morgia 2020) çalışmasında sanal para piyasalarında dolandırıcılığın nasıl işlendiğini gözlemlemiştir. Ardından kripto para piyasalarındaki düşük likiditeye dikkat çekmişlerdir. "Pump & Dump" dönemlerinde bu sahtekarlığı gerçek zamanlı olarak nasıl tespit ettiklerini paylaşarak yatırımcıların zarar görmesini engellemeye çalışmışlardır. Sonuç olarak "Pump & Dump" hareketlerini en baştan tespit etmenin mümkün olduğunu belirtmişlerdir. (Adeye 2018) çalışmasında 2004-2014 yılları arasında Nijerya'da gerçekleşmiş küresel finansal krizi ve borsa fiyat davranışlarını analiz etmişlerdir. Farklı zamanlardaki hacimsel değişiklikleri tahmin edilebilirlik açısından incelemişlerdir. Sonuç olarak, hisse senedi piyasasında fiyatların Genelleştirilmiş Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) yöntemiyle kriz öncesi dönemlerde tespit etme başarısının oldukça yüksek olduğunu belirtmişlerdir. Ayrıca kriz öncesi dönem sonuçlarındaki büyük başarının aksine, kriz sonrası dönemde ise bu başarının düştüğünü gözlemlemişlerdir. (Zhang 2020) çalışmasında, 2020 yılında Amerikan borsasında S&P 1200 Global sıralamasında yer alan hisse senetlerini "Minimum Spanning Tree" yöntemi kullanılarak incelemiştir. Ardından, araştırma süresince Covid-19'un borsaya herhangi bir olumsuz etkisinin olmadığını iddia etmişlerdir. (Kaya 2020) ile beraber yapılan çalışmamızda, Covid-19 öncesi ve sonrası fiyat tahmin yöntemlerinin başarımları karşılaştırılmıştır. Fiyatlar Covid-19 kısıtlamalarının konuşulmaya başlandığı dönemler için başarılı makine

öğrenimi ve istatistiksel tahminleme yöntemleri kullanılarak hesaplanmıştır. Deneysel sonuçlar tahmin için kullanılan çoğu yöntem için büyük başarıydı. Çalışmada 15 Mart 2020 ile 28 Haziran 2020 tarihleri arasındaki tüm piyasa yönü haftalık ve aylık olarak olacak şekilde tahmin edilmeye çalışılmıştır. Sonrasında yöntemlerin başarımları birbirleriyle karşılaştırılarak sonuçları paylaşılmıştır. Deneysel sonuçlara istinaden bir zaman serisi tahmin yöntemi olan Oto Regresif Bütünleşik Hareketli Ortalama (Box 2015) (ARIMA) 'nın en başarılı yöntem olduğu gözlemlenmiştir. (Ngai 2011) araştırmasında çeşitli finansal anormallikleri tespit etmek için yapay zekâ yöntemlerini kullanmıştır. Yapılan çalışmalar sonucunda bu alanda hala yeterli çalışma yapılmadığını belirtmişlerdir. Yanlış sınıflandırma sayısının fazla olması nedeniyle de sonuçların yeterince iyi olmadığı belirtilmiş ve bu sonuçların veri madenciliği yöntemleri kullanılarak iyileştirilmesi gerektiği sonucuna varmışlardır. Bu alanda toplanacak yeni verilerle birlikte gelecekte daha fazla akademik çalışmaya devam edilmesi gerektiğine dikkat çekmişlerdir. (Siering 2017) yaptığı çalışmada hisse senedi piyasası manipülasyonlarının güven ve bütünlük açısından finans dünyasını etkileyen en önemli tehditlerden biri olduğunu söylemiştir. Çalışmada TF-IDF ve K-means algoritmaları kullanılarak borsa külliyatında en çok ayrıştırılan kelimeler belirlenmeye ve bir sınıflandırma işlemi yapılmaya çalışılmıştır. Metin madenciliği, büyük veri analizi, yapay zekâ gibi yöntemler kullanılarak gelecekte bu sınıflama sonuçlarının iyileştirilmesi gerektiği sonucuna varmışlardır. (Cao 2014) ise, Adaptive Hidden Markov Model With Anomaly States (AHMMAS) olarak adlandırdıkları yöntemin borsadaki manipülasyonlar için çok başarılı olduğunu bildirmiştir. k-En Yakın Komşuluk (k-NN), One-class SVM (OCSVM), Gaussian Mixture Model (GMM) gibi yöntemlerden daha başarılı sonuçlar ürettiğini ifade etmişlerdir. Deneyler Google, Microsoft, Intel, Apple gibi şirketlerin borsa verilerine dayalı olarak yapılmıştır. (Lin 2015) araştırmasında, dolandırıcılıkta kullanılan 32 adet risk faktörünü değerlendirmiştir. Test için ortalama olarak 580 dolandırıcılık vakası kullanılmış ve Artificial Neural Network (ANN) ve Classification and Regression Tree (CART) yöntemlerinin en iyi sınıflandırma yöntemi oldukları ifade edilmiştir. (Kim 2016) çalışmasında Bitcoin'in değerinin 150\$'dan 1000\$'ın üzerine çıktığı 2013 yılının son iki ayındaki şüpheli faaliyetleri incelemiştir. 2011-2013 yılları arasında yaklaşık 18 milyon adet alım ve satım işlemi analiz edilmiştir. Sonuç olarak dijital para piyasasındaki

manipölasyonların gerçekten önemli bir rol oynadığını belirtmişlerdir. Bunun üzerine kripto para birimlerindeki dalgalanmaları tahmin eden bir yöntem önermişlerdir. Normal borsa ve kripto para piyasası arasındaki farklar nedeniyle borsayı tahmin etme yöntemlerinden birçoğunun başarısız olduğunu belirtmişlerdir. Çalışmada BTC, ETH ve XRP para birimlerini incelemişlerdir. Kullanıcı yorumları derecelendirilip ve etiketlenerek, Granger nedensellik testine tabi tutulmuştur. Bu sayede geleceğe yönelik tahminler yapan modüllerin, kullanıcıların tahminlerine göre fiyat tahmininde daha başarılı olduklarını savunmuşlardır.

2.2 Doğal Dil İşleme ve Duygu Analizi Yöntemleriyle İlgili Literatür Çalışmaları

(Fisher 2016) çalışmasında gelecekteki Natural Language Process (NLP) uygulamaları için bir yol haritası oluşturmaya çalışmıştır. Geçmiş çalışmaların sonuçlarını da karşılaştırarak en başarılı sınıflandırma yöntemlerini açıklamışlardır. Dolandırıcılık yöntemleri için NLP yöntemlerinin kullanılması gerektiğini belirtmişlerdir. (Albashrawi 2016), 2004 ve 2015 yılları arasında veri madenciliği yöntemlerini kullanılarak tespit edilmiş 65 dolandırıcılık vakasını incelemiştir. İncelemenin sonuçları, veri madenciliğinde finansal dolandırıcılıkları belirlemek için kullanılan tekniklerin %50'sinden fazlasında yapay sinir ağları, SVM ve Bayes ağlarının kullanıldığını ifade etmiştir. (Varol 2017), sosyal medyada trend olması muhtemel twitter hashtag'lerini belirlemeye çalışmıştır. Trend olacak hashtag'leri %75 başarı ile erkenden belirlemiştir. Trend sonrası gözetimli sınıflandırma yöntemleri kullanılarak bunların %95'ini doğru bir şekilde belirleyebileceklerini ifade etmişlerdir. Bu çalışmada Dynamic Time Warping k-Nearest Neighbors (k-NN-DTW) sınıflayıcı yöntemi kullanılmıştır. (Chheda 2018) çalışmasında, kripto para birimlerinin fiyat grafiğini insanların sosyal medyadaki düşünceleriyle ilişkilendirmeye çalışmıştır. Gözetimli makine öğrenme teknikleri kullanılarak duygu analizi kullanılmıştır. Naive Bayesian ve SVM yöntemlerinin sonuçları karşılaştırılmıştır. (Lamon 2017) araştırmasında, sosyal medyada Bitcoin, Litecoin ve Ethereum para birimleri hakkında yazılan metinleri duygu analizi ile sınıflandırmıştır. Bu duygulara göre fiyat tahmini gerçekleştirmeye çalışılmıştır. Bu çalışmada da SVM ve Naive Bayesian Sınıflandırma yöntemleri kullanılmıştır. Sonuç

olarak %75'in üzerinde başarılı sonuçlar alındığına dikkat çekmişlerdir. Belçika'da yapılan çalışmada (Karalevicius 2018), duygu analizini kullanarak Bitcoin üzerindeki günlük fiyat hareketlerini belirlemeye çalışmıştır. Sharpe oranı teknikleri kullanılarak gerçekleştirilen tahmin işlemlerinin tam olarak istenileni veremediği belirtilmiştir. (Laskowski 2016) tarafından kripto para piyasası için basit metin madenciliği yapan bir prototip uygulama geliştirilmiştir. Metinlerin analizinde para birimlerinin adlarının geçme sıklığı bu analizde esas alınmıştır. Sonuçlar para birimleri ve değerleri arasındaki ilişkiler incelenerek karşılaştırılmıştır. (Stenqvist 2017) çalışmasında, Bitcoin fiyatlarını duygu analizi kullanarak tahmin etmeye çalışmıştır. Bir aylık borsa verileri ile Tweeter'da yazılanlar arasında Lackluster tahmin yöntemi kullanılarak bir analiz yapmıştır. Sonuç olarak, tahminlerinin %83'lük bir kesinlik oranıyla doğru tahmin ettiklerini iddia etmiştir.

Bu bölüm sonrasında tez çalışması kapsamında deneylerde ve değerlendirme aşamalarında kullanılan teknik araçlar ve başarımlar metrikleri anlatılmaktadır.

3. DENEYLERDE KULLANILAN YÖNTEMLER

Bu bölümde anlatılan kısımlarda deneylerin öncesinde konu kapsamında incelenen yöntemler, deneyler esnasında ve başarımların değerlendirilmesi amacıyla kullanılmış yöntemler açıklanmıştır.

3.1 Başarım Metrikleri

Özellikle tahminleme ve yapay zekâ alanında en yaygın olarak kullanılan çalışmalar incelenmiştir. Bu çalışmalar içerisinde yöntemlerin başarımlarını ölçümlemek için en fazla kullanılan metriklerin F_1 skor, Matthews Correlation Coefficient (MCC) ve doğruluk değerlerinin oldukları gözlemlenmiştir. Bu nedenle de yukarıda bahsedilen yöntemlerin tez çalışmasının başarımlarını göstermek amacıyla kullanılması kararlaştırılmıştır. Yukarıda adı geçen hesaplamaların gerçekleştirilebilmesi için; doğru sınıflandırılmış pozitif örnek sayısı (TP), doğru sınıflandırılmış negatif örnek sayısı (TN), yanlış sınıflandırılmış pozitif örnek sayısı (FP) ve yanlış sınıflandırılmış negatif örnek sayısının (FN) bilinmesi gerekmektedir. Bu değerler standart olarak kullanılan yöntemler içerisinde yer alan sonuçlar kısmında paylaşılmaktadır.

3.1.1 F_1 skor hesaplama

F_1 skoru, bilim dünyasının kıyaslama olarak kullandığı en yaygın performans hesaplama tekniğidir. Bu değer hesaplanabilmesi için duyarlılık (precision) değeri ve anma (recall) değerinin hesaplanması gerekir. Bu sayılar hesaplandıktan sonra denklem numaralarıyla verilen duyarlılık (3.1) ve anma (3.2) değerleri bir arada yeniden hesaplanarak F_1 skor (3.3) değeri bulunur.

$$Duyarlılık = \frac{TP}{TP + FP} \quad (3.1)$$

$$Anma = \frac{TP}{TP + FN} \quad (3.2)$$

$$F_1 = 2 * \frac{Duyarlilik * Anma}{Duyarlilik + Anma} \quad (3.3)$$

3.1.2 Matthews korelasyon katsayısı (MCC) hesaplama

Bu yöntem özellikle istatistiksel hesaplamalarda ikili sınıflandırmaların kalitesini ölçmek için kullanılmaktadır. Denklem numarasıyla paylaşılan MCC (3.4) bu şekilde hesaplanmaktadır.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + TN) * (TP + FN) * (TN + FP) * (TN + FN)}} \quad (3.4)$$

3.1.3 Doğruluk değeri hesaplama

Doğruluk değeri, deneyde yapılan analizin gerçek değere ne kadar yakın olduğunu gösterir. Bu deney aynı koşullar altında tekrarlanırsa yeni sonucun önceki sonuca ne kadar benzer olacağını gösterir. Denklem numarasıyla paylaşılan doğruluk (3.5) değeri bu şekilde hesaplanmaktadır.

$$Dogruluk = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.5)$$

3.2 Borsa Teknik Analizinde Kullanılan Temel Araçlar

Sanal paraların ticaretinde bot diye tabir edilen ve belirli ekonomik göstergelere bağlı olarak ticaret yapan özel kendi kendine çalışan yazılımlar çok fazla kullanılmaktadır. Bu noktada trendin içindeki düzeltme fiyatlarını gösteren gösterge araçları kullanılmaktadır. Bu göstergeler sayesinde trend analiz edilerek, piyasadaki fiyatlama eğilimi görülmeye çalışılır. Böylelikle, yatırımcılar kendilerine en uygun olduğunu düşündükleri fiyatlardan istedikleri alım ve satım işlemlerini gerçekleştirebilmektedir. Fakat, bu yöntemlerin kullanılması fiyatlamaların her zaman doğru tahmin edileceğini garanti etmez. Piyasalarda kullanılan gösterge araçları izleyen kısımlarda anlatılmaktadır.

3.2.1 Trend çizgileri

Trend çizgileri, hissenin gidişatını tanımlamak ve onaylamak için kullanılır. Para hareketlerini gösteren grafik üzerinde en az iki fiyat noktasını birleştirilerek çizilirler. Bu çizgiler genellikle eğimli destek ve direnç alanlarını belirlemek için ileriye doğru düz bir şekilde uzanır. Eğer bu çizgiler pozitif bir eğime sahipse net talebin arttığını göstermektedir. Fiyatlar bu çizginin üzerinde kaldığı sürece yükseliş trendi devam ediyor demektir.



Şekil 3.1 Trend Çizgileri

Aynı şekilde fiyatlar aşağı eğimle çizilen trend çizgisinin altında kalan bölümde ilerlemeye devam ediyorsa net satışın arttığını yani düşüş trendinin devam ettiğini göstermektedir. Fiyat, genellikle eğimli bir trend çizgisini birkaç kez tekrar test eder, hatta bazı noktalarda kırabilir. Bu noktalarda belirlenmiş bir periyot boyunca keskin bir düşüş gözlenmeye devam ederse, bu trend çizgisinin kırıldığını ve yeni bir trendin başladığını ifade eder. Bu trendler farklı zamanlar için Şekil 3.1’de gösterildiği gibi farklı renklerde ayrı ayrı çizilebilmektedir.

3.2.2 Göreceli güç endeksi (RSI)

Dikey ekseninde 0 ile 100 arasında bir aralıkta hareket eden gösterge için 30 ve 70 fiyat seviyesinde iki sınır çizilmiştir. 30 seviyesinin altında bulunan Relative Strength Index (RSI); fiyatla uyum sağlamayan aşırı satış talebini gösterirken, 70 seviyesi üzerine çıktığında ise aşırı alım talebini gösterir. Bu sayede fiyat hareketinin yön değiştireceğine ilişkin önemli bilgiler verir. RSI bu hesabı yaparken belirtilen periyot içindeki günlere ait düşüşleri ve yükselişleri ayrı ayrı toplar, ortalamalarını hesaplar ve mevcut fiyat seviyesi ile karşılaştırır. Bu sayede yatırım aracının kendi iç potansiyelinin talebi

karşılayıp karşılayamayacağını anlamaya çalışır. Geçmiş talepler karşısında oluşan fiyatlamalar RSI için en önemli veridir. Bu göstergenin temeli de belirtilen bu iç potansiyele dayalıdır. Her yatırım aracı için oluşan alım veya satım talepleri farklı büyüklüklerdeki fiyat hareketlerine sebep olabilir. RSI, bu noktada yatırım aracının bu fiyat değişimlerini geçmiş günlerin ortalamasıyla hesaplar ve aynı gün içindeki taleplerin yaratacağı etkiyi tahmin etmeye çalışır. Dolayısıyla, RSI mevcut fiyatın talebi karşılama yeteneğini belirli aralıklar içerisinde gösterirken, bu sınırların dışına çıkan taleplerin aşırı olacağını ve bir süre sonra fiyatlamasının bu noktalarda kalıcı olamayacağını öngörür. Bunu da yukarıda 70, aşağıda 30 puan olacak şekilde Şekil 3.2'deki gibi limitler ile belirtir. Bu limitlerin dışındaki değerlerin kısa vadede bir değişim olacağını göstergesi olduğu kabul edilir.



Şekil 3.2 RSI İndikatörü

3.2.3 Hareketli ortalamaların yakınsama ayrımı (MACD)

Moving Average Convergence Divergence (MACD) göstergesi; kısa dönemdeki fiyat eğilimini, uzun dönemdeki fiyat eğilimi ile birlikte karşılaştıran ve son dönemdeki piyasa değişimini anlatan bir göstergedir. Standart olarak 12 günlük ortalamanın 26 günlük ortalamadan çıkarılması ile elde edilen sonuç, sıfır çizgisi üzerinde mavi ve kırmızı çizgiler ile betimlenir. Pozitif yöne kayan mavi çubuklar boğa piyasası gücünü gösterirken, negatif yöne ilerleyen kırmızı çubuklar ayı piyasasının gücünü göstermektedir. Piyasaların eğilimlerini anlaşılır bir şekilde sunan MACD bu nedenle piyasa analizlerinde sıkça tercih edilmektedir.



Şekil 3.3 MACD İndikatörü

MACD bu iki karşılaştırma dışında 9 günlük üssel ortalamayı da hesaplayabilmektedir. Bu nedenle de çok kısa vadedeki piyasa değişimini gösterebildiğinden, piyasada yaşanabilecek değişimleri erkenden gösterebilme özelliği vardır. Böylelikle yatırım yapılacak giriş noktalarının seçiminde fazlasıyla tercih nedeni olur. Üç farklı ortalamayı Şekil 3.3'te görülen grafiğe ait mumların alt tarafında tek bir görsel alanda rahatça okunabilir biçimde sonuçlar üretmesi ve ilişkilendirmesi nedeniyle tavsiye edilen MACD, gün sayıları değiştirilerek de arzuya göre özelleştirilebilmektedir.

3.2.4 Üssel hareketli ortalama (EMA)

Exponential Moving Average (EMA), borsada genelde bir hissenin eski fiyatlarına nazaran yeni fiyatlara daha fazla ağırlık ekler. Bu şekilde yatırımcı hızlı piyasa değişikliklerinde ani karar vermesi gerektiği durumlarda son verilere daha önemli olarak baktığından büyük yardımı olur. Teknik analiz araçlarının çoğu borsa hareketlerini geç vermesinden ötürü, bu araç tercih sebebi olmaktadır. Genelde 200 günlük veya senelik hareketli ortalamalar çok dikkate alınmaktadır.

3.3 Çalışma Konusu ve Kapsamında Değerlendirilen Yöntemler

Gerçekleştirilen tez çalışması bilimsel alanda başarısını ispat etmiş, güncel ve yaygın olarak kullanılan yöntemleri içermektedir. Duygu analizi yöntemleri, istatistiksel ve yapay zekâ tahminleme araçları deney süreçlerinde birçok değerlendirme yapılması için kullanılmıştır. Ayrıca, kapsam olarak günümüzde çok gündem teşkil eden sanal paraları içermesinin yanı sıra, literatürde kısa vadeler içerisinde fiyatlarda yaşanmış olan ani artış ve azalış konusundaki sorulara cevap verecek kapsamda bir çalışma olması sebebiyle de ön plana çıkmaktadır. Bu noktada; tahminleme yöntemlerinin yaşanmış bu fiyat dalgalanmalarını doğru bir şekilde analiz edebilme yetenekleri kanıtlanmaya çalışılmış, sonrasında ise manipülasyon konusunda ne kadar yetkin olarak kullanılabilecekleri araştırılmıştır. Bu çalışmaların yanında sosyal medya üzerinden gerçekleştirilen spekülasyon hareketleri de beraber araştırılmıştır. Araştırma esnasında

manip lat rler tespit edilirken, anomali tespiti kullanılması ne kadar anahtar bir rol oynayacağı da ortaya  ıkarılmıştır.

3.3.1 Duygu analizi y ntemleri

Veri madenciliđi alanında duygu analizi y ntemleri sık a kullanılmaktadırlar.  zellikle kurum ve kuruluřların kiřisel verileri saklamaya bařlamasıyla, b y k veri ambarları ve b y k veri kavramları ortaya  ıkmıştır. Bu kurum ve kuruluřlar ge miřten, g n m ze kadar hizmet kalitelerini y kseltebilmek i in ellerindeki verilerin anlamlandırılması konusunda  alıřmalar yapmıřlardır. Bunlar gerek el yordamıyla gerekse de bilgisayar y ntemleriyle yapılmaktadır. Duygu analizi; ses, metin, resim, video vb. t m verilerde yaygın olarak uygulanmaktadır. Ger ekleřtirilen bu  alıřmada metin madenciliđinde kullanılan duygu analizi y ntemleri incelenmiřtir. G n m zde yapay zek  algoritmalarının da geliřmesiyle beraber bilgisayar destekli duygu analizi y ntemleri daha hızlı ve daha bařarılı sonu lar vermeye bařlamıştır. Bu nedenle de yapılan  alıřmada, literat rde yer alan bařarılı duygu analizi y ntemleri incelenmiř ve tezin konusu bakımında en bařarılı olabilecek y ntemler deđerlendirilmiřtir.

3.3.2 Otomatik regresif entegre hareketli ortalama (ARIMA-SARIMAX) y ntemi

 zellikle zaman serilerinin analizinde  n plana  ıkan ve istatistiksel temalı bir tahminleme aracı olan Autoregressive Integrated Moving Average (ARIMA), bu tez  alıřmasında kullanılan verilerin t r  sebebiyle tercih edilmiřtir. Sonu ta Bitcoin i in kullanılacak olan verilerin de bir zaman aralıđını i ermesi, ARIMA y nteminin deneyler s recinde kullanılmasında rol oynamıştır. Bunun yanında derin  đrenme tekniklerinin g n m zde yaygın olarak kullanılmasına rađmen (Fu 2016) yaptığı bir  alıřmada ARIMA y nteminin daha bařarılı olabileceđini belirtmesi, modelin halen bařarılı sonu lar vermesi konusunda g ncelliđini koruduđunu kanıtlamaktadır. Kısaca ARIMA, aslında belirli bir zaman serisini kendi ge miř deđerlerine, yani kendi gecikmelerine ve gecikmeli tahmin hatalarına dayanarak a ıklayan bir model sınıfıdır.

Böylece gelecekteki değerleri tahmin etmek için bu değerleri kendi denklemi içerisinde kullanmaktadır. ARIMA'nın başka bir versiyonu olan Seasonal Auto-Regressive Integrated Moving Average (SARIMAX) da (p, d, q) ifadeleri şeklinde tanımlanmıştır. Trendin otoregresyon sırasını 'p', trendin farkının sırasını 'd' ve trendin hareketli ortalama sırasını 'q' belirtir. Tek fark olarak sezonsal parametreyi içermesidir.(3.6) ve (3.7) de gösterilen denklemlerde; y gözlem dizisini, ϕ ve θ modelin sırasıyla AR ve MA parçalarının parametrelerini, ϵ tahmin hatalarını, μ kesişim parametresini ve Δ fark alma operatörünü belirtir.

$$y_t^* = \Delta^d y_t \quad (3.6)$$

$$y_t^* = \mu + \underbrace{\sum_{i=1}^p \phi_i y_{t-i}^*}_{\text{AR}} + \underbrace{\sum_{i=1}^q \theta_i \epsilon_{t-i}}_{\text{MA}} + \epsilon_t \quad (3.7)$$

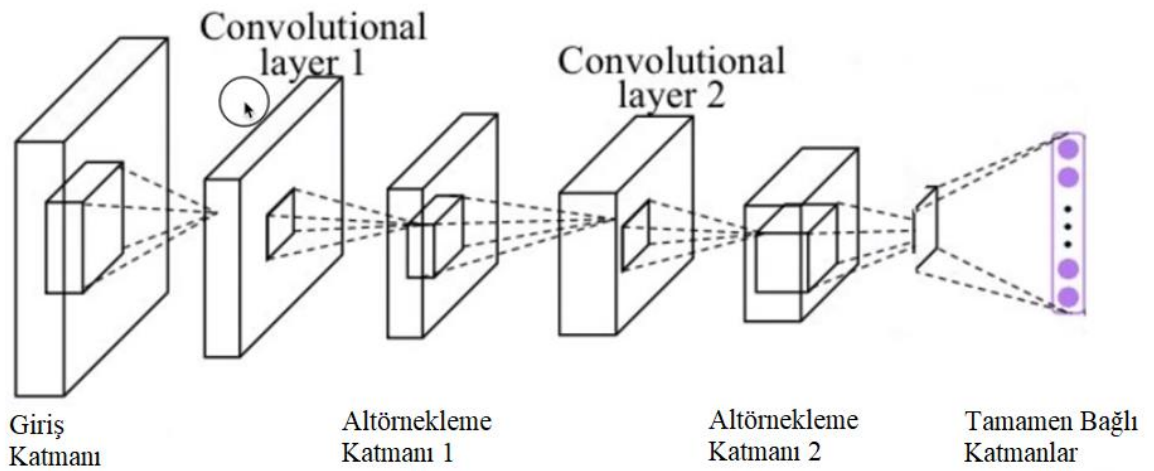
3.3.3 Derin öğrenme yöntemleri

Günümüzdeki en yaygın ve güncel makine öğrenimi yöntemlerinden biri olarak kullanılan derin öğrenme yöntemleri, yapay zekanın ilk temelleri olan perceptron (Algılayıcı) öğrenme algoritmasının özelleştirilmesiyle oluşmaya başlamıştır. Bu yöntemlerdeki temel mantık insanlardaki sinirlerin çalışması gibi her belirlenmiş perceptron için uygun olan eşik değerinin bilgisayar tarafından yapılan yüksek sayıdaki işlemlerle tespit edilip, bir matematiksel model oluşturulmasına dayanmaktadır. Sistem bu kümeler arasında bir ilişki kuracak ve sonrasında oluşturmuş olduğu model sayesinde başka verilerin ne gibi çıktılar üretmesi gerektiğini tahmin etmeye çalışacaktır. Yöntemin birçok probleme özgü optimizasyonu yapılmış algoritmaları mevcuttur. Herhangi bir derin öğrenme modeli hakkında hatırlanması gereken en önemli şeylerden biri, eğitmek için büyük miktarda veri ve çok fazla bilgi işlem kaynağı gerektirmesidir. Ayrıca, gürültünün olmadığı ve lineerliğin yüksek olduğu verilerde başarı oranları oldukça yüksektir. Araştırmacılar kendi problemlerinin çözümü için özelleştirilmiş derin öğrenme yöntemleri kullanarak veya kendi problemlerine uyacak

şekilde parametreleri değiştirerek ve yeni katmanlar ekleyerek daha başarılı sonuçlar elde edebilmektedirler. Yapılan tez çalışmasının duygu analizi kısmında, ele alınan problemin uygunluğuna göre iki farklı yöntem deneylerde kullanılmak üzere seçilmiştir. Bunlar Convolutional Neural Network (CNN) ve Deep Neural Decision Forests (DNDF) yöntemleridir. En son kısımda yer alan fiyat analizi için de zaman serileri problemlerinde başarılı çalışmalarda kullanılmış Long short-term memory (LSTM) yöntemi, tez çalışmasına uygunluğu açısından kullanılmıştır.

3.3.3.1 Evrişimli sinir ağları (CNN)

CNN ilk olarak 1980'lerde geliştirilip kullanılmaya başlanmıştır. İlk zamanlarda bir CNN algoritmasının yapabileceği en fazla şey elle yazılmış rakamları tanımaktı. Çoğunlukla da posta sektörlerinde posta kodlarını, pin kodlarını okumak için kullanılmıştır (Srihari 1989).



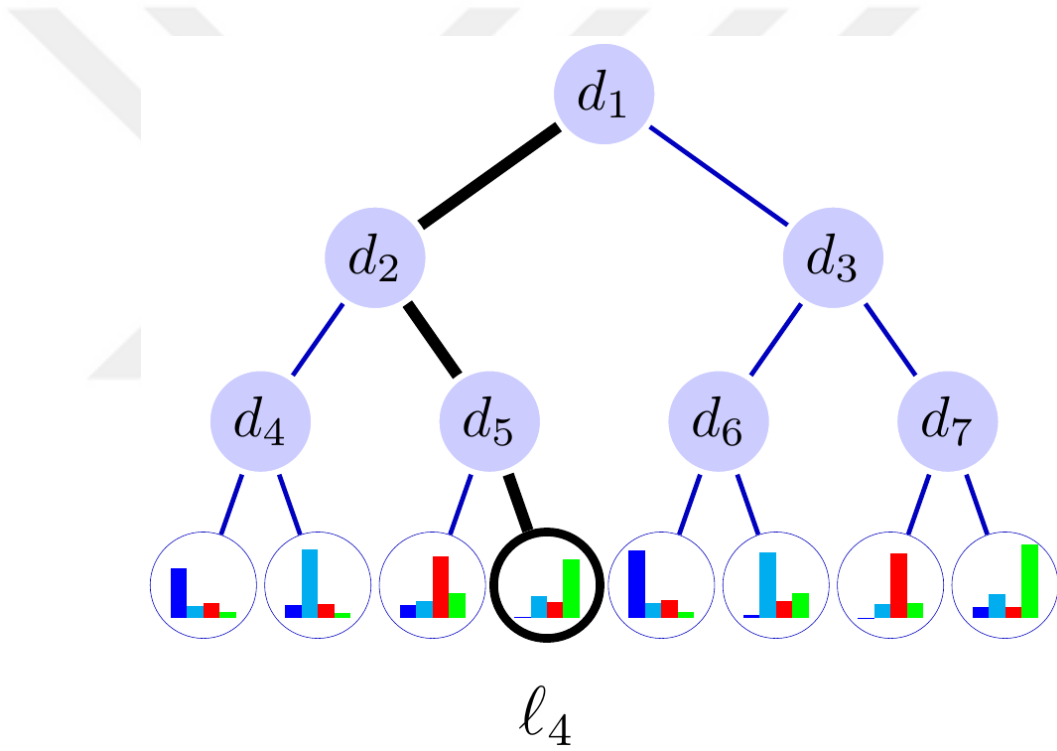
Şekil 3.4 Basit CNN Yapısı

Bu durum da o dönemde CNN'ler için büyük bir dezavantajdı ve bu nedenle CNN'ler yalnızca posta sektörleriyle sınırlıydı ve makine öğrenimi dünyasına giremedi. Bir sinir ağını düşündüğümüzde ilk başta matris çarpımlarını düşünürüz. Fakat gerçekte CNN'de durum böyle değildir. “Convolutional” (Evrişimli) adı verilen özel bir teknik kullanır.

Matematikte “Convolution” (Evrşim); birinin şeklinin diğeri tarafından nasıl değıştirildiğini ifade eden, üçüncü bir gizli fonksiyon üreten iki fonksiyon üzerinde gerçekleştirilen matematiksel bir işlemdir. Şematik gösterimi Şekil 3.4’te yer almaktadır.

3.3.3.2 Derin sinirsel karar ormanı (DNDF) yöntemi

İlk kez 2015 yılında ortaya konan DNDF (Kontschieder 2015) yöntemi, özellikle görsel veriler üzerinde almış olduğu başarılı sonuçlar ile ön plana çıkmaktadır.



Şekil 3.5 Örnek DNDF Yapısı (Kontschieder 2015)

Bu yöntemdeki temel amaç, girdilerdeki öğrenme kat sayısı düşük olan öznetelikleri hesaplayıp, bu dallarda yer alan sinirleri modelleme esnasında keserek mevcut başarımların oranlarını yükseltmektir. Bunu yaparken de Back-Propagation (Geri Yayılım) algoritması kullanarak, nöronlardaki ağaçların içinde bir hesaplama

gerçekleştirmektedir. Şekil 3.5'te gösterildiği üzere d karar ağacı numarasını, l ise yaprağı (leaf) ifade etmektedir.

3.3.3.3 Uzun kısa-süreli bellek (LSTM) yöntemi

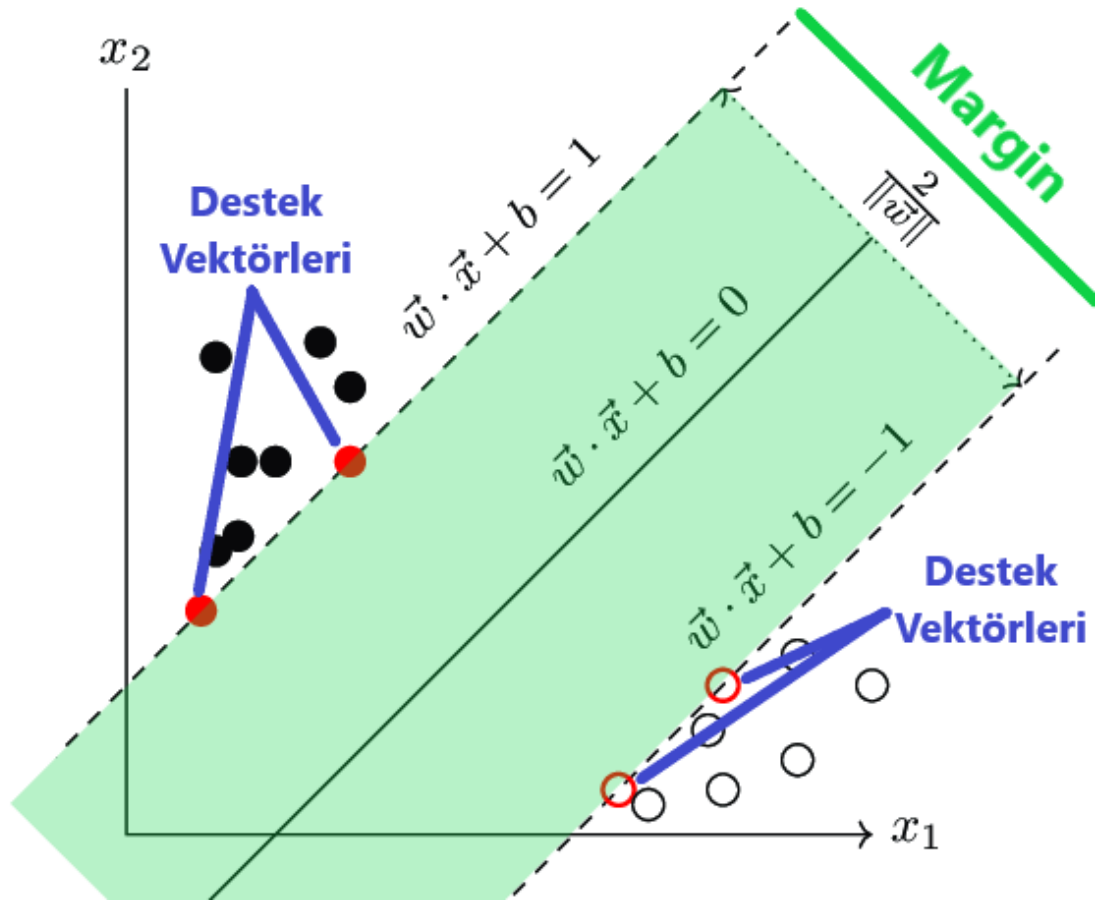
Recurrent Neural Network (RNN) yönteminin özelleştirilmesiyle ortaya çıkan LSTM (Hochreiter 1997), özellikle zaman serileri hesaplamalarında yaygın olarak kullanılmaktadır. Bunun nedeni de RNN'nin sahip olmuş olduğu kısa vadeli hafızanın geçmiş bilgileri hatırlamasında yetersiz kalmasıdır. LSTM bu noktada bilgileri daha iyi depolayarak, uzun zaman serileri problemlerinde daha başarılı sonuçlar üretebilmektedir. Giriş ve çıkış kapıları haricinde unutma kapısı değerleri gibi parametreler ile kullanılır. Giriş kapısı hücre durumunun, önceki ve mevcut verilerin sigmoid işlemi sonucuna göre güncellenip güncellenmeyeceğini belirler. Çıkış kapısı ise bir sonraki hücrenin girişini belirler. Ayrıca tahmin yapmak için de kullanılır. Unutma kapısı ise hangi bilgilerin unutulacağını belirleyen mekanizmadır. Kapılar, verileri 0 ile 1 aralığına sıkıştıran sigmoid aktivasyon fonksiyonunu kullanır. Aktivasyon fonksiyonu burada bir nöronun aktif olup olmayacağına karar vermek için kullanılmaktadır.

3.3.4 Diğer makine öğrenimi yöntemleri

Makine öğrenimi yöntemleri yapay zekâ ile ilgili konularda en yaygın kullanılan yöntemlerden biridir. Gözetimli, yarı-gözetimli ve gözetimsiz olmak üzere çeşitli makine öğrenimi literatürde yer almaktadır. Tez çalışmasının hem duygu analizi hem de fiyat tahmini kısımlarında kullanılmak üzere en başarıyla kullanılan yöntemler değerlendirmeye alınmıştır. Bu noktada Support Vector Machine (SVM) ve Random Forest (RF) algoritmalarının benzer alanlara ait literatürdeki yüksek başarımlarının özellikle ön plana çıkmasından ötürü, çalışmaların ileriki kısımlarında kullanılmışlardır.

3.3.4.1 Destek vektör makineleri (SVM)

SVM benzer özelliklere sahip, sınıflaması zor olan veriler için kullanılabilen başarılı bir gözetimli makine öğrenme algoritmasıdır. SVM algoritmasında her bir veri ögesi, her özneliğin değeri belirli bir koordinatın değeri olacak şekilde n-boyutlu uzayda birer nokta olarak temsil edilir. Daha sonra bu sınıfları belirleyebilmek için destek vektörleri kullanılır. Özelleştirilmiş kernel fonksiyonları üzerinden gerçekleştirilen bu hesaplamalar ile sınıflanacak kategoriler için hiper düzlemler bulunarak sınıflandırma işlemi gerçekleştirilir. Bu hiper düzlemlerin arasında kalan “margin” adı verilen bir boşluk bulunur.



Şekil 3.6 SVM Çalışma Mantığı (Akca 2020)

Bu boşluk ne kadar geniş olursa sınıflama işlemi o kadar başarılı olur. Şekil 3.6’da çoklu düzlemin ağırlık vektörünü w , veri noktasını x ve b ise dengeyi ifade etmektedir.

“Kernel” adı verilen çeşitli farklılaşmış SVM fonksiyonları yer almaktadır. Bu fonksiyonlar doğrusal olarak ayrılmayan sınıfları belirleyebilmek için kullanılırlar. Verilerin türüne göre birden fazla çekirdek (Kernel) fonksiyonu kullanılarak ihtiyaca en uygun yöntemin tespiti yapılabilmektedir. Gerçekleştirilen tez çalışmasında Linear SVC (Support Vector Classifier), Sigmoid SVC, Poly-SVC, Nu-SVR ve Epsilon SVR (Support Vector Regression) fonksiyonları en başarılı yöntemin tespiti amacıyla kullanılmıştır.

3.3.4.2 Rastgele orman (Random Forest) algoritması

Metin madenciliğinde yaygın olarak karşılaşılan kategorileri birbiri içine geçmiş veriler arasında oldukça başarılı sınıflama yapabilen RF (Breiman 2001) algoritması popüler makine öğrenmesi algoritmalarından biridir. Kökeni karar ağaçlarına dayanmaktadır. Fakat geliştirilen bu yeni algoritmada, ilham aldığı yöntemin sahip olduğu fazla öğrenme problemi giderilmiştir. Bunu aşmak için de öznitelikler kümesinden yüzlerce rastgele farklı alt kümeler seçerek veri ormanları oluşturur. Sonrasında her biri için oluşan karar ağacı sonuçlarını; eğer sorun regresyonsa tahmin ortalamasını, değilse tahminler arasında en çok oy alanı seçerek belirlemektedir. Bu algoritmanın başka bir avantajı da her bir özneliğin sınıflama için olan önemini çok kolay ölçebilmesidir. Bu şekilde verilerdeki kirlilik oranını azaltarak en iyi sonuçlara ulaşılmasının önü açılmaktadır. Bu tez çalışmasında özellikle duygu analizi kısmında sınıflanacak metin verilerindeki kirliliği azaltacağı düşünülerek, kullanılması kararlaştırılmıştır.

3.3.5 Öznitelik seçme metrikleri (Feature selection metrics)

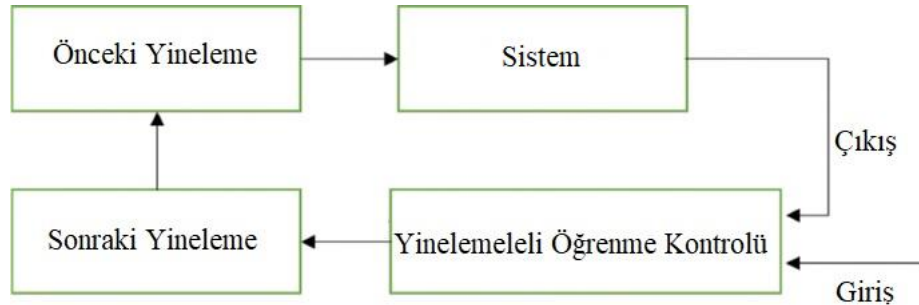
İstatistik bilim alanı içerisinde matematiksel denklemler ile ifade edilebilen öznitelik seçme metriklerinin amacı; sınıflandırma için kullanılacak veriler içerisinde sistemin öğrenmesine en çok katkı yapan öznitelikleri detaylı olarak puanlandırarak, en başarılı sonuçların alınmasını sağlamaktır. Bu aşamada birçok yöntem ön plana çıkmaktadır. Term Frequency-Inverse Document Frequency (TF-IDF), Information Gain (Bilgi

Kazanımı) ve Chi-Square (Kikare) (Wilson 1931) gibi algoritmalar oldukça popüler olarak literatürde kullanılmaktadır. Bu tez çalışmasında, daha önceki benzer metin verilerinden yüksek başarıyla öznitelikleri her bir kategori için ayrıştırma gücüne sahip Kikare algoritmasının kullanılması uygun görülmüştür. Öznitelikler başarıyla ayrıştırıldıktan sonra, borsa terimleri için en başarılı öznitelik sayısı ve bu kategoriler için en önemli kelimeler tespit edilmiştir. Böylelikle, daha başarılı sınıflama yapılmasının önü açılmıştır. Denklem 3.8’de bu tez çalışmasında kullanılan Kikare hesaplama yöntemi verilmektedir. 'X' Yunan alfabesinden gelmektedir. Öznitelik puanını ifade etmektedir. 'O' gözlemlenen değer anlamına gelir ve 'E' ise beklenen değer anlamına gelir. Her bir kategori için Kikare değeri her bir kelime için bu formül ile hesaplanmaktadır.

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (3.8)$$

3.3.6 Yinelemeli öğrenme kontrol teknikleri

Iterative Learning Control (ILC), belirli bir referans noktasını izleyen veya tekrar eden bir bozukluğu reddeden ileri beslemeli bir kontrol mekanizmasıdır. Genelde elektronik sinyal verilerinin işlendiği devreler üzerinde iyileştirmeler yapmak için kullanılırlar (Bondi 1988). Fakat, Şekil 3.7’de gösterilen bu iş modeli, bu tez çalışmasında metin verilerinin sınıflandırılması için uyarlanmıştır.



Şekil 3.7 Yinelemeli Öğrenme Kontrolü

ILC, iyi tasarlanmış bir geri besleme ve ileri besleme kontrolü sayesinde çeşitli avantajlara sahiptir. ILC önceki yinelemelerden öğrenerek tekrarlayan bozulmaları önceden telafi edebilir. Yalnızca bu verilerin her bir yinelenmeden sonra gerçekten iyileşip iyileşmediğini kontrol etmek gerektirir. Bir geri besleme denetleyicisi sistem modelindeki varyasyonları veya belirsizlikleri barındırabilirken, bir ileri beslemeli denetleyici yalnızca sistemin iyi çalıştığı ölçüde iyi performans gösterir. ILC, açık döngü kontrolünü uygulama (yineleme alanında geri bildirim) yoluyla oluşturduğundan, bu yüksek performanslı kontrol, sistem belirsizliklerine karşı da oldukça dayanıklıdır. Aslında, ILC sıklıkla lineer modeller varsayılarak tasarlanır ve genellikle sistem çözünürlüğü sırasına göre düşük izleme hataları veren doğrusal olmayan sistemlere uygulanır. Bununla birlikte, ILC her durumda mükemmel izleme sağlayamaz. En önemlisi, regresyon ve tekrar etmeyen rahatsızlıklar performansını engeller. Bununla birlikte, geri besleme kontrolünden farklı olarak, ILC'nin yinelenmeden yinelemeye öğrenmesi, ona gereksiz öznitelikleri filtreleme fırsatı sağlar. Örneğin, önemi azalan bir öznitelik bu sayede geri planlara itilerek değerlendirme kapsamından çıkarılır. Sonuç olarak sınıflama işlemlerindeki başarı oranları artmaya başlar. Tekrarlamayan bozulmaları önlemek için, ILC ile birlikte kullanılan bir geri besleme kontrolörü en iyi yaklaşımdır.

Bir sonraki bölümde tez ile ilgili gerçekleştirilmiş tüm çalışmalar detaylı bir şekilde anlatılmış ve sonraki adımlara geçilirken yaşanmış süreçler irdelenmiştir.

4. DENEY ÇALIŞMALARI VE SÜREÇLERİ

Bu kısımda yer alan alt başlıklar, deney aşamalarında kullanılan bakış açılarını, verilerin anlamlandırılması, deney aşamalarının hazırlanışını ve yöntemlerin eldeki veriler üzerinde nasıl uygulandığı ve deneylerde alınan sonuçlarla ilgili detayları anlatmaktadır.

4.1 Deney Ön Hazırlık Çalışmaları

Bu bölümde, deneyler esnasında izlenmiş yollar, yöntemler ve teknik olarak imkansız olan kısımlar anlatılmaktadır.

4.1.1 Ana fikir

Literatürdeki benzer çalışmaların konuyu ele alış biçimleri incelendiğinde; sadece duygu analizi üzerinden tespit edilmeye çalışıldığı, ya da fiyatlar üzerinden tahminleme yöntemlerinin sonuçlarına istinaden araştırmalar yapıldığı gözlemlenmiştir. Bu noktada yapılan çalışmaların başarımlarına bakıldığında ise büyük oranda farklılıklar yaşandığı dikkat çekmektedir. İlk etapta bu çalışmalarda başarımların nereden kaynaklanabileceğine dair incelemeler yapılmıştır. Özellikle sanal para piyasalarında sık sık manipülasyon yapıldığı başka çalışmalarda dile getirilmiş olsa da yapılmış çoğu çalışmada bu faktörün detaylı olarak ele alınmadığı ortaya çıkmıştır. Çoğunlukla bu problemden yüzeysel olarak bahsedilerek, manipülasyon konusun karanlık ve incelenmesi gereken bir nokta olarak bırakılmıştır. Bu aşamada çalışmaların derinleştirilmesi gerektiğini savunan çalışmalar da göz önünde bulundurularak; konu kapsamı manipülatörlerin tespitini de içerecek şekilde daha da genişletilip, doktora tez konusu olarak araştırılmaya başlanmıştır.

4.1.2 Teknik olarak imkânsız olan durumlar

Çalışmanın metodoloji kısımlarının belirlenmesinden önce, tez konusu kapsamı içerisinde yer alan tüm problemlerin çözümü noktasında elde kullanılabilecek tüm verilerin varlığı araştırılmıştır. Bu süreçte tahminleme yapılabilmesi için gerekli olan veriler aşağıda gösterildiği üzere listelenmiştir.

- Twitter üzerinde yer alan gerçek veya tüzel kişiler tarafından Bitcoin fiyat tahminleri ile ilgili paylaşılan gerçek kullanıcılar tarafından etiketlenmiş metin verileri
- Bitcoin üzerinde gerçekleştirilmiş bir manipülasyon periyodu
- Geçmişte yapılmış tüm Bitcoin alım-satım hareketlerini detaylarıyla beraber içeren veriler
- Manipülatör hesap bilgileri
- Geçmişte yer alan tüm Bitcoin'e ait günlük açılış, kapanış, en düşük ve en yüksek fiyatlama verileri

Yukarıda listede ihtiyaç duyulan bilgilerin tamamının gerçek yaşamdaki verilerden oluştuğu dikkat çekmektedir. Listedeki verilerin elde edilmesi esnasında bu bilgiler içerisinde sadece manipülatör hesapların bilgilerine ulaşamamıştır. Çünkü herhangi bir manipülatör verisi henüz resmi ve tamamen doğrulanabilir bir kaynak tarafından paylaşılmamıştır. Burada yaşanan doğrulama sorununun çözümü aşamasında çalışmanın metodoloji kısmında başka doğrulama yöntemlerin kullanılması gerektiği fark edilmiştir. Araştırmalar devam ederken, (Buy Bitcoin Worldwide 2019) çalışmasında belirtildiği üzere 2019 itibariyle yaklaşık 400 milyon adet Bitcoin hesabının olması nedeniyle kümeleme yöntemlerinin kullanılmasının hem harcanacak zaman hem de elde edilecek sonuçları doğrulayabilecek bir veri olmaması nedeniyle kullanılmasının uygun olmadığı anlaşılmıştır. Ayrıca çalışmanın spekülasyonlarla ilgili kısımlarında alternatif yöntem olarak başarı karşılaştırılmalarında kullanılmak üzere; (Liu 2018) tarafından yapılan çalışmada kullanılan Rough Set Exploration System (RSES) modellerinden (Bazan 2000) MODLEM (Grzymala-Busse 2001) yöntemleri test edilmek istenmiştir. RSES yöntemlerinin hesaplanmasının zaman ve kaynak açısından çok maliyetli

olduğunu yazarlar özellikle belirtmişlerdir. Eğitim hesaplaması için; veri kümesinin verilmesinden bir süre sonra yöntemin sistemin tüm kaynaklarını kullanarak uygulamanın hata verdiğine sebep olduğu görülmüştür. Bu nedenle, çalışmamızda kullanılacak verilerin boyutu düşünüldüğünde bu karşılaştırmanın mevcut kaynaklar ile yapılamayacağı sonucuna varılmıştır.

4.1.3 Metodoloji

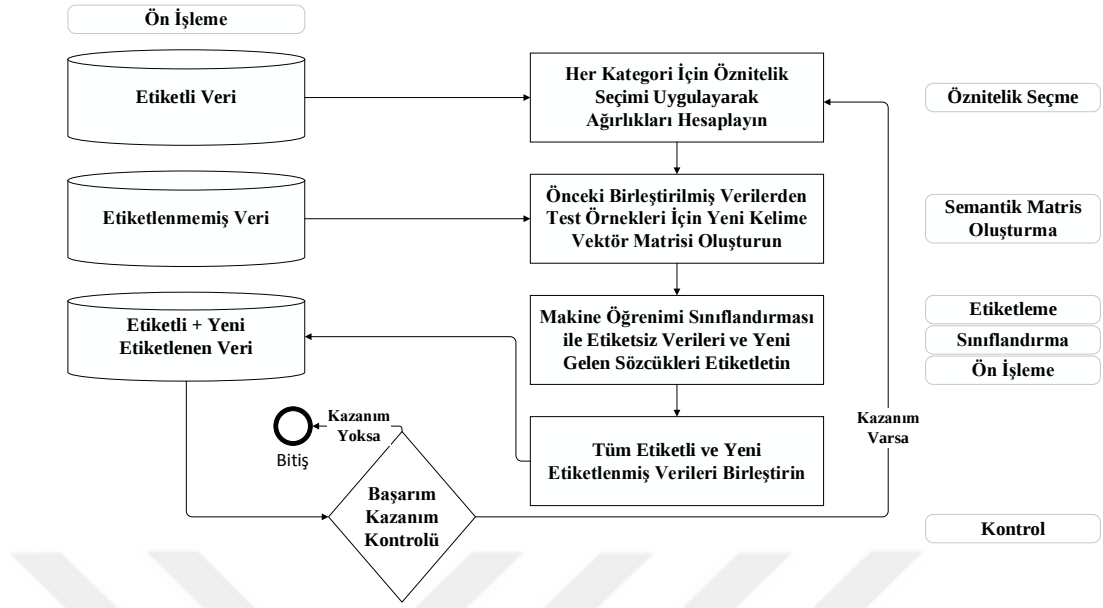
Çalışmanın ana fikrinde belirtilen problemin çözümüne katkı sunabilecek tüm argümanlar belirlendikten sonra, ilk etapta verilerin elde edilmesi için çalışmalara başlanmıştır. Sonrasında eldeki bu verilerin en iyi şekilde nasıl ana amaca hizmet edebileceği araştırılmaya başlanmıştır. Bu noktada, benzer çalışmalarda paylaşılan yol haritalarının bu problemin çözümünde yetersiz kaldığı sonucuna ileride gerçekleştirilen deneylerin ışığında varılmıştır. Böylelikle yeni bir yöntem ihtiyacı doğmuştur. Yeni oluşturulacak manipülatör tespit modeli, literatürde manipülasyon tespiti için en çok kullanılan yöntemleri içermeli ve başarımlara etkisi olduğu düşünülen iddialara doğrulanabilecek şekilde cevap verebilmelidir. Bunu sağlayabilmek içinse; duygu analizi, zaman serileri tahmini, makine öğrenimi gibi en yaygın kullanılan yöntemler deneylerin her bir aşamasında farklı çeşitleriyle beraber tek tek test edilip değerlendirilmiştir. Bu değerlendirme sonuçlarına istinaden; en başarılı yöntemlerin sonuçları modelin ilgili kısmında kullanılmak üzere sonuçlar kısmında paylaşarak ve modelin bir sonraki kısmında aynı süreçler tekrar edilerek deneylere devam edilmiştir. Ardı ardına yapılan en başarılıyı bulma deneylerinin sonucunda, tezde ortaya konan problemin çözümü için ortaya çıkan en son model paylaşılmaktadır.

Gerçekleştirilen deney çalışmalarında iki farklı noktada çalıştırmalar yoğunlaştırılmıştır. Bunlardan ilki; sosyal medya spekülasyonları etkisinin ölçülmesi amacıyla gerçekleştirilen çalışmalardır. Sosyal medya platformları üzerinde yazılmış olan bu verilerin ne kadar başarılı sınıflanabileceğine dair literatürdeki mevcut yöntemler gerçekleştirilen deneyler aracılığıyla test edilmiş ve bu yöntemlerin tek başına kullanılmasının iyi başarımlar sonuçları vermediği paylaşılmıştır. Sonrasında bu

sonuların iyileřtirilmesi amacıyla alıřmalar yapılmıř ve ISSFS yntemi bu tez alıřmasında ilk kez ortaya atılmıřtır. Yapılan alıřmalar neticesinde mevcut duygu analizi yntemlerine dahil edilerek gerekleřtirilen deney alıřmalarında yaklařık olarak %8 ile %10 F_1 skor deęeri artıřı saęlayan bu yntem hem iřlemleri hızlandırma aısından hem de bařarımı yksek oranda arttıran bir yntem olarak n plana ıkmıřtır. Alınan bařarılı sonulara istinaden speklasyonların deęerlendirilmesi, mevcut yntemlerin ISSFS yntemiyle harmanlanarak en bařarılı yntemin belirlendięi bir modelle ifade edilmiřtir. İkinci olarak ise; fiyatlama verilerinin analizi esnasında kullanılacak yntemlerin bařarımlarının deęerlendirilmesi amacıyla; ilk deney alıřmalarında sosyal medya speklasyonları ile ilgili en bařarılı sınıflamaya istinaden oluřan veriler bu adımdaki alıřmaya dahil edilmiřtir. Bylelikle, speklasyonların genel sistem bařarımına olan etkilerinin ortaya ıkarılması ile ilgili birtakım sonular paylařılmıřtır.

4.1.3.1 İteratif yarı gzetimli znitelik seme yntemi

Bu tez alıřmasında ilk kez ortaya konan bu yntem (ISSFS); aslında iteratif ęrenme tekniklerinin, yarı-gzetimli znitelik seme yntemleri ile beraber kullanılması ile oluřturulmuřtur.



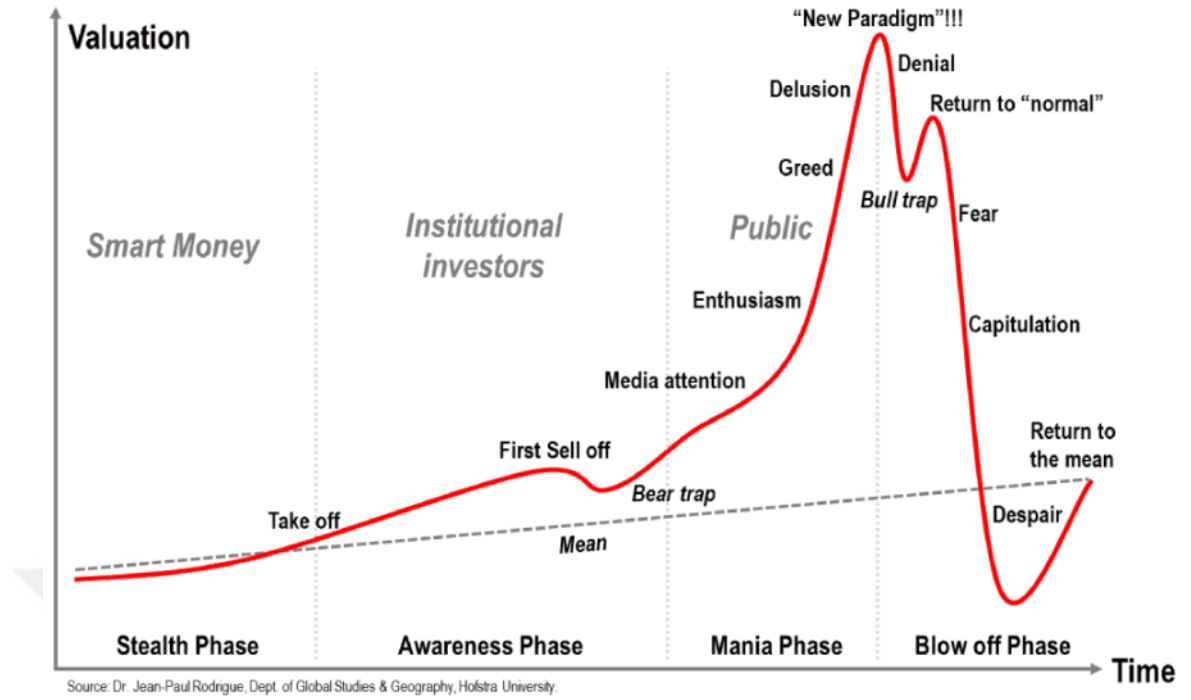
Şekil 4.1 ISSFS İş Akış Diyagramı

Bu yöntemin oluşturulmasındaki ana fikir; yarı-gözetimli olarak tetiklenen öğrenme sürecinden elde edilen yeni sınıflama sonuçlarının sisteme dahil edilerek, hem eğitim için en sağlıklı verilerin kullanılmasını sağlamak hem de sistemdeki sağlıklı veriler neticesinde başarımları daha yüksek noktalara kolayca taşıyabilmektir. Tez çalışmasının duygu analizi kısımları için kullanılan bu yöntem, Şekil 4.1’de iş akış diyagramıyla ifade edilmiştir. İş akışında belirtildiği üzere, ilk başta bir kısmı etiketlenen veriye öznitelik seçme metrikleri uygulanır. Sonrasında bu veri ile oluşturulan semantik matris sayesinde bilinmeyen verilerin sınıflama işlemi gerçekleştirilir. Daha sonrasında ilk baştaki etiketlenmiş veriyle, makine öğrenimiyle sınıflanmış bu yeni veriler bir araya getirilir. Tekrar öznitelik seçme metrikleriyle yeni bir semantik matris oluşturulur ve bilinmeyen başka bir veri daha bu yolla sınıflandırılır. Bu işlemler başarımlar artana değin tekrar ve tekrar gerçekleştirilir. Aşamalar arasında başarımın düştüğünü veya arttığını tespit edebilmek açısından yapılması gereken çok önemli bir nokta vardır. En başta etiketlenmiş olan veriler içerisinde hiçbir eğitim kümesine veya test kümesine katılmayacak, sadece bir defa seçilerek oluşturulmuş sabit bir veri olmalıdır. Böylelikle sistemin başarısını, her gerçekleştirilen yineleme sonucunda yeni semantik matrise göre hesaplayarak, bir önceki yöntemin başarımına göre oluşmuş farklılıklar tespit

edilebilmektedir. Deneyler sonucunda Bitcoin fiyatlamaları için en belirleyici semantik vektör boyutu veya diğer adıyla öznitelik sayısı 375 adet olarak belirlenmiştir (EK-1). Bu yöntemin ileri faz deney çalışmalarında başarımı olumsuz yönde etkileyen bazı sorunlar tespit edilmiştir. Sonrasında devam edilen deney çalışmalarıyla bu sorunların neden kaynaklandığı ve nasıl çözüleceği ile ilgili çözüm önerileri detaylarıyla paylaşılmıştır.

4.1.3.2 Manipülasyon tespiti

Manipulatörler ve Bitcoin için manipülasyon periyotlarının nokta atışı belirtecek bir şekilde verinin var olmayışı nedeniyle benzer çalışmalardaki yöntemler incelenmiştir. Fakat bu noktada literatürde bahsedilen çalışmalarda bahsettiğimiz eksiklik olan güvenilirlik problemi ortaya çıkmaktaydı. Çünkü gerçekleştirilen çalışmalarda (Wei 2020) kullanılan veriler rastgele bir zaman aralığı seçilerek oluşturulmuş ve bunların eğitim ve test kümesi periyotları rastgele seçilerek başarı sonuçları literatürde paylaşılmıştı. Bu aşamada deney sonuçlarının güvenilirliğini sağlayabilmek için katkısı olabilecek çalışmalar incelenmiştir. Bu çalışmalardan en önemli ipucunu veren (Adeyeye 2018), çalışmasında kriz dönemi öncesinde analizi yapılan verilerin yüksek oranda başarımlara sahip olduğunu belirtir. Bunun nedeni olarak da piyasaların kriz öncesi dönemler öncesinde yön tayini yapabilmek için piyasaların durağanlaştığını ve büyük yatırımcıların müdahalelerinin daha az olduğunu ifade etmiştir. Manipulatörlerin piyasadaki en yüksek likidite sağlayan yatırımcılar olduğunu ifade eden (Aggarwal 2003)'ün çalışmasında bahsettiği kişilerin de büyük yatırımcılar olduğu irdelenince, aslında bu durgunluğun yapılan manipülasyonların azalmasından kaynaklı olabileceği ihtimali ortaya çıkmıştır. Bu nedenle, bu iddiaların araştırılması için uygun bir verinin arayışı ihtiyacı doğmuştur.



Şekil 4.2 Köpüğün Aşamaları (Rodrigue, 2020)

Bu esnada tüm dünyada ortaya çıkan Covid-19 pandemisi bu iddiaların araştırılması için en uygun seçilecek periyot olarak belirlenmiştir. Böylelikle buradan elde edilecek sonuçlarla beraber bu iddialar ileride yapılacak deney sonuçlarıyla ispatlanacak ve (Rodrigue 2020) tarafından paylaşılan Şekil 4.2’de yer alan manipülasyon dönemlerini gösteren periyotlar üzerinde bu sonuçların ne doğrultuda sonuçlar vereceği gözlemlenebilmiştir. Köpüğün Aşamaları adı verilen grafik üzerinde verilen alanların manipülasyon bölgesi içinde yer alması sebebiyle; başarımlarının hangi periyotlarda ve ne düzeyde düştükleri gözlemlenmiş ve yaşanan düşüşlerin anomali tespiti yöntemi kullanılarak, manipülasyondan kaynaklı olabileceği ortaya çıkarılmıştır.

4.1.3.3 Manipulatör tespiti

Yukarıda anlatılan kısımlarda her bir farklı aşamada elde edilen sonuçlar ve (Aggarwal 2003)’in manipulatörleri piyasalara yüksek likidite sağlayan yatırımcılar olarak tanımlamış olduğu çalışmasından esinlenilerek, çalışmanın bir sonraki kısmı olan

manipulator tespiti yapılabilmektedir. Çalışmanın bir önceki manipülasyon kısmında anomali tespiti yapılarak yanlış tahmin edilmiş olan periyotlar, en başarılı olduğu tespit edilen yöntem ile işaretlenmiştir. En başarılı yöntemin tespiti içinse uygulanan yöntemlerin kriz öncesi ve kriz sonrasındaki başarımlarının bir ortalanması alınmıştır.

$i_1 \rightarrow \text{Total Period Number}$

$$\begin{aligned} & \sigma_{BTC_ADDRESS, F_{sum}(BTC_AMOUNT), F_{sum}(BTC_USD_VOLUME), time > period(1)_begin_date \text{ and } time < period(1)_end_date}(Transactions_Table) \cup \\ & \sigma_{BTC_ADDRESS, F_{sum}(BTC_AMOUNT), F_{sum}(BTC_USD_VOLUME), time > period(2)_begin_date \text{ and } time < period(2)_end_date}(Transactions_Table) \cup \\ & \sigma_{BTC_ADDRESS, F_{sum}(BTC_AMOUNT), F_{sum}(BTC_USD_VOLUME), time > period(3)_begin_date \text{ and } time < period(3)_end_date}(Transactions_Table) \cup \\ & \vdots \\ & \sigma_{BTC_ADDRESS, F_{sum}(BTC_AMOUNT), F_{sum}(BTC_USD_VOLUME), time > period(i)_begin_date \text{ and } time < period(i)_end_date}(Transactions_Table) \end{aligned}$$

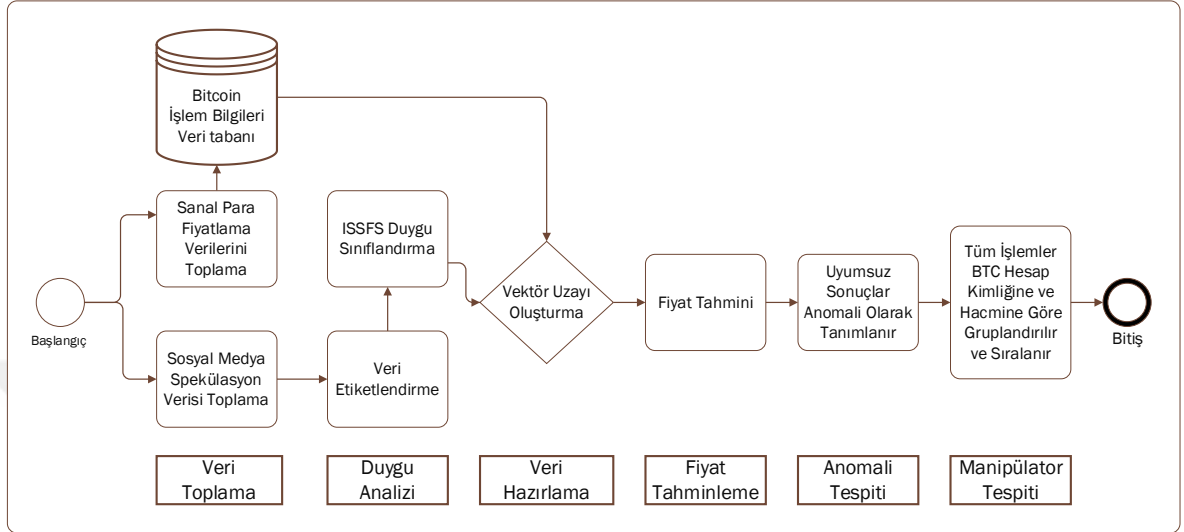
Şekil 4.3 Manipulator Tespiti için Kullanılan İlişkisel Cebir İfadesi

İki periyot için de bu ortalamanın üzerinde kalan yöntem en başarılı olarak belirlenmiştir. En başarılı yöntemle tespit edilen bu manipülasyon periyotlarına ait zamanlarda en yüksek hacimde alım ve satım işlemi gerçekleştirerek piyasalardaki likidite dengesini değiştiren bu hesaplar veri tabanına aktarılmış fiyatlama verileri üzerinden tespit edilmiştir. Bu hesapların tespiti için kullanılan sorgular Şekil 4.3'te ifade edilmektedir. Sonrasında en çok işlem yapan 500 hesap için özel bir analiz gerçekleştirilmiştir. Bu analiz sonuçları ileri kısımlarda paylaşılmıştır.

4.1.3.4 Önerilen sistem tasarımı

Tüm deney süreçleri tamamlandıktan sonra manipulator tespiti için bir sistem tasarımı tez sonucu olarak süreçlerle beraber Şekil 4.4'te gösterildiği şekilde genel hatlarıyla önerilmiştir. Fakat aşamalar içerisinde en başarılı sonuçları veren uygulamaların hangileri olduğu ileriki aşamalarda detaylarıyla paylaşılmaktadır. Genel hatlarıyla önerilen iş modeli, sırasıyla veri toplama, duygu analizi, veri hazırlama, fiyat tahminleme, anomali tespiti ve son olarak da manipulator tespiti adımlarından oluşmaktadır. Öncelikle manipulator tespiti için incelenmek istenen tarih aralığı

belirlenir. Sonrasında sosyal medyada o tarihe ait Bitcoin fiyat tahmini ile alakalı konu başlıklarıyla yapılan arama sonuçları bir veri tabanına kaydedilir. Sonrasında bu verilerin bir kısmı en az üç kişi tarafından etiketlenir.



Şekil 4.4 Önerilen Sistem Tasarımı

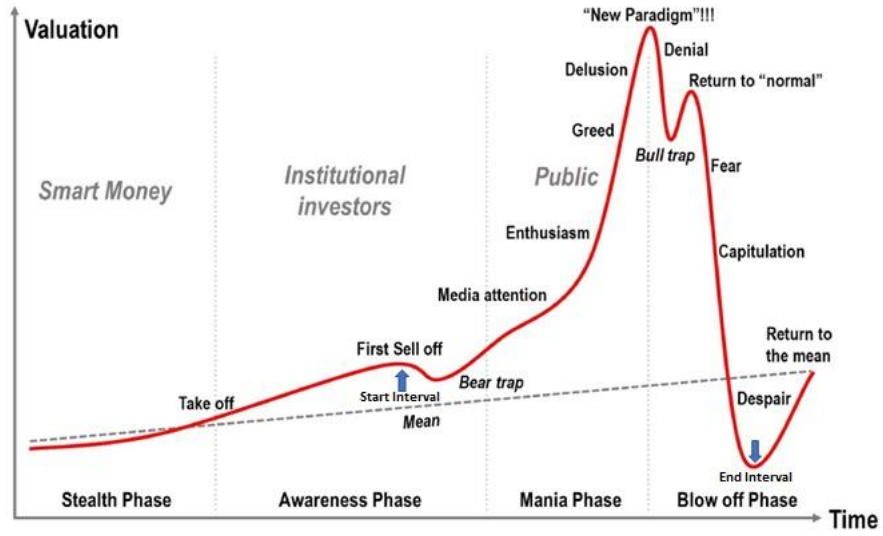
Tüm kategorilerde en çok oyu olan kategori o yorum için etiketlenir. Aynı tarih aralıkları için günlük açılış, kapanış, en düşük ve en yüksek fiyat verileri de bu süreçte kayıt altına alınır. Sonrasında fiyat verileri haftalık olarak açılış, kapanış, en yüksek ve en düşük fiyatları içerecek şekilde bu vektöre eklenir. Yanına da ISSFS yöntemiyle duygu analizi yapılan hafta için en yüksek çıkan fiyat yönü yükseliş, alçalış ve nötr (belirsiz) olacak şekilde belirteci eklenir. Bu sonuçlara istinaden, sistemin yanıldığı haftalar gerçek verilerle karşılaştırılarak tespit edilir. Sistemin yanıldığı haftalar için tüm Bitcoin işlemlerinin yer aldığı veri tabanı üzerinden anomali tespit edilen haftalar için belirlenmiş tarihlere göre en yüksek hacimde işlem yapan hesaplar tespit edilir ve bu hesaplar potansiyel manipülatörler olarak tespit edilir.

4.2 Veri Hazırlama

Bu tez çalışmasında kullanılan verilerin tamamı sosyal medya ve piyasalarda yaşanmış gerçek verilerden oluşmaktadır. Çalışılan konuya özel hazırlanmış bir veri kümesi olmaması nedeniyle, deney sürecine hazır olabilmesi için veriler ön işleme tabi tutulmuşlardır. Bu ön işlemler; sosyal medyadaki spekülasyon verilerinin elde edilmesi, metin verilerinde kelime-kök ayrıştırma, günlük Bitcoin fiyat verilerinden haftalık ve aylık sonuçların üretilmesi, “.txt” uzantılı ve her ay için ayrı olarak alınan tüm Bitcoin işlem verilerinin veri tabanına aktarılması gibi süreçleri içermektedir.

4.2.1 Manipülasyon verisi

Bu çalışma süreçlerinde en kritik verilerden birisi manipülasyon verisi olmuştur. Çünkü deney aşamalarında manipülasyon noktalarının tespiti aşamasında kullanılmıştır. Bu tür veriler internet ortamında çok yaygın bir şekilde paylaşılmaktadır. Fakat doğrulama konusunda eksik olduklarından, çalışmada kullanılmaları uygun görülmemiştir. Bu nedenle de yapılmış akademik çalışmalar içerisinde Bitcoin için özel üretilmiş bir veri arayışına girilmiştir. Bu noktada karşılaşılan Şekil 4.2’de gösterilen Köpüğün Aşamaları grafiğiyle karşılaşılmıştır. Bu çalışmada köpük olarak tasvir edilen şey aslında aşırı fiyatlamalar ve köpüğün patlamasıyla hemen ardından oluşan ani düşüşlerdir. Bu şekilde Bitcoin üzerinde yaşanmış olan manipülasyonlar, aşamaları verilerek grafik üzerinde paylaşılmıştır. Bu noktada elde bulunan spekülasyon verileri de göz önünde bulundurularak grafik üzerinde tarih aralıklarına karşılık gelen noktalar belirlenmiş ve test aşamasında kullanılacak alan için Şekil 4.5’te yer aldığı gibi mavi oklar ile gösterilerek bir tarih aralığı belirlenmiştir.



Şekil 4.5 Köpüğün Aşamaları Grafiği ve Tarih Belirteçleri

“Start Interval” ve “End Interval” yani tarih aralığının başlangıcını ve bitişini ifade etmektedirler. Bu aralığın bitiş kısmının en başta daha da uzun tutulması düşünülmüştür. Fakat elde bulunan spekülasyon verilerin en son tarihli verisi en fazla bu noktaya kadar ihtiyacı karşılamaktaydı. Burada dikkat çeken diğer bir önemli unsur da bu tarih aralıklarının belirlenmesi aşamasıydı.

Bu aşamada Şekil 4.6’da mum ifadeler ile gösterilen gerçek Bitcoin fiyat hareketi göz önünde bulundurulmuştur. Gerçek şablon üzerinde başlangıç ve bitiş aralığı eşleştirilmiş ve bu tarihler 12 Temmuz 2017-10 Temmuz 2019 olarak belirlenmiştir. Fakat elimizdeki en son tarihli spekülasyon verisinin 03 Ocak 2019 tarihine ait olması sebebiyle bu aralıktaki son kısımda yer alan altı ay değerlendirmeye tabi tutulmamıştır. Böylelikle toplamda yaklaşık 1,5 seneye denk gelen bir alan manipülasyon bölgesi olarak işaretlenmiştir.



Şekil 4.6 Bitcoin Gerçek Fiyat Hareketi

4.2.2 Bitcoin fiyat verileri

Gerçek borsalardaki hisseler için geçmiş veriler saklandığı gibi, aynı durum Bitcoin için de geçerlidir. Sanal para borsalarının hepsinde bu geçmiş fiyat verileri, günlük olarak kaydedilmekte ve teknik analiz yapmak için kullanacak yatırımcılar için bu verileri paylaşmaktadırlar. Bu nedenle de bu verilere kolayca ulaşılabilir. Bu tez çalışmasında kullanılabilmesi için gereken Bitcoin günlük açılış, kapanış, en düşük ve en yüksek fiyatlama bilgileri, (CoinMarketCap 2021) adlı web sayfası üzerinden sağlanmıştır. Bu veriler daha sonra işlenip, bir önceki haftaya göre fiyat yükselişi veya alçalışına göre de ayrıca etiketlenmiştir. Böylelikle ileride yapılacak analizlerde doğrudan fiyat tahmini yapılması yerine, haftalık yükseliş veya düşüş tahmini yapılabilmesi sağlanmıştır.

4.2.3 Tüm bitcoin işlem verileri

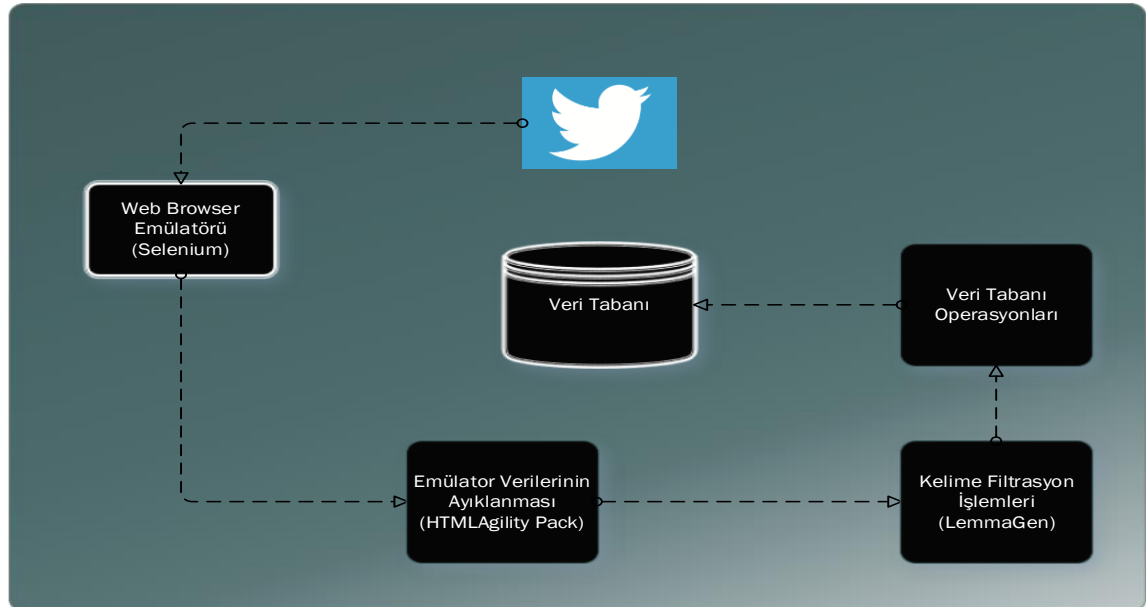
Tez çalışmasının son fazında manipulatörlerin tespit edilebilmesi için manipülasyon periyotlarına denk gelen tarih aralıkları için tüm geçmiş Bitcoin işlem verilerinin elde edilmesi gerekiyordu. Bitcoin işlem verilerinin çok fazla olmasından kaynaklı olarak bu verilerin bir yerden hazır olarak temin edilmesi sağlanmıştır. Bunun için de bilimsel

araştırma amacıyla yapılan çalışmalara ücretsiz olarak destek sağlayan (Blockchair 2021) ile iletişime geçilmiş ve oradan sağlanan API kullanıcı adı ve şifre sayesinde tez çalışması kapsamında ihtiyaç duyulan son beş seneye ait tüm işlem bilgileri elde edilmiştir. Bu aşamada elde edilen verilerin çok büyük olması sebebiyle aylık olarak indirilen bu veriler, daha sonrasında tez kapsamında oluşturulan bir aktarım yazılımı aracılığıyla MSSQL SERVER 2017 veri tabanı üzerinde kayıt altına alınmıştır. Yaklaşık olarak 350 GB büyüklüğünde olan bu veriler, işlem blok numarası, gönderici ve alıcı hesap adresleri, gönderim miktarları, gönderilen zamana ait Bitcoin USD cinsinden değeri gibi detaylı analiz yapmaya yetecek tüm bilgileri içermektedir.

4.2.4 Sosyal medyada yer alan spekülasyon verisi

Sosyal medya günümüzde internet kullanıcısı olan kesim tarafından yaygın olarak kullanılmaktadır. Sosyal medya insanların fikirlerini özgürce ifade edebildikleri bir ortam olarak bilinmektedir. Günde milyarlarca metin, resim, video vb. paylaşımlar bu platformlar aracılığıyla yapılmaktadır. Literatürde yer alan bazı çalışmalarda sosyal medya verileri üzerinden yapılan değerlendirmelerle fiyat tahminleri yapılmıştır. Ayrıca metin madenciliğinde sosyal medya platformlarında yapılmış olan çalışmaların sayıca çok fazla olması sebebiyle bu platformlardan en yaygın kullanılanları üzerinde odaklanılmıştır. (Pewinternet 2019) tarafından yayınlanmış ve Amerika’da yapılmış son anketine göre kullanıcıların %73’ü YouTube, %68’i Facebook, %35’i Instagram, %29’u Pinterest, %27’si Snapchat, %25’i LinkedIn, %24’ü Twitter kullanmayı tercih etmektedirler. Bu türde olan başka çalışmalar da incelenmiş ve en sık sosyal medya analizinde hedef alınan platformlardan biri olan (Twitter 2021)’in çalışmada kullanılacak sosyal medya platformu olması kararlaştırılmıştır. Bu noktada eğitim ve test verileri için yaklaşık 3,5 senelik bir yorum arşivinin elde edilmesine gerek duyulmuştur. İlk başta bu veriler platformun kendi ara yüzü ile elde edilmeye çalışılmıştır. Ama bu ara yüz ile o dönem çekilen bilgilerin 1,000 adetle sınırlandırılmış olması sebebiyle, daha fazla yorumun elde edilebilmesi için bir örümcek yazılım hazırlanmıştır. Bu örümcek yazılım sayesinde istenilen konu başlığına istinaden çekildiği günden en geçmiş veriye kadar yayında kalmış tüm paylaşımlar MSSQL

Server 2017 veri tabanına; metin verisi, tweet numarası, kullanıcı adı, beğeni sayısı, retweet (paylaşım) sayısı gibi veriler kaydedilmiştir. HTMLAgilityPack kütüphanesi verilerin HTML içerisinden alınmasını sağlamak için kullanılmıştır. LemmaGen kütüphanesi ise çok ayrıntılı bir dil kütüphanesi olup; İngilizcenin de olduğu toplam 12 adet dil için köklerine ayırıştırma fonksiyonu, kullanım kolaylığı ve açık kaynak kod desteği olması sebebiyle tez çalışmasında kullanılması tercih edilmiştir. Dinamik html sayfaları üzerinden bilgiye ulaşmak klasik bir html sayfasına göre çok farklıdır. Çünkü fare tekerleği ile aşağıya doğru gidildikçe sayfalar yüklenmektedir. Yapılan araştırmalarda bu tür yazılımların daha çok Python dilinde desteklerinin olması ve C# içinse popüler ve dokümantasyonu iyi başka bir alternatifin bulunamaması sebebiyle Selenium Webdriver kütüphanesinin bu yazılımda kullanılması tercih edilmiştir. Bu kütüphanenin işlevi web tarayıcı üzerinde fare emülasyonunun yapılmasını sağlamaktır. Şekil 4.7’de bu yazılımın nasıl çalıştığına dair iş akış şeması paylaşılmıştır. Tez çalışmasında kullanılan bu veriler İngilizce dilinde gerçekleştirildiği için 'Bitcoin', 'Bitcoin Price', 'Bitcoin Forecast', 'BTC', 'BTC / USD', 'BTCUSD' arama cümleleri sonucunda erişilen paylaşımlar tez çalışmasında kullanılmak üzere kayıt altına alınmıştır. Toplamda 86,077 adet yorum bu şekilde elde edilmiştir.



Şekil 4.7 Örümcek Yazılım İş Akış Diyagramı

Bu verilerden 28,808 adeti eğitim ve test kümelerinin bulunduğu tarih aralıklarının dışında kalmıştır. Fakat bu veriler yine de ISSFS yönteminin deney çalışmalarında verilerin etiketlenmesi ve sistemin iyileştirilmesi amacıyla kullanılmıştır. Deneyin fiyat tahmini aşamaları tarihlerine denk gelen 3,5 senelik kısım, toplamda 57,269 adet yorumun etiketine göre değerlendirilmiştir. Bu verilerin sadece 6,500 adeti deney çalışmalarında kullanılmak üzere toplamda üç kişi tarafından etiketlenmişti. Kalan etiketlenmemiş tüm yorumlar elle etiketlenen bu veriler içerisinde seçilen eğitim kümelerinin ISSFS yöntemiyle sonradan etiketlenmesiyle elde edilmiştir.

4.3 Tahminleme yöntemlerinin gerçek başarımlarının tespiti

Fiyat tahminleme başarımlarının kriz öncesi ve sonrası dönemlerde farklı sonuçlar verebileceği daha önceki kısımlarda anlatılmıştı. İlk Aşama testler için (Kaya 2020) ile beraber gerçekleştirilen çalışmada Covid-19 döneminde fiyat tahmini için literatürdeki en gelişmiş yöntemlerin başarımları birbirleriyle karşılaştırıldı. Bu tahminleri gerçekleştirebilmek için haftalık Bitcoin fiyat verileri kullanıldı. Bu verilerde pazartesi günü açılış ve pazar günü kapanış günü olarak belirtilmiştir.

05 Şubat 2018 ile 27 Ekim 2019 tarihleri arasında kalan 90 haftalık periyot, küresel pandemi haberlerinin yayılması öncesine denk gelmesi nedeniyle eğitim kümesi olarak kullanılmıştır. 27 Ekim 2019 ile 28 Haziran 2020 tarihleri arasında kalan tüm 30 haftalık dilim ise hem 200 günlük üssel ortalamaya (EMA) denk gelmesi hem de bu olayın pandeminin dönem olarak kullanılabilmesinden kaynaklı olarak test kümesinde bütün olarak değerlendirilmesi planlanmıştır. Tüm bu sonuçlar çizelgenin sol tarafında en iyi sonuç veren parametreleriyle birlikte paylaşılmıştır. Aynı deneyleri tam krizin öncesine denk gelen dönemler olan 15 Mart 2020 ve 28 Haziran 2020 dönemleri için daraltıldığında ise başarımların daha da yükseldiği dikkatimizi çekmiştir. Bu sonuçlar da en iyi sonuç veren parametreleriyle birlikte tablonun sağ tarafında paylaşılmıştır. ARIMA yöntemi için alınan sonuçlar %93 F_1 skor değerini göstermekteydi. Bu da aslında (Adedeye 2018)'in iddialarını onaylar nitelikteydi. ARIMA yöntemlerinin çok başarılı olması, çalışmaların sonraki adımlarında ARIMA'nın farklı bir kullanım

yöntemi olan SARIMAX’ın da sonuçlarının merak edilmesine sebep oldu. Aynı veriler kullanılarak SARIMAX yöntemi de bu süreçte değerlendirildi.

Çizelge 4.1 Kriz Öncesi Fiyat Tahmini Başarım Değerleri

Covid-19 Dönemi (Tüm Periyotlar)								Covid-19 Dönemi (Kriz Öncesi)						
Yöntem	Trend	F1 Skor (%)	Duyarlılık (%)	Anna (%)	Ort. F1 Skoru (%)	MCC (0-1)	Doğruluk (%)	Yöntem	F1 Skor (%)	Duyarlılık (%)	Anna (%)	Ort. F1 Skoru (%)	MCC (0-1)	Doğruluk (%)
LSTM (5,20,64)	Y	73	63	86	70	0,434	70	LSTM (10,75,256)	75	67	86	73	0,491	73
	D	67	82	56					71	83	62			
ARIMA (2,1,1)	Y	35	100	21	55	0,356	63	ARIMA (2,1,4)	92	100	86	93	0,873	93
	D	74	59	100					94	89	100			
SARIMAX (11,0,9)	Y	60	64	56	60	0,205	60	SARIMAX (6,1,2)	71	67	75	66,5	0,33	67
	D	60	56	64					62	67	57			
SVM Linear SVC	Y	78	100	64	82	0,7	83	SVM Linear SVC	83	100	71	86	0,756	87
	D	86	76	100					89	80	100			
SVM Sigmoid SVC	Y	75	90	64	79	0,614	80	SVM Sigmoid SVC	83	100	71	86	0,756	87
	D	83	75	94					89	80	100			
SVM Poly-SVC	Y	78	100	64	82	0,7	83	SVM Poly-SVC	83	100	71	86	0,756	87
	D	86	76	100					89	80	100			
SVM NuSVR	Y	78	64	100	73	0,56	73	SVM NuSVR	67	62	71	67	0,34	67
	D	67	100	50					67	71	62			
SVM Epsilon SVR	Y	78	100	64	82	0,7	83	SVM Epsilon SVR	82	70	100	80	0,66	80
	D	86	76	100					77	100	62			

Fakat alınan sonuçlar Çizelge 4.1’de gösterildiği üzere kardeşi ARIMA kadar başarılı değildi. Trend kısmında belirtilen ‘D’ ve ‘Y’ harfleri, sırasıyla düşüş ve yükselişi belirtmektedir. Parametreler bu tez çalışması kapsamında (Kaya 2020) ile birlikte paylaşılan makalede yer almaktadır. Alınan bu başarılı sonuçların aslında kriz öncesi dönemlerde manipülasyonların azalmasından kaynaklanmış olabileceğinden şüphelenilmiştir. Sonrasında yapılacak deney çalışmalarında, bu başarılı değerlerin manipülasyon dönemleri için de geçerli olup olmayacağının tespit edilmesi kararlaştırılmıştır.

4.4 Sosyal Medya Spekölasyonlarının Duygu Analizi Teknikleriyle Değerlendirilmesi

Sosyal medyada Bitcoin fiyatları ile ilgili yapılan paylaşımların anlamlandırılabilmesi için duygu analizi yöntemleriyle bir analiz gerçekleştirilmiştir. Bu analiz gerçekleştirilirken elde edilen verilerde birtakım problemler tespit edilmiştir. Bu problemlerin çoğu ön işleme süreçlerinde giderilebilmiştir.

4.4.1 Verilerin etiketlenilmesi

Veriler ön işleme süreçlerinden sonra tekrar incelendiğinde; bazı yorumlar fiyatların tek başına artış veya azalışını belirtmesinin haricinde, uzun ve kısa dönemde yaşanan fiyat farklılıklarını da barındırmaktaydı. Bu durumda oluşan senaryoda yorumlar hem fiyat artışını hem de düşüşünü içerebiliyordu. Bu nedenle de yapılacak kategori sınıflamalarının başarıları büyük oranda düşecekti. Bu noktada verilerin etiketlenmesi aşamasında fiyatlama belirteci olarak kullanılan iki tür kategoriyi (Yükseliş-Alçalış) de aynı anda içeren yorumlar Belirsiz (Nötr) olarak etiketlenmiştir. Toplamda üç farklı kişi tarafından gerçekleştirilen bu etiketleme işlemlerinde toplamda 6,500 adet yorum çoğunluk etiketlenmiş kategoriye göre ana kategorisi belirlenmiştir. 6,500 adet verinin, tezin ileriki aşamalarında kullanılacak analiz değerlendirmeleri için yeterli sayıda olmayacağı da düşünülmüştür. Bunun için de gözetimli öğrenme teknikleri yardımıyla etiketlenmemiş geri kalan verinin etiketlenmesi düşünülmüştür. İlk etapta, etiketlenen bu verilerden toplamda 3000 adedi, her biri için 1000'er adet olacak şekilde artış, azalış ve belirsiz kategorilerinin eğitimi için ayrılmıştır. Kalan etiketli verilerden de eşit olacak şekilde dört adet farklı test kümesi ileride kullanılacak çalışmalar için hazırlanmıştır. Bunların dağılımları da 358'er adet fiyat artışı, 208 adet fiyat düşüşü, 89'er adet belirsiz kategori olacak şekilde bölüştürülmüştür.

4.4.2 ISSFS yönteminin geliştirilmesi ve verilerin çoğaltılması

Verilerin karmaşıklık içermesi sebebiyle sınıflama işleminin düzgün gerçekleştirilebilmesi için; öznitelik seçme metrikleri etkili bir şekilde kullanılarak, başarımların düşüklükleri mümkün olacak en iyi şekilde engellenmeye çalışılmıştır. Bu noktada maksimum öznitelik sayısı, veri tabanı sütun limitinin 4094 olması sebebiyle rastgele alınan kelimelere göre gerçekleştirilmiştir. Ayrıca deneylerin ilk aşamalarında literatürde metin verileriyle ilgili en başarılı çalışmalardan biri olan (Akba 2014) öznitelik seçme metriklerinin makine öğrenimi sınıflayıcılarıyla harmanlanarak kullanıldığı yöntemler de kullanılarak bir başarımlar karşılaştırılması yapılmıştır. En iyi başarı-temsili oranı için çalışmalar da bu deneyin sonraki süreçlerinde gerçekleştirilmiştir. Yapılmış olan bu çalışmada (Akba 2014), sınıflayıcılar içerisinde en iyi başarımların SVM yöntemiyle alındığı sonuçlar kısmında paylaşılmıştır. Bu noktada Rastgele Orman yönteminin iç içe geçmiş veriler üzerinde daha iyi sınıflama yeteneğinin olabileceği değerlendirilerek deney çalışmalarda kullanılması kararlaştırılmıştır.

Çizelge 4.2 Öznitelik Seçme Uygulanmadan Elde Edilen Sınıflama Sonuçları

Öznitelik Seçme Metrikleri Yok (Rastgele 4094 Öznitelik)				
Kategori	SVM	CNN-5 Katman	DNDF	Rastgele Orman
	F ₁ Skor Ölçümü (0-1)			
Fiyat Artışı	0,59825	0,5435	0,534	0,63125
Fiyat Azalışı	0,5455	0,482	0,581	0,592
Belirsiz/Nötr	0,292	0,32175	0,296	0,342
Toplam (Ort.)	0,54725	0,494	0,470	0,57925

Bunun yanında başarılı olabileceği benzer çalışmalardan anlaşılmış ve CNN-5 katman yöntemi ile DNDF yöntemi de bu çalışmalara dahil edilmiştir. Gerçekleştirilen ilk çalışmada kullanılacak etiketli verilerde toplamda 12700 farklı kelime bulunduğu tespit edilmiştir. Bunlardan 4094 adeti rastgele seçilerek bir kelime vektör uzayı oluşturulmuş, test için ayrılan ilk kümeyle bir sınıflandırma deneyi gerçekleştirilmiştir. Tüm deney sonuçları Çizelge 4.2’de paylaşılmıştır. İlk gerçekleştirilen deney sonuçlarında

başarıların çok düşük olduğu fark edilmiştir. Sonrasında yapılacak deney sonuçlarıyla bu başarımların öznelik seçme yöntemlerinin beraber kullanılmasıyla artması beklenmiştir. Sonrasında eğitim kümesine bir öznelik seçme işlemi uygulanmış ve bir önceki deney aşamasında kullanılmış aynı test kümesiyle işlemler tekrarlanmıştır.

Çizelge 4.3 Öznelik Seçme Yöntemi Kullanılarak Elde Edilen Sınıflama Sonuçları

Öznelik Seçme (Kikare-375 Öznelik)				
Kategori	SVM	CNN-5 Katman	DNDF	Rastgele Orman
	F ₁ Skor Ölçümü (0-1)			
Fiyat Artışı	0,686	0,664	0,5784	0,6755
Fiyat Azalışı	0,596	0,569	0,69	0,6415
Belirsiz/Nötr	0,425	0,391	0,4171	0,43675
Toplam (Ort.)	0,622	0,596	0,5618	0,63225

En başarılı 375 adet özneliğe göre gerçekleştirilmiş ve sonuçları Çizelge 4.3'te paylaşılmıştır. 375 adet öznelik sayısının neden seçildiğine dair yapılan çalışmalar ileriki aşamalarda detaylarıyla anlatılmıştır. Alınan deney sonuçlarında gözlemlendiği gibi tüm başarımlar yüzde oranlarında 5 ile 10 arasında bir iyileşme görülmüştür. Burada alınan sonuçlardaki iyileşmeler ne kadar iyi görünmüş olsa da bunu daha da iyileştirebilmek adına çalışmalar devam ettirilmiştir. Bu aşamada kullanılan gözetimli öğrenme işlemi yinelemeli bir yapıya dönüştürerek, öğrenme işleminin sürekli bir hale getirilmesi sağlanmıştır. Yapıyı bu şekilde dönüştürerek, sadece kategori belirlenmesi için en önemli olan öznelikler üzerinden öğrenme yapılması sağlanmıştır. Bununla birlikte her bir yeni yinelemede öğrenilen öznelikler; bir önceki yinelemelerde kazanılmış bilgiyi güncelleyerek hem sistemin daha isabetli tahminler yapmasına olanak sağlayacak, hem de gereksiz özneliklerin süreçte değerlendirilmemesi sebebiyle sınıflama işlemlerinde çok büyük avantaj sağlayacaktır. Bu düşünceyle birlikte ortaya ISSFS yöntemi çıkmıştır. Önceki verilerin analizinde kullanılan eğitim ve test kümeleri bu yöntem üzerinde uygulanmış ve ilk sonuçları önceki sonuçlarla beraber karşılaştırılmalı olarak Çizelge 4.4'te paylaşılmıştır. Paylaşılan tablonu ilk satırında karşılaştırma amaçlı olarak Çizelge 4.3'te yer alan sonuçlar gösterilmektedir.

Çizelge 4.4 ISSFS Yöntemi Aşama Karşılaştırmalı Sınıflama Sonuçları

Yinelemeli Yarı Gözetimli Öznitelik Seçme (ISSFS)					
(Kikare-375 Öznitelik)					
F ₁ Skor Ölçümü (0-1)		SVM	CNN-5 Katman	DNDF	Rastgele Orman
Gözetimli Öznitelik Seçme Yöntemi	Fiyat Artışı	0,69	0,664	0,5784	0,68
	Fiyat Azalışı	0,60	0,569	0,69	0,64
	Belirsiz/Nötr	0,43	0,391	0,4171	0,44
	Toplam (Ort.)	0,62	0,596	0,5618	0,63
(1. Yineleme) Yarı-Gözetimli Aşama 1	Fiyat Artışı	0,70	0,696	0,6531	0,71
	Fiyat Azalışı	0,61	0,574	0,6636	0,65
	Belirsiz/Nötr	0,41	0,43	0,3778	0,41
	Toplam (Ort.)	0,63	0,613	0,5648	0,65
(2. Yineleme) Yarı-Gözetimli Aşama 2	Fiyat Artışı	0,72	0,711	0,6228	0,75
	Fiyat Azalışı	0,64	0,618	0,70425	0,67
	Belirsiz/Nötr	0,42	0,399	0,40445	0,43
	Toplam (Ort.)	0,66	0,639	0,5771	0,67
(3. Yineleme) Yarı-Gözetimli Aşama 3	Fiyat Artışı	0,70	0,694	0,5958	0,76
	Fiyat Azalışı	0,64	0,581	0,7428	0,68
	Belirsiz/Nötr	0,45	0,392	0,4037	0,43
	Toplam (Ort.)	0,64	0,617	0,5807	0,69
(4. Yineleme) Yarı-Gözetimli Aşama 3 + 5000 Etiketlenmemiş Veri	Fiyat Artışı	0,72	0,694	0,5958	0,75
	Fiyat Azalışı	0,63	0,635	0,7628	0,66
	Belirsiz/Nötr	0,44	0,44	0,4137	0,47
	Toplam (Ort.)	0,65	0,641	0,5907	0,68

Sonrasında deney aşamaları esnasında hazırlanmış dört adet test kümesi bu sistemin testleri aşamasında kullanılmıştır. Tüm bu sistemlerin gerçek başarımlarını ölçebilmek amacıyla ILC yönteminde de bahsedilmiş olan referans noktası olarak kullanılmak üzere

4. test kümesi hiçbir şekilde eğitim sürecine dahil edilmemiştir. Çizelge 4.4'te verilen sonuçların hepsi yine 4. test kümesi kullanılarak gerçekleştirilmiştir. 1. yinelemede, 1. test kümesi eğitim kümesinin içine dahil edilmiş ve bir öznitelik seçme işlemine tâbi tutulmuştur. Sonrasında kelime vektör uzayı yenilenip tekrardan sınıflama başarımı 4. test kümesi kullanılarak gerçekleştirilmiştir. Bu işlemler sırasıyla 2. ve 3. test kümeleri de eğitim kümesine eklenerek devam ettirilmiştir. Her aşama esnasında yeni öğrenilen öznitelikler ve değişen kıkare puanları nedeniyle kelime vektör uzayı değişmektedir. Bu şekilde sonuçların iyileştirilmesi amaçlanmıştır. En son olarak hiç bilinmeyen 5000 adet veri bu sisteme dahil edildiğinde sonuçların ne şekilde değişeceğine yönelik bir deney çalışması gerçekleştirilmiştir. 4. yineleme aşamasında alınan sonuçlar, etiketlenmeyen sonuçlar üzerinden başarımların arttırabileceğini göstermiştir.

Çizelge 4.5 DNDF Değişen Öznitelik Sayısına Göre Başarım Değişimi

Deney Numarası	Öznitelik Sayısı	Döngü Sayısı	Yığın Boyutu	Ağaç Sayısı	Ağaç Derinliği	F ₁ Skoru (0-1)
1	375	1000	1000	5	10	0.655
2	375	1000	1000	10	80	0.633
3	500	1000	1000	5	10	0.611
4	500	1000	1000	10	80	0.577
5	750	1000	1000	5	10	0.6
6	750	1000	1000	10	80	0.567
7	1000	1000	1000	5	10	0.623
8	1000	1000	1000	10	80	0.611
9	4094	1000	1000	5	10	0.556
10	4094	1000	1000	10	80	0.504

Bu noktada etiketlenmemiş diğer verilerin etiketlenilmesi için bu yöntemin güvenilir olarak kullanılabileceği ortaya çıkarılmıştır. Rastgele orman algoritmasının sınıflayıcı olarak en başarılı sonuçlar verdiği Çizelge 4.4'de görülmektedir. Ayrıca önceki gözetimli yönteme kıyasen yarı-gözetimli önerdiğimiz bu yöntemin başarımlarının %3 ile %5 aralığında iyileştiği de fark edilmiştir. Bu aşamalar esnasında, uyguladığımız yönteme benzerliği nedeniyle DNDF için bir deney haritası

oluşturulmuştur. DNDF mantık olarak öğrenmeye engel teşkil eden ağaç dallarını keserek çalışmasından ötürü, öznitelik sayıları arttırılarak deneyler tekrarlanmıştır.

Bu deneyin sonuçları da Çizelge 4.5'te gösterilmektedir. Bu noktada alınan sonuçlar incelendiğinde öznitelik sayısının arttırılmasının hem başarımların düşmesine hem de hesaplama sürelerinde yüksek artışa sebep olduğu gözlemlenmiştir. Ayrıca 375 adet öznitelik sayısından daha düşük sayılarda seçilmesi, değerlendirme için kullanılan kelime vektör matrislerinde boş gelen yorumların sayısını arttırmaktaydı. Bu da doğal olarak sistemin başarımını doğrudan düşürecek bir faktör olmasından dolayı 375 rakamının altında bir öznitelik sayısı seçilmemesine neden olmuştur.

Çizelge 4.6 DNDF Öznitelik Sayısına Göre Başarım Değişimi (Düşük Döngü Sayısı)

Deney Numarası	Öznitelik Sayısı	Döngü Sayısı	Yığın Boyutu	Ağaç Sayısı	Ağaç Derinliği	F ₁ Skoru (0-1)
1	375	55	1000	5	10	0,59
2	500	55	1000	5	10	0,44
3	750	55	1000	5	10	0,44
4	1000	55	1000	5	10	0,41
5	4094	55	1000	5	10	0,35

Rastgele orman algoritmasıyla beş dakikanın altında süren işlem süresi bu yöntemler uygulanırken; NVidia RTX 2080Ti ekran kartlı, Intel i7 7700k ve 32 GB Ram belleğe sahip bir bilgisayar üzerinde yaklaşık iki saatte tamamlanabilmektedir. DNDF yönteminin düşük döngü sayısına sahip en düşük ayarlarında bu işlemler gerçekleştirildiğinde alınan sonuçlar da Çizelge 4.6'da gösterilmektedir. Parametreler düşük döngülerde ayarlandığı için işlemler 15-20 dakika aralığında tamamlanmıştır. Fakat önceki sonuçlara kıyasen daha düşük başarım sonuçları gözlemlenmiştir.

Çizelge 4.7 Bütün Yöntemlerin Öznitelik Sayılarına Göre Başarım Değişimi

Öznitelik Sayısı	SVM	CNN 5-LAYER	DNDF	Random Forest	Yöntem Ortalamaları
375	0.613	0.585	0.56019	0.637	0.5987975
500	0.608	0.285	0.5741	0.644	0.527775
750	0.584	0.535	0.5479	0.62	0.571725
1000	0.588	0.584	0.5385	0.627	0.584375
4096	0.559	0.507	0.4761	0.558	0.525025

DNDF yöntemi için öznitelik sayılarına göre sistemin başarım skorları incelendiğinde 375 adet özneliğin vektör matrisi için kullanılmasının en doğru tercih olacağı Çizelge 4.7'de gösterilen en iyi ortalama başarım sonuçlarına istinaden kararlaştırılmıştır. Bu çizelgede gösterilen veriler; sıralı olarak 375, 500, 750, 1000 ve 4094 adet özneliğe göre hazırlanmış vektörlere göre, tüm yöntemler için gerçekleştirilen duygu sınıflama başarımlarını ve ortalama sonuçlarını içermektedir. Eldeki metin verilerinin küçük bir kısmıyla elde edilen başarılı sonuçlar sonrasında tamamına uygulanması yönünde motivasyonu arttırmıştır. Bu nedenle, deney setleri bu deneylerin büyük çapta gerçekleştirilebilmeleri için yeniden hazırlanmıştır. Önceki deneylerde kullandığımız 4. test kümesi haricindeki kalan tüm etiketli veriler eğitim kümesi olacak şekilde kullanılmıştır. 5675 adet olan bu verilere kalan veriler eşit parçalara bölünerek sınıflatılmıştır. Fakat elde edilen sonuçlar her bir yineleme sonrasında sonuçlarda büyük dalgalanmalar oluştuğunu göstermekteydi. Bu noktada bu regresyona sebep olabilecek iki farklı senaryo tespit edildi. İlk senaryo; az bir veriyle kendisinden daha fazla bilinmeyen bir veriyi sınıflayıp, birleştirme aşamasında iyi sınıflama yapılmadığı için tüm deney aşamalarını kötüleştirmesiydi. İkinci senaryoda ise; her bir öğrenme sonrasında kikare puanına göre yapılan hesap neticesinde özniteliklerin belirleniyor oluşunun, diğer kategorilerde daha düşük puana sahip ama önem derecesi yüksek olan öznitelikleri geri plana atıyor olmasıydı. Böylelikle kategoriler için ayrıştırıcı terim sayıları dengesiz olduğundan doğru sınıflandırma gerçekleştirilemiyordu.

Çizelge 4.8 ISSFS Yöntemi Büyük Verilerdeki Sistem Dengesi Çalışmaları

1. Ayar	Yükseliş	Düşüş	Belirsiz/Nötr	Ort. F ₁ Skoru (0-1)
Yineleme 1	0,674	0,562	0,654	0,633
Yineleme 2	0,637	0,537	0,582	0,59
Yineleme 3	0,377	0,186	0,25	0,282
Yineleme 4	0,628	0,437	0,603	0,561
Yineleme 5	0,607	0,356	0,588	0,523
2. Ayar	Yükseliş	Düşüş	Belirsiz/Nötr	Ort. F ₁ Skoru (0-1)
Yineleme 1	0,714	0,638	0,608	0,661
Yineleme 2	0,678	0,587	0,613	0,631
Yineleme 3	0,667	0,51	0,603	0,6
Yineleme 4	0,67	0,489	0,628	0,602
Yineleme 5	0,639	0,452	0,569	0,561
3. Ayar	Yükseliş	Düşüş	Belirsiz/Nötr	Ort. F ₁ Skoru (0-1)
Yineleme 1	0,621	0,603	0,667	0,628
Yineleme 2	0,654	0,637	0,623	0,64
Yineleme 3	0,655	0,646	0,588	0,634
Yineleme 4	0,659	0,642	0,563	0,627
Yineleme 5	0,601	0,621	0,484	0,575

Bu sorunların ne kadar giderildiğini görebilmek amacıyla 2. ayar olarak gösterilen kısımlarda önce eğitim ve test kümelerinin sayıları eşit olacak şekilde tekrardan ayarlandı. Yani ilk eğitim verisi 5675 adet ise, karşılığında aynı sayıda bilinmeyen etiketli yorum verisi sınıflandırılacaktı. Böylelikle bir sonraki eğitim kümesi birleştiğinde 11350 adet olacaktır. Sonrasında bu şekilde sırasıyla 22700 ve 55400 olacak şekilde veriler sınıflandırılmıştır. Sadece sonda kalan etiketlenmemiş veri sayısı 29677 adet kalmasından ötürü son verilen sınıflanacak test kümesi sayısı bu şekilde kullanılmıştır. Son olarak 2. ayarda uygulanan iyileştirmeye ek olarak; 3. ayarda öznelitliklerin değiştiği kısımda bir değişikliğe gidilmiş ve 375 adet olan öznelitlik sayısı, her bir kategori için sabit 125'er adet gelecek şekilde denge altına alınmıştır. Böylelikle sistemin eğitim ve test verilerinin aşırı dengesizliğinden kaynaklanan düşük başarımlar

sorunu ortadan kaldırılmıştır. Deneylerin sonuçlarına istinaden, sistemin dengede olduğu durumu içeren sonuçlar arasından en yüksek başarıma sahip olan 2. yinelemeye ait öznitelik seçme modeline göre tüm etiketsiz veriler sınıflandırılmıştır. Bu çalışmaların bütünü için kullanılan sınıflandırma algoritması da önceki deneylerde ön plana çıkmış olan rastgele orman algoritmasıdır. Yukarıda anlatılan tüm bu deneylere ait sonuçlar Çizelge 4.8’de gösterilmektedir. Bu sonuçlar sonrasında elde bulunan tüm etiketlenmiş değerler, çalışmanın ileri aşamalarında kullanılmak üzere veri tabanında kayıt altına alınmıştır. Alınan deney sonuçlarına istinaden İngilizce dili için Bitcoin piyasaları için kullanılmış en önemli 375 adet öznitelik EK-1’de paylaşılmıştır. Tüm burada anlatılan ve tez kapsamında gerçekleştirilen çalışmalar (Akba 2020) ve (Akba 2021) yayınlarıyla paylaşılmıştır.

4.5 Manipülasyonların Fiyat Tahmin Modellemeleriyle Değerlendirilmesi

Bu aşamada gerçekleştirilen çalışmalar spekülasyonların tahmin sonuçlarına olan etkisinin görülmesi amacıyla iki farklı doğrultuda ilerlemiştir. İlk yaklaşımda sadece fiyat verileri tahmin yapmak için kullanılmıştır. Diğer yaklaşımda ise, önceki deney aşamalarında elde edilen, ISSFS yöntemiyle etiketlenilmiş olan sosyal medya verileri de vektör uzayına eklenerek tahmin sonuçlarına olan etkileri izlenmiştir. Bu etkiler manipülasyon dönemi olarak belirtilen alanlar için haftalık ve aylık olarak gerçekleştirilmiştir. Bu noktada kendi kategorisine ait en başarılı yöntemler fiyat tahmini yapmak için kullanılmıştır. Çizelge 4.9’da gösterildiği üzere duygu analizi (DA) dahil ve dahil olmadan ne kadar başarıyla bu tahminlerin gerçekleştirilebileceği gözlenmiştir. Deneyler Python, Anaconda, scikit-learn, Statsmodels, Tensorflow ve Keras arayüz ve kütüphaneleri kullanılarak gerçekleştirilmiştir.

Çizelge 4.9 Manipülasyon Dönemi için Gerçekleştirilen Fiyat Tahmini Sonuçları

Yöntemler		Haftalık Parametreler	Haftalık Ortalama MCC Skoru (0-1)	Haftalık Ortalama F1 Skoru (%)	Aylık Parametreler	Aylık Ortalama MCC Skoru (0-1)	Aylık Ortalama F1 Skoru (%)
ARIMA	DA Dahil	(4,1,5)	0,0956	54,5	(1,1,5)	0,1791	58,5
	DA Hariç	(2,1,7)	0,15891	58	(5,1,5)	0,2567	62,5
SARIMAX	DA Dahil	(9,1,9)	0,1546	57,5	(8,0,6)	0,3144	64,5
	DA Hariç	(4,0,1)	0,2529	62,5	(5,0,5)	0,2516	55
SVM	DA Dahil	SVM Epsilon SVR	0,2917	64,5	SVM Linear SVC	0,2423	60
	DA Hariç	SVM Epsilon SVR	0,2695	63,5	SVM NuSVR	0,3186	60
LSTM	DA Dahil	(128, 4, 150, 0.1)	0,1093	54	(512, 15, 10, 0.1)	0,2423	60
	DA Hariç	(512, 2, 20, 0.1)	0,0825	54	(64, 5, 20, 0.1)	0,2423	60

Çizelge 4.9'da gösterilen tüm deney sonuçlarına göre; daha önceki deneylerde finansal kriz dönemlerinde yüksek başarı elde eden bütün yöntemler, olası manipülasyon noktaları olarak verilen tarih aralıklarında yeniden uygulanınca beklendiği üzere başarısız olmuştur. Bu noktada elde edilen sonuçlarda, SARIMAX Covid-19'un hem erken dönemleri için hem de manipülasyon bölgesi için benzer başarımlar vermesinden kaynaklı deneylerimizde sonuca yönelik bir katkı sağlayamamıştır. En çarpıcı sonuçlardan biri ise ARIMA için gerçekleştirilen deneylerde haftalık alınan başarımların %93 F₁ skoru gibi yüksek değerden %54 gibi düşük değerlere düşmüş

olmasıdır. Ayrıca bu durum sadece ARIMA için değil, diğer uygulanan tüm sonuçlar içinde gözlemlenmiştir.



Şekil 4.8 Manipülasyon Bölgeleri için Tespit Edilmiş Anomali Periyotları

Bu noktada yüzde 40'a yakın bir başarımlı düşüşü gözlemlenmesi ortada bir anomali söz konusu olduğunu kanıtlamaktadır. Çalışmanın ilk ana fikrinde de belirtildiği üzere manipülasyonların gerçekte de bu anomalilere sebep olduğu anlaşılmıştır. Sonrasında gerçekleştirilen tüm deney başarımlı sonuçlarının kriz öncesi ve sonrası dönemler için ayrı ayrı ortalamaları alınmıştır. Böylelikle belirtilen bu ortalama başarımlı değerlerini, deneylerin gerçekleştirilmiş olduğu iki farklı dönem için de gerçekleştiren SVM yöntemi en başarılı tespit yöntemi olarak belirlenmiştir. Bu durumda Çizelge 4.1'de kriz öncesi dönem için hesaplanan başarımlı ortalaması olan %79,69 F_1 değeri ve kriz sonrası dönem için hesaplanan %58,56 F_1 değerini karşılayan tek yöntemin, haftalık duygu analizi ile birlikte değerlendirilen SVM yöntemi olduğu kararlaştırılmıştır. Sonrasında sistemin yanlış sınıflandırmış olduğu haftalar manipülasyon yapılmış dönem olarak işaretlenmiştir. Bu dönemler Şekil 4.8'de gösterildiği gibi sarı alanlarla işaretlenerek gösterilmektedir.

4.6 Manipölasyon Verisi Üzerinden Manipölätör Tespitinin Gerçekleştirilmesi

Çizelge 4.10 Manipölasyon Dönemi Anomalileri Tarih Aralıkları

Başlangıç Tarihi	Bitiş Tarihi	Gerçek Sonuç	Tahmin Sonucu	Durum
12/06/2017	18/06/2017	Düşüş	Yükseliş	Anomali
19/06/2017	25/06/2017	Düşüş	Yükseliş	Anomali
26/06/2017	02/07/2017	Yükseliş	Düşüş	Anomali
03/07/2017	09/07/2017	Düşüş	Yükseliş	Anomali
17/07/2017	23/07/2017	Yükseliş	Düşüş	Anomali
31/07/2017	06/08/2017	Yükseliş	Düşüş	Anomali
14/08/2017	20/08/2017	Düşüş	Yükseliş	Anomali
21/08/2017	27/08/2017	Yükseliş	Düşüş	Anomali
27/11/2017	03/12/2017	Yükseliş	Düşüş	Anomali
15/01/2018	21/01/2018	Düşüş	Yükseliş	Anomali
22/01/2018	28/01/2018	Yükseliş	Düşüş	Anomali
29/01/2018	04/02/2018	Düşüş	Yükseliş	Anomali
05/02/2018	11/02/2018	Yükseliş	Düşüş	Anomali
12/02/2018	18/02/2018	Yükseliş	Düşüş	Anomali
19/02/2018	25/02/2018	Düşüş	Yükseliş	Anomali
26/02/2018	04/03/2018	Yükseliş	Düşüş	Anomali
05/03/2018	11/03/2018	Düşüş	Yükseliş	Anomali
26/03/2018	01/04/2018	Düşüş	Yükseliş	Anomali
02/04/2018	08/04/2018	Düşüş	Yükseliş	Anomali
30/04/2018	06/05/2018	Yükseliş	Düşüş	Anomali
07/05/2018	13/05/2018	Düşüş	Yükseliş	Anomali
02/07/2018	08/07/2018	Yükseliş	Düşüş	Anomali
10/09/2018	16/09/2018	Yükseliş	Düşüş	Anomali
24/09/2018	30/09/2018	Yükseliş	Düşüş	Anomali
01/10/2018	07/10/2018	Yükseliş	Düşüş	Anomali
22/10/2018	28/10/2018	Düşüş	Yükseliş	Anomali
29/10/2018	04/11/2018	Yükseliş	Düşüş	Anomali
05/11/2018	11/11/2018	Düşüş	Yükseliş	Anomali
10/12/2018	16/12/2018	Yükseliş	Düşüş	Anomali
31/12/2018	06/01/2019	Yükseliş	Düşüş	Anomali

Manipölasyon yapılan haftalar, en yüksek başarıma sahip yöntemin yanıldığı periyotlar üzerinden tespit edilmiştir. Toplamda 30 haftalık dilime denk gelen, anomali tespiti yöntemiyle elde edilmiş bu haftalar Çizelge 4.10'da paylaşılmıştır. Sonrasında

(Aggarwal 2003)'ın çalışmasında belirttiği üzere manipölatörlerin piyasaları yönlendirmek için kullandıkları en büyük aracın likidite sağlamak olması nedeniyle, deneyler bu aşamadan sonra bu tarihler aralığında en yüksek hacimde işlem yapmış hesapların tespitine yönelik devam etmiştir. Böylelikle Çizelge 4.10'da belirtilen periyotlar içerisinde en çok işlem yapan 500 adet Bitcoin hesabı, önceden Şekil 4.3'te de ilişkisel cebir yöntemiyle gösterildiği şekilde tespit edilmiştir. Gerçekte kullanılan veri tabanı sorguları ise EK-2'de yer almaktadır. Elde edilen sonuçlar analiz edildiğinde manipölasyon olarak tanımlanan tarihler içerisinde en fazla işlem yapan 500 hesap için ortalama Bitcoin maliyetinin 4.721,388 dolar olduğu gözlemlendi. Manipölasyon sayılan dönemlerde en düşük satış fiyatının 1934 dolar, en yüksek satış bedelinin ise 11312 dolar olduğu belirlendi.

Bu 30 haftalık dönem için Bitcoin ve dolar bazında gerçekleştirilen işlemlerin hacmi 19.978.954 adet Bitcoin ve 81.018.924.212 \$ olarak belirlenmiştir. Bu da özünde hesap sahiplerinin manipüle ederek yarattıkları ortamdaki fiyat farklılıklarından ne kadar daha fazla kazanç elde ettiğini ortaya koymaktadır.

Bir sonraki kısımda tüm bu bölümde anlatılmış deneylerin özeti ve sonuçları tartışılıp değerlendirilmiştir. Daha sonrasında neler yapılabileceği paylaşılmıştır.

5. ÖZET, TARTIŞMA, DEĞERLENDİRME VE GELECEK ÇALIŞMALAR

Bu bölümde gerçekleştirilen tüm deneyler özetlenmiş, tartışılmış ve değerlendirilmiştir. Deney sonuçlarına istinaden, uygulanan çözümlerin en başta tanımlanmış problemlere ne kadar cevap verebildiği de burada açıklanmıştır. Son olarak bu çalışmanın devamında gerçekleştirilebilecek çalışmalar da anlatılmıştır.

5.1 Çalışma Özeti

Manipülasyon alanlarının tespiti ve manipülatörlerin doğrulanabilmeleri aşamalarında literatürdeki çalışmalar yeterli olmamaktaydı. Bu nedenle veri madenciliği yönteminde kullanılan en başarılı örnekler incelenmiş ve çözümün bu noktalarda anomali tespiti yapılmasının en iyi yöntem olacağı kararlaştırılmıştır. Ayrıca literatürde diğer bir soru olarak karşımıza çıkan spekülasyonların, manipülasyonlar üzerinde olan yöntemleri bu deney aşamaları esnasında irdelenmiştir. Deneylerin başlamasından önce kriz öncesi ve diğer dönemler için fiyat tahmin modellerinin başarılarının değiştiğine ait iddialar ardı sıra gerçekleştirilen deneylerle incelenmiştir. Sonrasında bu hipotezin doğruluğunun deneylerdeki yüksek başarımla sonuçlarıyla kanıtlanmasından sonra, manipülasyon bölgesinde aynı deneyler tekrarlanmıştır. Başarımla sonuçlarının hipotezimizde beklendiği üzere bu bölgeler için çok düşük olması (yaklaşık olarak %40 F_1 Skor) manipülasyonların bu başarımla düşüşleri ile doğrudan alakalı olduğu sonucunu ortaya koymuştur. Haftalık fiyat verileri içerisinde yanlış fiyat tahmini yapılan periyotlarda en yüksek işlem hacmini gerçekleştiren Bitcoin hesapları tespit edilmiştir. Tespiti yapılan bu hesaplar manipülatör hesaplar olarak tanımlanmıştır. Tüm bu aşamalarda gelecekte yapılacak veri madenciliği çalışmalarında çok faydası olacağı düşünülen ISSFS yönteminin temelleri bu tez çalışması sırasında ortaya çıkarılmış ve ilk kez gerçek senaryo üzerinde denenme fırsatı olmuştur. Sonuçlar, ISSFS yönteminin diğer yöntemlerle birlikte kullanıldığında başarımla sonuçlarını her yöntem için yükselttiğini göstermiştir. Tezin sonuç kısımlarında İngilizce dilinde Bitcoin borsaları için en önemli 375 kelimenin tayini yapılmış ve sonuçları EK-2’de paylaşılmıştır. Son olarak Bitcoin için gerçek bir manipülatör verisi paylaşılmamasına rağmen incelenen akademik

alıřmalardan elde edilen yol haritalarıyla maniplatrler tespit edilmiř ve etik olmaması nedeniyle sonular burada paylařılmamıřtır. Bunun yerine bu iřlemin nasıl yapılabileceęi deney ařamaları kısımlarında tez sonucu olarak paylařılmıřtır. Sonu olarak, maniplasyonları tespit edebilmek iin iki farklı adımın gerekleřtirilmesi gerektięi deneyler neticesinde ngrlmektedir. İlk adım olarak speklasyonlar, n iřleme srelerinden geirilen sosyal medya verilerinin bir kısmı eęitilecek řekilde Rastgele Orman ve ISSFS ynteminin harmanlanmasıyla etiketlenerek tespit edilmelidir. Sonrasında Bitcoin kapanıř fiyatları bir nceki hafta ile srekli karřılařtırılacak řekilde dřř ve ykseliř olarak etiketlenmelidir. Daha sonrasında speklasyon verilerinden gelen haftalık fiyat yn analizleri, gerek haftalık fiyat ynleri, Bitcoin aılıř ve kapanıř fiyatları ve haftalar bir uzay-zaman vektr olarak kullanılarak beraber SVM sınıflayıcı ile tahmin edilmek istenen periyotlar iin gerekleřtirilmelidir. Son olarak da gerek verilerle kıyas yapıldığında, sistemin yanıldıęı haftalar tespit edilmeli ve bu haftalar ierisinde en yksek hacimde iřlem yapan hesaplar EK-2’de rnek veri tabanı sorgusunda ifade edildięi řekilde tespit edilmelidir. Burada elde edilen sonuların resmi makamlarca halen doęrulanamaması sebebiyle potansiyel maniplatr listesi olarak ihtiya duyulduğunda kullanılabilecektir. Literatrde ilk defa bu kadar detaylı ve kapsamlı olarak gerekleřtirilen tez alıřmasında, tezin problem tanımı ařamasında yer alan tm sorulara bylelikle cevap verilmiřtir. Elde edilen dięer sonulara istinaden inceledięimiz speklatif yorumlar SVM ve SARIMAX yntemlerine olumlu etkiler yapmıřtır. LSTM’nin duygu analizi sonularıyla beraber deęerlendirilmesinde hibir olumlu veya olumsuz etkiye rastlanılmamıřtır. Speklatif veriler, ARIMA’nın hem haftalık hem de aylık verdięi tahmin bařarımlarını dřrmřtr. Verilen sonulara ek olarak maniplatrlerin, anomalilerin tespit edildięi tm dnemlerde eęilimlerin deęiřmesinde nemli bir etkiye sahip oldukları tespit edilmiřtir. Bu etkilerin genellikle fiyatları ykseltici, dřrc ve duraęanlařtırma ynnde olduęu gzlemlenmiřtir. Son olarak tez sonucunda literatre katkı olarak sunulan ISSFS ynteminin; zaman verimlilięi ve bařarı oranları aısından DNDF gibi benzer bakıř aısına sahip derin ęrenme alternatifleri ile karřılařtırıldığında, duygu sınıflandırmasında ok daha verimli olarak kullanılabileceęi kanıtlanmıřtır. Ayrıca maniplatrlerin fiyat maniplasyonları sayesinde byk karlar elde ettikleri de periyotlardaki fiyat hareketliliklerinden tespit edilebilmiřtir. Bu tez

alışmasının sonuçları toplamda 200 adetten fazla gerekleřtirilen deneyler neticesinde paylařılmıştır. Bitcoin piyasalarında manipölätör tespiti ile ilgili bilimsel ve gereki özölmler sunulması aısından řimdiye kadar gerekleřtirilmiş bu kapsamdaki bir arařtırmanın, sonrasında gerekleřtirilecek alışmalar iin de bir yol haritası olacaėı düřünölmektedir.

5.2 Tartıřma ve Deėerlendirme

Tezin hedef ve amaları doėrultusunda gerekleřtirilen tüm deneyler neticesinde birtakım sonuçlara varılmıştır. İlk olarak literatürde, krizin hemen öncesinde yapılan fiyat tahminlerinin daha başarılı olduėuna yönelik yer alan iddialar arařtırılmış ve elde edilen deney sonuçlarına istinaden bu farklılıkların aslında hipotezde de belirtildiėi üzere manipölasyonlardan kaynaklandığı ortaya ıkarılmıştır. Sonrasında makine yöntemlerinin güvenilirliğini ölçmek amalı, en az manipölasyon olduėu düřünölen bölgelerde birçok tahmin yöntemlerinin başarımları karşılařtırılarak bir deėerlendirme yapılmıştır. Bu deėerlendirmeler neticesinde tıpkı literatürdeki başka alışmalarda da belirtildiėi üzere krizin hemen öncesinde yapılan fiyat tahminlerinin gerekten yüksek başarılar elde ettikleri gözlemlenmiştir. Literatürde yapılan benzer alışmalar incelendiėinde, sonuçların doėrulaması ařamasında finansal para politikalarında söz sahibi olan küresel saygınlıktaki kamu kurum ve kuruluşlarının paylařmış olduėu veriler kullanılmaktadır. Fakat, tez alışmasında kullanılan Bitcoin'i halen gerek bir menkul kıymet olduėunu kabul eden ve bu alanda piyasalarda düzenleyici olmaları aısından bilimsel veriler paylařan resmi bir kurum veya kuruluş halen mevcut deėildir. Bu nedenledir ki deney sonuçlarında kullanılan yöntemlerin başarımlarının ok yüksek olması, alışmanın sonraki ařamalarında alınan sonuçlar iin büyük bir önem arz etmekteydi. Manipölasyon bölgelerindeki yüksek yanılma oranları da bu bölgelerdeki müdahalelerin varlığını kanıtlamış olmaktadır. Hipotezimize göre bu müdahalelerin kaynaėı da gerekleřtirilen manipölasyonlardır.

5.3 Gelecekte Yapılması Planlanan Çalışmalar

Bitcoin blok zincir teknolojisinin de getirmiş olduğu avantajı içinde barındırması nedeniyle gün geçtikçe insanlar arasında kullanımı yaygınlaşmaktadır. Fakat içinde her ne kadar kullanışlı bir teknolojiyi barındırıyor olsa da küresel para piyasası düzenleyiciler tarafından halen güvenilir bir yatırım aracı olarak listelenmemektedir. Bunun en temel nedeni de tasarımında sahip olmuş olduğu merkeziyetsiz yapısıdır. Merkeziyetsiz olması nedeniyle gerçekleştirilen işlemler için bir geri dönüş mekanizmasının neredeyse imkansızına yakın olması en büyük etkenlerden sadece bir tanesidir. Ayrıca Bitcoin hesabına sahip olmak için gerçek bir kişi, kurum veya tüzel bir kimlik gerekmediğinden dolayı, nereden geldiği belli olmayan paraların sirkülasyonu çok kolaylaşmaktadır. Bu nakit akış dengesini sağlayabilmek amacıyla, bugün çoğu ülkede kripto para ile doğrudan alışveriş işlemi yapılmasına ardı ardına yasaklar gelmektedir. Fakat yapılan bu kadar düzenlemelere rağmen henüz piyasa düzenleyiciler tarafından kabul görmemektedir. Bu nedenle de Bitcoin ile ilgili manipülatif hareketlere dair akademik çalışmalar çok kısıtlı gerçekleştirilebilmektedir. Bu aşamada ileride yapacağımız bu alandaki çalışmalar, bu alanda paylaşılacak yeni ve doğrulanabilir veriler ışığında gerçekleştirilecektir. Elde edilen manipülasyon verileri bu noktada doğrulanma imkanına sahip olacaktır.

KAYNAKLAR

- Adeyeye, P. O., Aluko O. A., Migiro S. O., 2018. The global financial crisis and stock price behaviour: time evidence from Nigeria, *Global Business and Economics Review*, 20(3), 373-387.
- Aggarwal, R. K., Wu G., 2003. Stock market manipulation-theory and evidence, *AFA 2004 San Diego Meetings*.
- Akba, F., et al., 2014. Assessment of feature selection metrics for sentiment analyses: Turkish movie reviews, *8th European Conference on Data Mining*, Vol. 191. No. 2002. 2014.
- Akba, F., Medeni, I. T., Guzel, M. S., & Askerzade, I., 2020. Assessment of Iterative Semi-Supervised Feature Selection Learning for Sentiment Analyses: Digital Currency Markets, *IEEE 14th International Conference on Semantic Computing (ICSC)* (pp. 459-463).
- Akca M. F, 2021. Web Sitesi: <https://medium.com/deep-learning-turkiye/nedir-bu-destek-vektör-makineleri-makine-öğrenmesi-serisi-2-94e576e4223e>, Nedir Bu Destek Vektör Makineleri? (Makine Öğrenmesi Serisi-2) (Son Erişim: 28 Ocak 2021).
- Albashrawi, M. 2016. Detecting financial fraud using data mining techniques: A decade review from 2004 to 2015, *Journal of Data Science*, 14(3), 553-569.
- Babayan, D. 2019. Trader: If Bitcoin Declines in February, it Would be First 7-Month Consecutive Decline, Web Sitesi: <https://www.newsbtc.com/analysis/trader-if-bitcoin-declines-in-february-it-would-be-first-7-month-consecutive-decline/>, (Son Erişim: 28 Ocak 2021).
- Bazan, J. G., Szczuka M., 2000. RSES and RSESlib-a collection of tools for rough set computations, *In International Conference on Rough Sets and Current Trends in Computing* (pp. 106-113). Springer, Berlin, Heidelberg.
- Blockchair, 2021. Web Sitesi: <https://blockchair.com/>, (Son Erişim: 28 Ocak 2021).
- Bondi, P., Casalino, G., Gambardella, L., 1988. On the iterative learning control theory for robotic manipulators, *IEEE Journal on Robotics and Automation*, 4(1), 14-22.
- Box, G. E., Jenkins G. M., Reinsel G. C., Ljung G. M., 2015. Time series analysis: forecasting and control, John Wiley & Sons.
- Breiman, L., 2001. Random forests. *Machine learning*, 45(1), 5-32.
- Buy Bitcoin Worldwide, 2019. How Many Bitcoin Users Are There?, Web Sitesi: <https://www.buybitcoinworldwide.com/how-many-bitcoin-users/>, (Son Erişim: 02 Şubat 2021).

- Cao, Y., Li Y., Coleman S., Belatreche A., McGinnity T. M., 2014. Adaptive hidden Markov model with anomaly states for price manipulation detection, *IEEE transactions on neural networks and learning systems*, 26(2), 318-330.
- Chahine, S., Malhotra N. K., 2018. Impact of social media strategies on stock price: the case of Twitter, *European Journal of Marketing*.
- Chen, W., Wu J., Zheng Z., Chen C., Zhou Y., 2019a. Market manipulation of bitcoin: Evidence from mining the Mt. Gox transaction network, *IEEE INFOCOM 2019-IEEE Conference on Computer Communications* (pp. 964-972).
- Chen, W., Xu Y., Zheng Z., Zhou Y., Yang J. E., Bian J., 2019b. Detecting "pump & dump schemes" on cryptocurrency market using an improved apriori algorithm, *2019 IEEE International Conference on Service-Oriented System Engineering (SOSE)* (pp. 293-2935).
- Chheda, S. R., Singh A. K., Singh P. S., Bhole A. S., 2018. Automated trading of cryptocurrency using twitter sentimental analysis, *Int. J. Comput. Sci. Eng*, 6, 209-214.
- CoinMarketCap, 2021. Web Sitesi: <https://coinmarketcap.com/>, (Son Erişim: 28 Ocak 2021).
- Desjardins J. 2020. All of the World's Money and Markets in One Visualization, Web Sitesi: <https://www.visualcapitalist.com/all-of-the-worlds-money-and-markets-in-one-visualization-2020/>, (Son Erişim: 28 Ocak 2021).
- Dunning, T., Friedman E., 2014. Practical machine learning: a new look at anomaly detection, O'Reilly Media, Inc..
- Fisher, I. E., Garnsey M. R., Hughes M. E., 2016. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research, *Intelligent Systems in Accounting, Finance and Management*, 23(3), 157-214.
- Fontanills, G. A., Gentile T., 2002. The stock market course (Vol. 117), John Wiley & Sons.
- Fu, R., Zhang Z., Li L., 2016. Using LSTM and GRU neural network methods for traffic flow prediction, *IEEE 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)* (pp. 324-328).
- Gerace, D., Chew C., Whittaker C., Mazzola P., 2014. Stock market manipulation on the Hong Kong stock Exchange, *Australasian accounting, Business and finance journal*, 8(4), 105-140.
- Golmohammadi, K., Zaiane O. R., 2017, Sentiment analysis on Twitter to improve time series contextual anomaly detection for detecting stock market manipulation, *International Conference on Big Data Analytics and Knowledge Discovery* (pp. 327-342), Springer, Cham.

- Golmohammadi, K., Zaiane, O. R., Díaz D, 2014. Detecting stock market manipulation using supervised learning algorithms, *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE.
- Grzymala-Busse, J. W., Stefanowski J., 2001. Three discretization methods for rule induction, *International Journal of Intelligent Systems*, 16(1), 29-38.
- Hochreiter, S., Schmidhuber J., 1997. Long short-term memory Neural computation 9.
- Karalevicius, V., Degrande N., De Weerd J., 2018. Using sentiment analysis to predict interday Bitcoin price movements, *The Journal of Risk Finance*.
- Kaya, U., Akba F., Medeni I. T., Medeni T. D., 2020. Covid-19 Öncesi ve Sonrasındaki Bitcoin Fiyat Değişimlerinin Makine Öğrenmesi, Zaman Serileri Analizi ve Derin Öğrenme Yöntemleriyle Değerlendirilmesi, *Bilişim Teknolojileri Dergisi*, 13 (3), 341-355. DOI: 10.17671/gazibtd.648424.
- Keidar, R., Blemus S., 2018. Cryptocurrencies and Market Abuse Risks: It's Time for Self-Regulation, *Lexology, Forthcoming*.
- Kim, Y. B., Kim J. G., Kim W., Im J. H., Kim T. H., Kang S. J., Kim C. H., 2016. Predicting fluctuations in cryptocurrency transactions based on user comments and replies, *PloS one*, 11(8), e0161197.
- Kontschieder, P., Fiterau M., Criminisi A., Buló S. R., 2015. *Deep neural decision forests*, *Proceedings of the IEEE international conference on computer vision* (pp. 1467-1475).
- La Morgia, M., Mei A., Sassi F., Stefa J., 2020. Pump and Dumps in the Bitcoin Era: Real Time Detection of Cryptocurrency Market Manipulations, *IEEE 2020 29th International Conference on Computer Communications and Networks (ICCCN)* (pp. 1-9).
- Lamon, C., Nielsen E., Redondo, E., 2017. Cryptocurrency price prediction using news and social media sentiment, *SMU Data Sci. Rev*, 1(3), 1-22.
- Laskowski, M., Kim H. M., 2016. Rapid prototyping of a text mining application for cryptocurrency market intelligence, *In 2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)* (pp. 448-453). IEEE.
- Lin, C. C., Chiu A. A., Huang S. Y., Yen D. C., 2015. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments, *Knowledge-Based Systems*, 89, 459-470.
- Liu, K., Yang X., Yu H., Mi J., Wang P., Chen, X., 2018. Rough set based semi-supervised feature selection via ensemble selector, *Knowledge-Based Systems*.
- Mirtaheri, M., Abu-El-Haija S., Morstatter, F., Ver Steeg G., Galstyan A., 2021. Identifying and analyzing cryptocurrency manipulations in social media, *IEEE Transactions on Computational Social Systems*, 8(3), 607-617.

- Ngai, E., Hu W., Wong Y. H., Chen Y., Sun X., 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, *Decision support systems*, 50(3), 559-569.
- Pewinternet, 2019. <http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/>, (Son Erişim: 21 Ocak 2019)
- Rodrigue, J. P., 2020. The geography of transport systems, Routledge.
- Siering, M., Clapham B., Engel O., Gomber P., 2017. A taxonomy of financial market manipulations: establishing trust and market integrity in the financialized economy through automated fraud detection, *Journal of Information Technology*, 32(3), 251-269.
- Sridhar, S., Mootha S., Subramanian S., 2020. Detection of Market Manipulation using Ensemble Neural Networks, *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)* (pp. 1-8).
- Srihari, S. N., Cohen E., Hull J. J., Kuan L., 1989. A system to locate and recognize ZIP codes in handwritten addresses, *IJRE*, 1, 37-45.
- Stenqvist, E., Lönnö J., 2017. Predicting Bitcoin price fluctuation with Twitter sentiment analysis, *KTH ROYAL Institute of Technology school of computer science and communication*.
- Twitter, 2021. Web Sitesi: <https://twitter.com/>, (Son Erişim: 28 Ocak 2021).
- Varol, O., Ferrara E., Menczer F., Flammini A., 2017. Early detection of promoted campaigns on social media, *EPJ Data Science*, 6, 1-19.
- Wei W., Zhang Q., Liu L., 2020. Bitcoin Transaction Forecasting with Deep Network Representation Learning, *IEEE Transactions on Emerging Topics in Computing*.
- Wilson, E. B., Hilferty M. M., 1931. The distribution of chi-square, *Proceedings of the National Academy of Sciences of the United States of America*, 17(12), 684.
- Zhang, D., Hu M., & Ji Q., 2020. Financial markets under the global pandemic of COVID-19, *Finance Research Letters*, 36, 101528.