



**SOSYAL MEDYANIN YAPAY ZEKA
ÖNGÖRÜ MODELLERİ ÜZERİNDEKİ ETKİSİ:
YÜKSEK FREKANSLI FİNANS VERİLERİ
ÜZERİNE BİR UYGULAMA**

Doktora Tezi

Hakan GÖKDAŞ

Eskişehir 2025

**SOSYAL MEDYANIN YAPAY ZEKA ÖNGÖRÜ MODELLERİ ÜZERİNDEKİ
ETKİSİ: YÜKSEK FREKANSLI FİNANS VERİLERİ ÜZERİNE BİR
UYGULAMA**

Hakan GÖKDAŞ

DOKTORA TEZİ

Endüstri Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Onur KAYA

(İkinci Danışman: Doç. Dr. Emre ÇİMEN)

Eskişehir

Eskişehir Teknik Üniversitesi

Lisansüstü Eğitim Enstitüsü

Şubat 2025

Bu tez çalışması BAP Komisyonu tarafından kabul edilen 21DRP107 no.lu proje kapsamında desteklenmiştir.

Bu tez çalışmasının yazarı YÖK 100/2000 kapsamında “Yapay Zeka ve Makine Öğrenmesi” alanında Doktora Bursiyeri olarak desteklenmiştir.

JÜRİ VE ENSTİTÜ ONAYI

Hakan GÖKDAŞ'ın "SOSYAL MEDYANIN YAPAY ZEKA ÖNGÖRÜ MODELLERİ ÜZERİNDEKİ ETKİSİ: YÜKSEK FREKANSLI FİNANS VERİLERİ ÜZERİNE BİR UYGULAMA" başlıklı çalışması 13/02/2025 tarihinde aşağıdaki jüri tarafından değerlendirilerek "Eskişehir Teknik Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca, Endüstri Mühendisliği Anabilim dalında Doktora Tezi olarak kabul edilmiştir.

<u>Jüri Üyeleri</u>	<u>Unvan Adı Soyadı</u>	<u>İmza</u>
Üye (Tez Danışmanı)	: Prof. Dr. Onur KAYA	
Üye	: Prof. Dr. Abdullah YALAMAN	
Üye	: Prof. Dr. Gürkan ÖZTÜRK	
Üye	: Doç. Dr. Eyüp ÇİNAR	
Üye	: Dr. Öğr. Üyesi Zeliha ERGÜL AYDIN	

Prof. Dr. Harun BÖCÜK
Lisansüstü Eğitim Enstitüsü Müdürü

13/02/2025

DANIřMAN ONAYI

Daniřmanlıęını yrttęm Doktora ęrencisi Hakan GKDAř, ‘‘Sosyal Medyanın Yapay Zeka ngr Modelleri zerindeki Etkisi: Yksek Frekanslı Finans Verileri zerine Bir Uygulama’’ bařlıklı tez alıřmasını tamamlamıřtır. Hazırlamıř olduęu tez tarafımca incelenmiř ve ęrencinin tez savunma sınavına alınması bilimsel ve etik aıdan uygun grlmřtır.

Tez Daniřmanı
Prof. Dr. Onur KAYA

ÖZET

SOSYAL MEDYANIN YAPAY ZEKA ÖNGÖRÜ MODELLERİ ÜZERİNDEKİ ETKİSİ: YÜKSEK FREKANSLI FİNANS VERİLERİ ÜZERİNE BİR UYGULAMA

Hakan GÖKDAŞ

Endüstri Mühendisliği Anabilim Dalı

Eskişehir Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Şubat 2025

Danışman: Prof. Dr. Onur KAYA

(İkinci Danışman: Doç. Dr. Emre ÇİMEN)

Piyasada gelecekte oluşacak hareketleri öngörebilmek birçok yatırımcı ve araştırmacının hedefidir. Finansal piyasaların davranışını etkileyen çok fazla etmen olması ve bu etmenlerin birbirleriyle olan karmaşık ilişkileri farklı veri kaynaklarından elde edilen birçok değişkenin birlikte modellenmesini gerektirmektedir. Finansal verilerle birlikte sosyal medya verilerini kullanan araştırmaların sayısı artmaktadır. Günümüzde sosyal medya en önemli iletişim araçlarından biridir ve birçok alana yön verir haldedir. Bu etkiden finansal piyasalarda etkilenmektedir.

Bu çalışma kapsamında NASDAQ borsasında yer alan beş adet teknoloji firması seçilmiştir. Sosyal medya verileri X aracılığıyla toplanmış ve hisse senedi verileriyle birlikte birçok öznitelik hesaplanarak analizler gerçekleştirilmiştir. X paylaşımını paylaşan kişinin etkisinin ve paylaşımın ulaştığı etki büyüklüğünün algoritmaya dahil edilmesi amacıyla gönderi istatistikleri kullanılarak yazar etki değeri ve post etki değeri öznitelikleri hesaplanmış ve bu öznitelikler gönderilerin duygu değerleriyle birlikte analize dahil edilmiştir. İkili ve üç sınıflı sınıflandırma analizlerinde karar ağacı, rastgele orman, destek vektör makinesi ve lojistik regresyon olmak üzere dört farklı makine öğrenmesi algoritması ve LSTM ağları kullanılmıştır. Algoritmaları kullanılarak borsanın hareket yönü tahmin edilmiş ve bu tahminler market simülasyonunda kullanılarak toplam karlılık değerleri hesaplanmıştır. Yapılan analizler sonucunda hem sosyal medya verilerinin etkisi gösterilmiş hem de en karlı parametre kombinasyonları belirlenmiştir.

Anahtar Sözcükler: Makine öğrenmesi, Derin öğrenme, Sınıflandırma, Hisse senedi hareket tahmini, Market simülasyonu, Sosyal medya duygu analizi.

ABSTRACT

THE IMPACT OF SOCIAL MEDIA ON ARTIFICIAL INTELLIGENCE FORECASTING MODELS: AN APPLICATION TO HIGH-FREQUENCY FINANCIAL DATA

Hakan GÖKDAŞ

Department of Industrial Engineering

Eskişehir Technical University, Institute of Graduate Programs, February 2025

Supervisor: Prof. Dr. Onur KAYA

(Co-Supervisor: Assoc. Prof. Dr. Emre ÇİMEN)

Predicting future market movements is a goal of many investors and researchers. The fact that there are many factors affecting the behavior of financial markets and the complex interrelationships between these factors requires the joint modeling of many variables obtained from different data sources. The number of studies using social media data together with financial data is increasing. Today, social media is one of the most important communication tools and is influencing many areas. Financial markets are no exception.

Five technology companies listed on the NASDAQ Stock Market were selected for this study. Social media data was collected through X and analyzed by calculating many attributes along with stock data. To include the influence of the person who shared the X post and the impact size reached by the post in the algorithm, author influence value and post influence value attributes were calculated using post statistics and these attributes were included in the analysis along with the sentiment values of the posts. Four different machine learning algorithms, namely decision tree, random forest, support vector machine and logistic regression, and LSTM networks were used in the binary and tri-class classification analyses. The algorithms are used to predict the direction of movement of the stock market and these predictions are used in the market simulation to calculate total profitability values. As a result of the analysis, both the effect of social media data was shown, and the most profitable parameter combinations were determined.

Keywords: Machine learning, Deep learning, Classification, Stock movement prediction, Market simulation, Social media sentiment analysis.

TEŞEKKÜR

Tez çalışmam sırasında kıymetli bilgi ve tecrübeleri ile bana yol gösterici olan ve desteğini esirgemeyen Sayın Danışman Hocalarım Prof. Dr. Onur KAYA ve Doç. Dr. Emre ÇİMEN'e, verdikleri değerli geribildirimler ile tezimi zenginleştiren Sayın Tez İzleme Komitesi Üyesi Hocalarım Prof. Dr. Abdullah YALAMAN ve Prof. Dr. Gürkan ÖZTÜRK'e, görüş ve önerileri ile tezime katkıda bulunan Sayın Jüri Üyesi Hocalarım Doç. Dr. Eyüp ÇİNAR ve Dr. Öğr. Üyesi Zeliha ERGÜL AYDIN'a sonsuz teşekkürlerimi sunarım.

Teşekkürlerimin en özelini ise, tüm eğitim hayatım boyunca benden desteklerini esirgemeyen ve her zaman yanımda olan sevgili anne ve babama sunuyorum.

Hakan GÖKDAŞ

ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ

Bu tezin bana ait, özgün bir çalışma olduğunu; çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ve kurallara uygun davrandığımı; bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi; bu çalışmanın Eskişehir Teknik Üniversitesi tarafından kullanılan “bilimsel intihal tespit programıyla tarandığını ve hiçbir şekilde “intihal içermediğini” beyan ederim. Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

Hakan GÖKDAŞ

İÇİNDEKİLER

	<u>Sayfa</u>
BAŞLIK SAYFASI	I
JÜRİ VE ENSTİTÜ ONAYI.....	II
DANIŞMAN ONAYI	III
ÖZET	IV
ABSTRACT.....	V
TEŞEKKÜR	VI
ETİK İLKE VE KURALLARA UYGUNLUK BEYANNAMESİ.....	VII
İÇİNDEKİLER	VIII
TABLOLAR DİZİNİ.....	X
ŞEKİLLER DİZİNİ	XIII
SİMGELER VE KISALTMALAR DİZİNİ	XIV
1. GİRİŞ	1
2. LİTERATÜR ARAŞTIRMASI	6
3. YÖNTEM	23
3.1. Finansal Zaman Serilerinin Analizi	23
3.1.1. Temel analiz.....	23
3.1.2. Teknik analiz.....	24
3.1.3. Matematiksel analiz	25
3.2. Yapay Zeka, Makine Öğrenmesi ve Derin Öğrenme	25
3.2.1. Karar ağaçları.....	27
3.2.2. Rastgele orman.....	28
3.2.3. Destek vektör makinesi	29
3.2.4. Lojistik regresyon	29
3.2.5. Uzun kısa vadeli hafıza.....	31
3.3. Duygu Analizi.....	32
4. VERİLER	34
4.1. Sosyal Medya Verileri	34
4.2. Sosyal Medya Verilerinden Öznitelik Oluşturma.....	36

4.2.1. Metin önişleme	36
4.2.2. Metnin vektörel olarak ifade edilmesi	39
4.2.3. Emoji analizi.....	40
4.2.4. Paylaşım yapan yazar sayısı.....	40
4.2.5. Yazar ve post etki değerleri	41
4.2.6. Metin ve emoji duygu durum değerleri	41
4.3. Hisse Senedi Verileri.....	42
4.4. Hisse Senedi Verilerinden Öznitelik Oluşturma	42
4.4.1. Keltner channel (KC) ve keltner channel percentage (KCP)	44
4.4.2. Ease of movement (EMV)	44
4.4.3. Commodity channel index (CCI).....	45
4.4.4. Stochastic relative strength index (Stoch RSI)	45
5. UYGULAMA	47
5.1. Market Simülasyonu ile Kazanç Analizi	61
5.2. İki Sınıflı Analiz için Sonuçlar	61
5.3. Üç Sınıflı Analiz için Sonuçlar	78
5.4. Analiz Sonuçlarının Karşılaştırılması.....	94
6. SONUÇ VE ÖNERİLER.....	98
KAYNAKÇA	101
ÖZGEÇMİŞ	

TABLÖLER DİZİNİ

	<u>Sayfa</u>
Tablo 2.1. Makine öğrenmesi yöntemleriyle yapılan çalışmalar	12
Tablo 2.2. Derin öğrenme yöntemleriyle yapılan çalışmalar	18
Tablo 2.3. Market simülasyonu gerçekleştirilen çalışmalar.....	22
Tablo 4.1. Şirketler için toplanan gönderi sayıları	35
Tablo 4.2. Duygu değer değişkenleri	42
Tablo 5.1. Beş şirketin üçüncü test dönemi başlangıç ve bitiş tarihleri	57
Tablo 5.2. Beş şirketin pozitif ve negatif dağılım yüzdeleri	57
Tablo 5.3. Beş şirketin pozitif, nötr ve negatif dağılım yüzdeleri.....	58
Tablo 5.4. Sınıflandırma değişken kombinasyonları	59
Tablo 5.5. İkili sınıflandırmada öznitelik seçim dönemi için en yüksek toplam kar değerleri	62
Tablo 5.6. İkili sınıflandırmada toplam kar-zarar değerinin T-testi değerleri	63
Tablo 5.7. İkili sınıflandırmada birinci test dönemi analiz sonuçları	64
Tablo 5.8. İkili sınıflandırmada birinci test dönemi yöntem karşılaştırması	64
Tablo 5.9. İkili sınıflandırmada ikinci test dönemi analiz sonuçları.....	66
Tablo 5.10. İkili sınıflandırmada ikinci test dönemi yöntem karşılaştırması.....	66
Tablo 5.11. İkili sınıflandırmada üçüncü test dönemi analiz sonuçları	67
Tablo 5.12. İkili sınıflandırmada üçüncü test dönemi yöntem karşılaştırması	68
Tablo 5.13. İkili sınıflandırmada öznitelik seçim dönemi için en yüksek doğruluk yüzdeleri.....	69
Tablo 5.14. İkili sınıflandırmada ortalama doğruluk yüzdesi değerinin T-testi değerleri	70
Tablo 5.15. Algoritma Hiper-parametre kombinasyonları.....	71
Tablo 5.16. LSTM hiper parametreleri	72
Tablo 5.17. İkili sınıflandırmada birinci test dönemi RF algoritması sonuçları	73
Tablo 5.18. İkili sınıflandırmada birinci test dönemi RF- ARIMA yöntem karşılaştırması	73
Tablo 5.19. İkili sınıflandırmada birinci test dönemi LSTM algoritması sonuçları	73
Tablo 5.20. İkili sınıflandırmada ikinci test dönemi RF algoritması sonuçları	75
Tablo 5.21. İkili sınıflandırmada ikinci test dönemi RF- ARIMA yöntem karşılaştırması	75

Tablo 5.22. İkili sınıflandırmada ikinci test dönemi LSTM algoritması sonuçları.....	75
Tablo 5.23. İkili sınıflandırmada üçüncü test dönemi RF algoritması sonuçları	76
Tablo 5.24. İkili sınıflandırmada üçüncü test dönemi RF- ARIMA yöntem karşılaştırması	77
Tablo 5.25. İkili sınıflandırmada üçüncü test dönemi LSTM algoritması sonuçları	77
Tablo 5.26. Üç sınıflı sınıflandırmada öznitelik seçim dönemi için en yüksek toplam kar değerleri	78
Tablo 5.27. Üç sınıflı sınıflandırmada toplam kar-zarar değerinin T-testi değerleri	80
Tablo 5.28. Üç sınıflı sınıflandırmada birinci test dönemi analiz sonuçları	81
Tablo 5.29. Üç sınıflı sınıflandırmada birinci test dönemi yöntem karşılaştırması	81
Tablo 5.30. Üç sınıflı sınıflandırmada ikinci test dönemi analiz sonuçları	83
Tablo 5.31. Üç sınıflı sınıflandırmada ikinci test dönemi yöntem karşılaştırması	83
Tablo 5.32. Üç sınıflı sınıflandırmada üçüncü test dönemi analiz sonuçları	84
Tablo 5.33. Üç sınıflı sınıflandırmada üçüncü test dönemi yöntem karşılaştırması	85
Tablo 5.34. Üç sınıflı sınıflandırmada öznitelik seçim dönemi için en yüksek doğruluk yüzdeleri.....	86
Tablo 5.35. Üç sınıflı sınıflandırmada ortalama doğruluk yüzdesi değerlerinin T-testi değerleri	87
Tablo 5.36. Üç sınıflı sınıflandırmada birinci test dönemi LR algoritması sonuçları....	89
Tablo 5.37. Üç sınıflı sınıflandırmada birinci test dönemi LR- ARIMA yöntem karşılaştırması	89
Tablo 5.38. Üç sınıflı sınıflandırmada birinci test dönemi LSTM algoritması sonuçları	90
Tablo 5.39. Üç sınıflı sınıflandırmada ikinci test dönemi LR algoritması sonuçları	91
Tablo 5.40. Üç sınıflı sınıflandırmada ikinci test dönemi LR- ARIMA yöntem karşılaştırması	91
Tablo 5.41. Üç sınıflı sınıflandırmada ikinci test dönemi LSTM algoritması sonuçları	92
Tablo 5.42. Üç sınıflı sınıflandırmada üçüncü test dönemi LR algoritması sonuçları...	93
Tablo 5.43. Üç sınıflı sınıflandırmada üçüncü test dönemi LR- ARIMA yöntem karşılaştırması	93
Tablo 5.44. Üç sınıflı sınıflandırmada üçüncü test dönemi LSTM algoritması sonuçları	93
Tablo 5.45. Toplam kar-zarar değeri için test dönemlerinin karşılaştırması	95

Tablo 5.46. ML algoritması ile ortalama doğruluk değerleri için test dönemlerinin karşılaştırması	96
Tablo 5.47. LSTM algoritması ile ortalama doğruluk değerleri için test dönemlerinin karşılaştırması	96



ŞEKİLLER DİZİNİ

	<u>Sayfa</u>
Şekil 3.1. LSTM hücre yapısı	32
Şekil 4.1. Metin önişleme akış şeması.....	37
Şekil 5.1. Algoritma akış şeması	47
Şekil 5.2. Karmaşıklık matrisi	49
Şekil 5.3. Algoritma verilerinin zaman çizelgesi	51
Şekil 5.4. Eğitim dönemi beş şirketin fiyat grafikleri a)Tesla b)Microsoft c)Apple d)Netflix e)Nvidia	52
Şekil 5.5. Öznitelik dönemi beş şirketin fiyat grafikleri a)Tesla b)Microsoft c)Apple d)Netflix e)Nvidia	53
Şekil 5.6. Birinci test dönemi beş şirketin fiyat grafikleri a)Tesla b)Microsoft c)Apple d)Netflix e)Nvidia	54
Şekil 5.7. İkinci test dönemi beş şirketin fiyat grafikleri a)Tesla b)Microsoft c)Apple d)Netflix e)Nvidia	55
Şekil 5.8. Üçüncü test dönemi beş şirketin fiyat grafikleri a)Tesla b)Microsoft c)Apple d)Netflix e)Nvidia	56
Şekil 5.9. Sosyal medya ile finansal verilerin ilişkilendirilmesi	61

SİMGELER VE KISALTMALAR DİZİNİ

AB	: AdaBoost
AID	: Automatic Interaction Detector
ANN	: Artificial Neural Network
ATR	: Average True Range
Bİ-GRU	: Bi-directional Gated Recurrent Unit
CART	: Classification And Regression Tree
CCI	: Commodity Channel Index
CNN	: Convolutional Neural Network
DGAP	: Deutsche Gesellschaft für Adhoc-Publizität
DJIA	: Dow Jones Industrial Average
DRSE	: Deep Random Subspace Ensembles
DT	: Decision Tree
ELM	: Extreme Learning Machine
EMA	: Exponential Moving Average
EMV	: Ease of Movement
ET	: Extra Tree
FLM	: Fast Large Margin
GBDT	: Gradient Boosting Decision Tree
GBM	: Gradient Boosting Classifier
GLM	: Generalized Linear Model
GRU	: Gated Recurrent Unit
HSCI	: Hang Seng Composite Index
IDF	: Inverse Document Frequency
INDU	: Dow Jones Industrial Average
KC	: Keltner Channel
KCP	: Keltner Channel Percentage
KNN	: K-Nearest Neighbor
LDA	: Linear Discriminant Analysis
LR	: Logistic Regression
LSTM	: Long-Short Term Memory

MACD	: Moving Average Convergence Divergence
MAE	: Mean Absolute Error
MFI	: Money Flow Index
MKL	: Multiple Kernel Learning
MLE	: Maximum likelihood Estimation
MLP	: Multi-Layer Perceptron
NYSE	: New York Stock Exchange
OBV	: On-balance Volume
OR	: Odds Ratio
RCNN	: Recurrent Convolutional Neural Network
RF	: Random Forest
RMSE	: Root Mean Square Error
RNN	: Recurrent Neural Network
ROC	: Rate of Change
RSI	: Relative Strength Index
SMA	: Simple Moving Average
Stoch RSI	: Stochastic relative strength index
SSWN	: Stock Senti Word Net
SVM	: Support Vector Machine
TF	: Term Frequency
TF-IDF	: Term Frequency- Inverse Document Frequency
TGRU	: Two-stream Gated Recurrent Unit
X API	: X Uygulama Programlama Arayüzü
XGBoost	: Extreme Gradient Boosting

1. GİRİŞ

Finansal teknolojinin hızla gelişmesi araştırmacıların, yatırımcıların, teknoloji ve finans şirketlerinin yapay zeka ve onun alt alanlarına yoğunlaşmasına neden olmuştur. Bu durum finans alanında oluşan büyük veriden faydalı bilgiler çıkarma ve bu bilgileri kullanma isteğinden kaynaklanmaktadır. Piyasada gelecekte oluşacak hareketleri öngörebilmek yatırımcılar ve piyasa yapıcılar için kritik bir avantaj sağlamaktadır. Geleneksel tahmin yöntemleri düşük frekanslı verilerden faydalanırken, yüksek frekanslı veriler piyasa dinamiklerini daha iyi anlamaya ve kısa vadeli hareketleri de öngörmeye olanak sağlamaktadır. Yüksek frekanslı veriler dakikalar içinde gerçekleşen değişimleri analiz etme imkanı sunarak, daha hassas modeller oluşturulmasına olanak sağlar. Bu yüksek frekanslı veriler, algoritmik işlem stratejileri geliştirme açısından fırsatlar sunmaktadır. Teknolojinin gelişmesi yüksek frekanslı verilerin işlenmesini kolaylaştırmıştır, buna bağlı olarak da yüksek frekanslı veri kullanımına artan bir ilgi bulunmaktadır. Bu çalışma kapsamında yüksek frekanslı verilerden yararlanılarak hisse senetlerinin fiyat değişim yönlerinin tahmin edilmesine yönelik bir yaklaşım ele alınmaktadır.

Yapay zeka, makine öğrenmesi ve derin öğrenme birbirleriyle ilişkili kavramlardır. Makine öğrenmesi yapay zekanın bir alt alanıyken, derin öğrenmede makine öğrenmesinin bir alt alanı olmaktadır. Teknolojinin gelişmesiyle birlikte makine öğrenmesi algoritmaları ve derin öğrenme algoritmalarının finans sektörü ve finansal tahminlerde kullanımına yoğun bir ilgi bulunmaktadır. Yüksek miktardaki finansal veriye ulaşımın internet kullanımıyla kolaylaşması, araştırmacı ve şirketlerin bu alana olan ilgisini arttırmaktadır. Hem doğrusal olmayan ilişkileri modelleyebiliyor olması hem de büyük miktardaki verinin içerisindeki öznitelikleri keşfedebiliyor olması, makine öğrenmesi ve derin öğrenme yöntemlerinin yoğun olarak tercih edilmesinin temel sebebidir. Finansal piyasaların davranışını etkileyen çok fazla etmen olması ve bu etmenlerin birbirleriyle olan karmaşık ilişkileri farklı veri kaynaklarından elde edilen birçok değişkenin birlikte modellenmesini gerektirmektedir. Son yıllarda finansal verilerle birlikte sosyal medya verilerinin kullanımı ile ilgili araştırmaların sayısı giderek artmaktadır. Bu çalışma kapsamında yüksek frekanslı finansal verilerin sosyal medya verileri de dahil edilerek, makine öğrenmesi ve derin öğrenme yöntemleriyle analizi gerçekleştirilecektir.

Günümüzde sosyal medya sahip olduğu kullanıcı sayılarıyla en önemli iletişim araçlarından biridir ve ekonomik, ticari, kültürel ve bilimsel birçok alana yön verir haldedir. Sosyal medyanın toplum üzerindeki bu etkisinden finansal piyasalarda etkilenmektedir. Örneğin bir hisse senedinin değerinde oluşan ani bir değişim bu hisse senedinin sosyal medyada konuşulmasına sebep olurken, sosyal medyada çok konuşulmaya başlaması hisse senedinde daha büyük değişimler görülmesine sebep olabilmektedir ve böylece zincirleme olarak artan bir etki ortaya çıktığı görülmektedir.

Sosyal medyada yapılan paylaşımın kimin tarafından yapıldığı da büyük önem taşımaktadır. Paylaşım yapan kişinin etkileyebileceği kitlenin büyüklüğü paylaşım etkisini de değiştirmektedir. Fenomen olarak tanımlanan sosyal medyadaki etkili kişilerin sahip oldukları yüksek takipçi sayıları sayesinde çok hızlı bir şekilde geniş kitlelere ulaşabilmekte ve bu kitleleri etkileyebilmektedirler. Bu kişilerin yaptıkları paylaşımlar finansal piyasalarda önemli etkilere neden olmaktadır. Mevcuttaki çalışmalarda yapılan paylaşımların kimin tarafından yapıldığının herhangi bir önemi olmaksızın tüm paylaşımlar aynı önem derecesinde algoritmaya katılmaktadır fakat gerçekte paylaşımı yapan kişinin etkileyebildiği kitlenin büyüklüğü paylaşımın etkisini değiştirmektedir. Fenomen olarak tanımlanan sosyal medyadaki etkili kişilerin sahip oldukları yüksek takipçi sayıları sayesinde çok hızlı bir şekilde geniş kitlelere ulaşabilmekte ve bu kitleleri etkileyebilmektedirler. Bu kişilerin yaptıkları paylaşımlar finansal piyasalarda önemli etkilere neden olmaktadır.

Fenomenlerin finansal piyasalardaki üzerindeki etkilerine birkaç örnek vermek gerekirse, BBC haber sitesinin paylaştığı habere göre, Tesla şirketinin CEO'su Elon Musk'ın X üzerinden yaptığı bir paylaşım Tesla Hisselerinin değerinin 14 milyar dolar düşmesine neden olmuştur ([http-1](#)). NTV haber sitesindeki habere göre Elon Musk'ın X üzerinden yapmış olduğu Signal iletişim programını kullanmayı tavsiye eden paylaşım, Signal uygulaması ile ilişkisi bulunmayan Signal Advances isimli şirketin hisse değerinin %1300 artmasına neden olmuştur ([http-2](#)). Independent haber sitesinin paylaştığı haberde sosyal medyadaki etkili kişilerin paylaşımlarının hisse senetleri üzerinde yaptıkları etkiler paylaşılmıştır ([http-3](#)). İlk olarak 24,5 milyon takipçisi bulunan Kylie Jenner'ın X üzerinden Snapchat şirketi hakkında yapmış olduğu paylaşım şirketin hisse değerinde %7'lik bir düşüş yaşanmasına neden olmuştur. X sitesini yoğun olarak kullanan ABD başkanı Donald J. Trump'ın F-35 programı hakkında yapmış olduğu X paylaşımı programda yer alan şirketlerin hisselerinin 3,5 milyar dolar değer kaybetmesine neden

olmuştur. Donald J. Trump'ın X üzerinden Toyota şirketi hakkında yapmış olduğu bir paylaşım şirket hisselerinin 1,2 milyar dolar değer kaybetmesine neden olmuştur. Hoven [1] yapmış olduğu yüksek lisans tez araştırmasında Donald J Trump'ın X platformundaki paylaşımlarının hisse senetleri üzerinde etkili olduğunu göstermiştir. Kinyua ve ark. [2] yapmış oldukları çalışma ile Donald J Trump'ın X platformundaki paylaşımlarının S&P 500 ve Dow Jones üzerinde negatif etkisi olduğunu göstermişlerdir. Marie Claire dergisinin paylaştığı habere göre, Jeremy Jordan'ın Chipotle şirketinin ürünlerinden yaşamış olduğu gıda zehirlenmesi hakkında yapmış olduğu “Az kalsın ölüyordum” şeklindeki X paylaşımı sonrası şirket hisselerinin değeri tarihinin en düşük seviyesine düşmüştür (http-4). Piñeiro-Chousa ve ark. [3] yaptıkları çalışmada sosyal medyanın yatırımcıların kararları üzerinde etkili olduğunu göstermişlerdir. Bu gibi nedenlerle bu çalışmada, X paylaşımını paylaşan kişinin etkisinin algoritmaya dahil edilmesi amacıyla gönderi istatistikleri kullanılarak yazar etki değeri ve post etki değeri öznitelikleri hesaplanmış ve bu öznitelikler gönderilerin duygu değerleriyle birlikte analize dahil edilmiştir. Bu çalışmada paylaşan kişiyi etkisi de göz önüne alınarak hesaplanan sosyal medya paylaşımlarına ait duygu durum değerlerinin, hisse senedi fiyat değişim yönü tahmininde kullanılmasının toplam kar ve ortalama doğruluk yüzdesi üzerindeki etkisi incelenecektir.

Literatürde makine öğrenmesi ve derin öğrenme algoritmalarıyla, sosyal medya verileri de dahil edilerek, hisse senedi tahmini gerçekleştirilen çalışmalarda şu eksiklikler tespit edilmiştir:

- Çalışmalarda bir veya birden fazla günlük periyotlarda analizler gerçekleştirilmektedir.
- Sosyal medya paylaşımları çalışmalara dahil edilirken paylaşıma ait yazar etki ve post etki istatistikleri göz ardı edilmektedir
- Sosyal medya paylaşımlarının sadece metin verileri için duygu analizi gerçekleştirilirken, paylaşımlardaki emojilerin etkisi göz ardı edilmektedir.
- Çalışmalarda tek bir zaman dönemi için analiz yapılmakta ve farklı zaman dilimlerindeki değişimler karşılaştırılmamaktadır.
- Çalışmalarda iki sınıflı ya da üç sınıflı analiz yöntemlerinden birisi için analizler gerçekleştirilmektedir.

Literatürde tespit edilen bu eksikliklerden yola çıkarak çalışma kapsamında şu geliştirmeler yapılmıştır:

- Çalışma kapsamında gün içi yüksek frekanslı veriler için analizler gerçekleştirilmiştir.
- Çalışma kapsamında paylaşılan kişini etkisinin de analize dahil edilebilmesi için yazar etki ve post etki değerleri hesaplanmış ve bunlar paylaşımların duygu durum değerleriyle birleştirilmiştir.
- Çalışma kapsamında sosyal medya paylaşımlarının hem metin verileri hem de emoji verileri için duygu analizleri gerçekleştirilmiştir ve metin duygu durum değeri ile emoji duygu durum değerinin birleştirildiği bir alternatif değer oluşturulmuştur.
- Çalışma kapsamında borsanın yükseliş, düşüş ve dengeli olduğu 3 zaman dilimi belirlenmiş ve bu 3 dönem için gerçekleştirilen analizlerin sonuçları karşılaştırılmıştır.
- Çalışma kapsamında literatürden farklı olarak hem iki sınıflı hem de üç sınıflı olarak analizler gerçekleştirilmiş ve bu iki durumun karşılaştırması gerçekleştirilmiştir.
- Çalışma kapsamında öznitelik seçimi dönemi belirlenmiş ve bu dönem için tüm öznitelik ve algoritma kombinasyonları için sonuçlar hesaplanarak, en iyi sonucu veren kombinasyon belirlenmeye çalışılmıştır.

Bu çalışma kapsamında paylaşılan yazarın etki gücünün ve postun aldığı etkileşim miktarının da göz önüne alındığı duygu durum değerleri kullanılarak yüksek frekanslı hisse senedi değişim yönünün tahmin edilmesi ile daha yüksek ortalama doğruluk yüzdesine ulaşılabileceğinin ve daha yüksek toplam karlılık elde edilebileceğinin gösterilmesi hedeflenmektedir.

Çalışma 6 ana bölümden oluşmaktadır. Bu bölümde tez çalışmasıyla ilgili kısa giriş yapılmıştır. İkinci bölümde çalışılan konu ile ilgili literatürde yer alan araştırmalar incelenmiştir. Üçüncü bölümde finansal zaman serisi verilerinin analiz yöntemleri ile yapay zeka ve makine öğrenmesi yöntemleri ele alınmıştır. Dördüncü bölümde çalışma kapsamında kullanılan hisse senedi ve sosyal medya verileri ile bu veriler kullanılarak elde edilen öznitelikler detaylı olarak açıklanmıştır. Beşinci bölümde tez çalışmasının uygulaması açıklanmıştır. Bu bölümde toplanan tüm veriler ve öznitelikler kullanılarak hisse senedi fiyat değişim yönünün tahmini gerçekleştirilmiş ve bu tahminler kullanılarak market simülasyonu gerçekleştirilmiştir. Gerçekleştirilen analizler sonucunda ulaşılan toplam kar-zarar, toplam yatırım değeri ve ortalama doğruluk yüzdesi değerleri

paylaşmıştır. Son bölüm olan 6. Bölümde ise tez çalışmasının sonuçları değerlendirilmiştir ve önerilerde bulunulmuştur.



2. LİTERATÜR ARAŞTIRMASI

Birçok farklı alandan araştırmacılar sosyal medya verilerinin finansal zaman serilerinin tahmininde kullanılması üzere araştırmalar gerçekleştirmiştir. Literatürdeki araştırmalar incelendiğinde, sosyal medya verilerinin hisse senedi tahminlerinde kullanılmasıyla ilgili hem sınıflandırma hem de regresyon analizleri gerçekleştirildiği görülmektedir. Bu alanda gerçekleştirile çalışmalar kullanılan algoritmalar açısından makine öğrenmesi ve derin öğrenme algoritmaları olarak 2 gruba ayrılmaktadır.

Makine öğrenmesi algoritmaları kullanılarak sınıflandırma analizi gerçekleştirilen çalışmalara bakıldığında; Zhou ve ark. [4] Sosyal medya ve haber verileri de dahil olmak üzere 5 farklı veri grubunu kullanarak 2017-2018 yılları arasındaki 10 aylık süre için 7 farklı hisse senedinin fiyatlarının hareket yönünü 1gün, 2gün ve 3 günlük periyodlar halinde analiz etmişlerdir. Çalışma kapsamında hisse senedi fiyatlarının hareket yönü ikili sınıflandırma olacak şekilde Support Vector Machine (SVM) algoritması kullanılarak tahmin edilmiştir. Tahminlenecek 7 şirket Eastmoney sitesinde en çok konuşulan 30 şirket arasından farklı sektörlerden olacak şekilde rasgele biçimde seçilmiştir. Araştırmacılar, teknik indikatörler, geçmiş işlem verileri, sosyal medya verisi, haber verisi ve arama motoru hacim verisinden oluşan 5 veri grubundan 14 farklı kombinasyon oluşturularak analizlerde girdi olarak kullanmışlardır. Sosyal medya verileri GUBA platformundaki paylaşımlardan, haber verileri ise Eastmoney, Hexun ve Sina Weibo sitelerindeki hisse senetlerine ait verilerden oluşturulmuştur. Sosyal medya ve haber verileri sözlük tabanlı yaklaşım kullanılarak sınıflandırılmıştır. Teknik indikatör olarak Stochastic oscillator (%K), RSI ve Moving average convergence/divergence (MACD) teknik indikatörleri araştırmacılar tarafından seçilmiştir. Seçilen 7 şirket, 4 tanesi aktif şirketler, 3 tanesi ise aktif olmayan şirketler olarak 2 gruba ayrılmıştır ve grup içerisindeki şirketlerin doğruluk oranlarının ortalaması alınarak sonuç tabloları oluşturulmuştur. Aktif şirketler için SVM algoritması gerçekleştirilen tahminlerin doğruluk oranları %50.39 - %74.91 aralığında elde edilmiştir. Aktif şirketler için en yüksek doğruluk oranına 2 günlük periyotta sadece sosyal medya ve haber duygu durum verilerinin girdi olarak kullanıldığı kombinasyonda ulaşılabilmektedir. Aktif olmayan şirketler için ise doğruluk oranları %48.83 - %69.94 aralığında elde edilmiştir ve en yüksek doğruluk oranına haber duygu durum verileri ve arama motoru hacim verisinin kullanıldığı kombinasyonda ulaşılmıştır. Ren ve ark. [5], sosyal medya verilerini kullanarak SVM algoritmasıyla SSE50 endeksinin değişim yönünü tahmin etmeye çalışmışlardır. SSE50 endeksinde yer

alan 50 hisse senedine ait paylaşımlar EastMoney ve Sine Stock Forum sitelerinden toplamış ve duygu analizleri gerçekleştirilmiştir. Araştırmacılar haftanın günü etkisi de göz önüne alan duygu durumu değerlerini modifiye etmişlerdir. Haftanın günü etkisi finansal piyasalarda pazartesi günleri ortalama getirinin diğer günlerden daha düşük olacağı prensibine dayanmaktadır. 4 tanesi duygu durumlarına dayanan 4 tanesi ise hisse senedi geçmiş verilerine dayanan toplam 8 tane öznitelik oluşturulmuş ve 2 sınıflı sınıflandırma analizi SVM algoritması kullanılarak gerçekleştirilmiştir. Analizler sonucunda Sosyal medya verileri kullanılmadığı durumda %71.33 doğruluk oranına ulaşılabilirken, sosyal medya verilerinin analize dahil edilmesi durumunda %89.93 doğruluk oranına ulaşılmıştır. Araştırmacılar sosyal medya verilerinin hisse senedi değişim yönü tahmininde %18.6 oranında bir iyileşme sağladığını tespit etmişlerdir.

Liu ve ark. [6] StockTwits sosyal medya platformundan toplanan verilere duygu analizi gerçekleştirdikten S&P500 borsasında yer alan SPDR fonunun değişim yönünü tahmin etmişlerdir. Araştırmacılar çalışma kapsamında SVM algoritmasını stok tahminlerine daha uygun hale getirmek için ensembleSVM algoritmasını geliştirmişlerdir. StockTwits platformundan toplanan paylaşımların FinBERT ve VADER yöntemleriyle duygu analizleri gerçekleştirilerek duygu durumları belirlenmiştir. Sonuçlar değerlendirildiğinde çalışma kapsamındaki en iyi F1 skoru olan 0,653 değerine, FinBERT yöntemi ensembleSVM algoritmasıyla ulaşılmıştır. Duygu durumunu içeren özniteliklerin kullanılmasının F1 skoru değerinde %10.3 iyileşme sağladığı görülmüştür. Ghanavati ve ark. [7], Hang Seng Composite Index’de (HSCI) yer alan hisse senetleri için toplanan haber verilerini ve teknik indikatörleri kullanarak hisse senedi hareketlerini 7 günlük ve 30 günlük periyodlar için tahmin etmişlerdir. 9 teknik indikatörle birlikte duygu durum değerlerinden pozitiflerin sayısından negatif olanları çıkardıktan sonra 5 ve 22 günlük basit hareketli ortalamaları hesaplanarak analizlere dahil edilmiştir. Haber kaynaklarından toplanan haber makaleleri ve haber özetlerin duygu analizleri sözlük tabanlı duygu analiz yöntemiyle gerçekleştirilmiştir. Araştırmacılar çalışma kapsamında Fuzzy-Based Local Metric Learning extension based on SVM (FuzyyML-SVM) algoritmasını önermişlerdir. HSCI indeksinde yer alan hisse senetleri küçük, orta ve büyük olmak üzere 3 farklı boyuttaki gruplara ayrılıp bu 3 farklı grup için sonuçlar karşılaştırılmıştır. Sonuçlar değerlendirildiğinde 30 günlük periyod için en iyi sonuçlar önerilen FuzyyML-SVM algoritması ile elde edilirken, büyük grup için 0.18 Hata oranına ve 0.85 F değerine ulaşılmıştır. Kinyua ve ark. [2] Random Forest (RF), Decision Tree

(DT) ve Logistic Regression (LR) algoritmalarını kullanarak Donald J. Trump'ın X platformunda 2017 - 2019 yılları arasında paylaşmış olduğu 2048 gönderinin S&P 500 endeksi (SPX) ve Dow Jones Industrial Average (INDU) endeksi üzerindeki 30 dakikalık (15 dakika öncesi ve sonrası) etkisini incelemişlerdir. Paylaşımların vektör haline getirilmesinde Term Frequency- Inverse Document Frequency (TF-IDF) yöntemi kullanılırken, duygu değerleri VADER kütüphanesi kullanılarak elde edilmiştir. Gerçekleştirilen sınıflandırma analizi sonucunda endeks verileriyle paylaşımların duygu değerlerinin birlikte kullanılması halinden daha yüksek doğruluk değerlerine ulaşıldığı görülmüştür. Çalışma sonucunda INDU için ulaşılan doğruluk oranı %70 - %97 arasında değişirken SPX için %68 - %91 aralığındadır. RF ve DT algoritmaları en doğru sonuçları verirken, LR algoritmasının doğruluğu daha düşük kalmıştır. Yapılan araştırma sonucunda Donald J. Trump'ın X platformundaki paylaşımlarının piyasalar üzerinde belirgin negatif etkisi olduğu saptanmıştır.

Bouktif ve ark. [8] X sosyal medya paylaşımlarını duygu durum değeri ile SVM, RF, Artificial Neural Network (ANN), LR, Naive Bayes (NB) ve Extreme Gradient Boosting (XGBoost) algoritmalarını kullanılarak 2008-2018 yılları arasındaki Amazon, Apple, Microsoft, IBM, Facebook ve Tesla şirketlerinin hisse senedi fiyatlarının 14 günlük değişim yönünü tahmin etmişlerdir. Araştırmacılar sosyal medya paylaşımlarının duygu durum değerlerini belirlemek için Textblob ve VADER kütüphanelerinden faydalanmışlardır. Çalışma kapsamında belirlenen dönem içerisindeki toplam paylaşım sayısı, geçmiş dönem duygu durum değeri ve geçmiş dönem hisse senedi fiyatı gibi özniteliklerde analize dahil edilmiştir. Sonuçlar değerlendirildiğinde 6 farklı şirket için %57.8 - %60.3 aralığında doğruluk oranlarına ulaşılmıştır. Saurabh ve Dey [9], 2010 – 2014 yılları aralığı için X üzerinden toplanan sosyal medya verileriyle sosyal medyadaki ruh halinin FTSE 100 endeksi üzerindeki etkisini incelemişlerdir. Çalışma kapsamında sinirli, korkmuş, mutlu, aşık, üzgün ve gergin olmak üzere 6 farklı duygu durumunu belirten kelime belirlenmiş ve X platformunda saatlik olarak bu kelimeleri içeren paylaşımların oranı tespit edilmiştir ve günlük ortalamaları hesaplanmıştır. FTSE100 endeksinin fiyatları değerlendirilerek düşük, orta ve yüksek olmak üzere 3 sınıfa bölünmüştür. ANN, Discriminant Analysis (DA), SVM ve DT algoritmaları ve duygu durumu verileriyle endeks getirisi kullanarak üç sınıflı sınıflandırma analizi gerçekleştirilmiştir. Analiz sonucunda ANN ile %80.22 doğruluk oranına ulaşılırken, SVM ile %76.58, DT ile %73.3 ve DA ile %61 doğruluk oranına ulaşılabilmektedir. Çalışma

kapsamında sadece mutlu duygu durumu ile endeks getirisi arasında iki yönlü nedensellik olduğu tespit edilmiştir.

Khan ve ark. [10] hem X sosyal medya verilerini hem de Business Insider ile haber verilerini toplayarak duygu analiz yöntemleri ile işlemiştir. Toplanan bu veriler SVM, LR, DT, RF, Multi-Layer Perceptron (MLP), AdaBoost (AB), Extra Tree (ET), Gradient Boosting (GB), Linear Discriminant Analysis (LDA), NB ve K-Nearest Neighbor (KNN) algoritmaları fiyat değişimlerinin yönünü tahmin etmek için kullanılmıştır. Karachi Stock Exchange (KSE), London Stock Exchange (LSE) ve New York Stock Exchange (NYSE) borsa endeksleri ve HP, IBM, Microsoft, Oracle, X, Nokia, Motorola ve Red Hat şirketleri için analizler gerçekleştirilmiştir. Sosyal medya verileri X platformunda hisse senetleri ve endekslerin \$cashtag'leri kullanılarak toplanmıştır. Sosyal medya ve haber verileri Stanford NLP paketi kullanılarak çok negatif, negatif, nötr, pozitif ve çok pozitif olmak üzere 5 sınıflı olacak şekilde duygu durum değerleri belirlenmiştir. Paylaşımların ve haber verilerinin duygu durum değerlerinin toplamı analizlerde girdi olarak kullanılmıştır. Hisse senedi ve endekslere ait açılış, kapanış, en yüksek, en düşük fiyat verileri ve hacim verisi de analizlerde girdi verisi olarak dahil edilmiştir. Hisse senedi ve endekslerin fiyat değişimleri pozitif, negatif ve nötr olmak üzere 3 sınıfa ayrılmıştır. Çalışma kapsamında 1 – 10 gün aralığında tahminler gerçekleştirilmiştir. Hem sosyal medya hem haber verilerinin kullanılması halinde doğruluk değerleri %27.33 – %79.86 aralığında elde edilirken en yüksek değere ET, en düşük değere GB algoritmasıyla ulaşılmıştır. Sadece sosyal medya verilerinin kullanılması halinde doğruluk değerleri %43.04 – %80.53 aralığında elde edilirken en yüksek değere RF, en düşük değer MLP algoritmasıyla ulaşılmıştır. Sadece haber verilerinin kullanılması halinde doğruluk değerleri %22.66 – %75.16 aralığında elde edilirken en yüksek değere RF, en düşük değer GB algoritmasıyla ulaşılmıştır. Sonuçlar değerlendirildiğinde NYSE endeksi, IBM hisse senedinin sosyal medya verilerinden daha çok etkilenirken, LSE endeksi ve Microsoft hisse senedinin haber verilerinden daha çok etkilendiği, X hisse senedinin ise hem sosyal medya hem haber verilerinden etkilendiği tespit edilmiştir.

Qui ve ark. [11] Eastmoney platformundan topladıkları sosyal medya verilerinin duygu analizini Baidu AI Cloud hizmetinde yer alan ön-eğitilmiş duygu analiz algoritmasını kullanarak 3 sınıflı olacak şekilde gerçekleştirmişlerdir. Toplanan bu veriler SVM, RF, LR, KNN, NB, DT, Gradient Boosting Decision Tree (GBDT) algoritmaları ile 2017-2019 yılları arasında SSE 50 endeksinin hareket yönünün pozitif ne negatif

olmak üzere 2 sınıflı olarak tahmin edilmesinde kullanılmıştır. Araştırmacılar bu çalışma kapsamında ağırlıklı duygu endeksi ve haftanın günü etkisini birleştirerek modifiye edilmiş duygu endeksini önermişlerdir. Sonuçlar değerlendirildiğinde doğruluk değerlerinin %62,24 - %68,37 aralığında değiştiği ve en yüksek doğruluk değerine SVM algoritması ile ulaşıldığı görülmektedir. Xiao ve Ihnaini [12] X platformundan ve 6 farklı haber kaynağından topladıkları verileri duygu analiz yöntemleri ile analiz ettikten sonra KNN, DT, SVM, RF, LR ve NB algoritmalarını kullanarak Apple, Amazon, Microsoft ve Netflix şirketlerine ait hisse senetlerinin ertesi günkü hareket yönü analiz etmişlerdir. Toplanan sosyal medya verilerinde 24 saatlik olacak şekilde 2 farklı başlangıç zamanı (00:00 ve 8:30) için gruplar oluşturarak bunları karşılaştırmışlardır. Araştırmacılar çalışma kapsamında duygu durum değerleri daha önceden belirlenmiş X paylaşımları ve haber verilerini kullanarak VADER ve FinBERT yöntemlerini karşılaştırmışlardır. X paylaşımlarının duygu durumları belirlendiğinde VADER yöntemi ile %68 doğruluk oranına ulaşılırken, FinBERT yöntemi ile %53 doğruluk oranına ulaşılabilmiştir. Haber verilerinin duygu durumları için analiz yapıldığında ise FinBERT %86 doğruluk oranına ulaşırken, VADER yöntemi %54 doğruluk oranına ulaşabilmiştir. Yapılan bu analizler sonrasında araştırmacılar X paylaşımlarının duygu durumlarını VADER yöntemiyle, Haber verilerinin duygu durumlarını ise FinBERT yöntemiyle değerlendirmişlerdir. Araştırmacılar X paylaşımları için belirlenen duygu durum değerlerini tekrar paylaşım sayısının bir fazlasıyla çarpmış ve ağırlıklandırılmış duygu durum değerini analizlerde girdi olarak kullanmışlardır. Sonuçlar değerlendirildiğinde Apple hisse senedi için %60.4, Amazon hisse senedi için %62.4 ve Microsoft hisse senedi için %60 doğruluk oranlarına NB algoritmasıyla ulaşılırken Netflix hisse senedi için %58.8 doğruluk oranına LR algoritmasıyla ulaşılmıştır.

Coyne ve ark. [13], StockTwits sosyal medya platformundan toplanan verilere duygu analizi gerçekleştirdikten sonra Linear Regression algoritmasını kullanarak hisse senedi hareketlerini tahmin etmişlerdir. Araştırmacılar sosyal medya paylaşımlarını sınıflandırmak için MLP algoritmasını kullanmışlardır ve her sosyal medya kullanıcısının aynı olmadığı fikrinden yola çıkarak akıllı kullanıcı fikrini önermişlerdir. Toplanan sosyal medya verilerindeki her kullanıcının paylaşımının sınıflandırma sonucuna bakılarak hisse senedi değişim yönünü ne oranda doğru bildiğine bakılmış ve %80'den fazla doğruluk oranına sahip olanlara Akıllı Kullanıcı denilmiştir. Seçilen 9 şirket için belirlenen akıllı kullanıcı sayıları 2-36 arasında değişirken, toplanan paylaşımların sadece

%1.19 - %7.9 oranı kullanılabilmiş ve sadece 7 – 84 gün aralığındaki gün sayısında analiz yapılabilmektedir. Sonuçlar incelendiğinde 9 şirket için doğruluk oranlarının %57.8 - %78.6 aralığında değiştiği görülmüştür. Nam ve Seong [14], haber paylaşımlarını kelime kalıplama yöntemleri ile işledikten sonra Multiple Kernel Learning (MKL) algoritmasını kullanarak 2014-2016 yılları arasındaki Kore borsasındaki şirketlerin hisse senetlerinin fiyat değişim yönünü tahmin etmişlerdir. Kore borsasında yer alan 62 farklı şirkete ait haber verileri Naver platformu kullanılarak 26 farklı haber sitesinden toplanmıştır. Toplanan haber verileri TF-IDF yöntemi kullanılarak vektörel hale getirilerek analize dahil edilmiştir. Çalışma kapsamında 62 adet şirket 3 farklı sektöre ayrılmıştır. MKL algoritmasıyla gerçekleştirilen sınıflandırma analizleri sonucunda sektörel ortalama doğruluk değerleri %55.9 - %62.35 aralığında elde edilmiştir. Ayrıca araştırmacılar hedeflenen şirkete ait haber verileri bulunamasa bile sektördeki ilişkili firmalara ait haber verilerinin hisse senedi fiyatının yönünün tahmininde kullanılabileceğini göstermişlerdir.

Makine öğrenmesi yöntemlerinin kullanıldığı regresyon analizi gerçekleştirilen çalışmalara bakıldığında; Chakraborty ve ark. [15] Stocktwits platformundan topladıkları sosyal medya verilerini ve Boosted Regression Tree algoritmasını kullanarak DJIA endeksi ve Apple hisse senedi günlük kapanış fiyat değişimini tahminlemede kullanmıştır. Sosyal medya paylaşımlarının sınıflandırılmasında SVM algoritması kullanılmış ve algoritmanın eğitilmesi için daha önceden etiketlenmiş sosyal medya paylaşımlarını içeren Sentiment140 veri seti kullanılmıştır. Sonuçlar incelendiğinde Apple hisse senedi için 1.202 Mean Absolute Error (MAE), 1.732 Root Mean Square Error (RMSE) değerlerine ulaşılırken, DJIA endeksi için 82.951 MAE, 116.105 RMSE değerlerine ulaşılmıştır. Maqsood ve ark. [16] 4 ülkenin her birinden 4 farklı hisse senedi seçip X paylaşımlarından yola çıkarak 2012 – 2016 yılları arasındaki yerel ve uluslararası olayların hisse senetleri üzerindeki etkilerini analiz etmişlerdir. Paylaşımlara ait duygu durum değerleri hesaplandıktan sonra gün içerisindeki pozitif paylaşımlarının oranından negatif paylaşımların oranı çıkarılarak olayların günlük duygu durum değerleri elde edilmiş ve nötr olan paylaşımlar göz ardı edilmiştir. Linear Regression, SVM ve Convolutional Neural Network (CNN) algoritmalarını kullanarak hisse senetlerinin günlük fiyatlarının tahmin edilmesi için regresyon analizleri gerçekleştirilmiş ve duygu durum değerinin dahil edildiği sonuçlar, dahil edilmediği sonuçlar ile karşılaştırılmıştır. SVM algoritması ile en düşük MAE ve RMSE değerlerine ulaşılmıştır. Araştırma sonuçları değerlendirildiğinde her yerel olay hisse senetleri üzerinde etkili olmasa da

2012 Amerikan başkanlık seçimi gibi uluslararası etkisi olan olayların hisse senedi tahmin sonuçlarını ciddi oranda iyileştirdiği görülmüştür. Mndawe ve ark. [17] Vodacom Group Limited şirketine ait hem X platformundan sosyal medya verilerini hem de Moneyweb sitesinden haber başlıklarını toplayarak duygu analizi gerçekleştirmişlerdir. Daha sonrasında hem duygu değerleri hem de teknik indikatörlerden oluşan veri seti ile analizler gerçekleştirmişlerdir. LDA, SVM, DT ve RF algoritmalarını kullanarak şirkete ait hisse senedinin değişim yönünü 14 günlük periyotlarda tahmin etmek üzere sınıflandırma analizi gerçekleştirilmiş daha sonrasında 6 farklı Long-Short Term Memory (LSTM) algoritması kullanılarak şirketin kapanış fiyatının 21 günlük periyotlarda tahmin etmek üzerine regresyon analizi gerçekleştirilmiştir. Çalışma kapsamında haber başlıkları ve X paylaşımlarının duygu analizinde Stanford Twitter Sentiment Corpus ve BERT yöntemleri kullanılmıştır. Haber başlıkları, X paylaşımları ve bunların duygu durum değerlerinin kullanıldığı sınıflandırma analizi sonucunda SVM için %49, RF için %74, DT için %82 ve LDA için ise %96 doğruluk oranına ulaşılmıştır. Hisse senedi geçmiş verileri ve 3 farklı teknik indikatörün (MA, MACD, Bollinger Bands) kullanıldığı regresyon analizinde LSTM, Bidirectional LSTM, Stacked (Deep) LSTM, ConvLSTM ve CNNxLSTM katmanları kullanılmış ve analiz sonucunda en iyi sonuca LSTM ile ulaşılmış 5706.57 MSE ve 75.54 RMSE değerleri elde edilmiştir. Araştırmacılar yaptıkları karşılaştırmalı analizlerde duygu durum değeri ve teknik indikatörlerin tahmin doğruluğunu arttırmaya yardımcı olduğunu ortaya koymuşlardır.

Tablo 2.1. Makine öğrenmesi yöntemleriyle yapılan çalışmalar

Yazarlar	Yayın Tarihi	Veri Yılı	Veri Kaynağı 1	Veri Kaynağı 2	Yöntem	Performans Kriteri
Zhou ve ark. [4]	11.2019	2017-2018	GUBA, Eastmoney, Hexun, Weibo	Çin borsasındaki 7 farklı şirket	SVM	Doğruluk
Ren ve ark. [5]	03.2019	2014-2016	Eastmoney, Weibo	SSE 50 Index	SVM	Doğruluk
Liu ve ark. [6]	01.2023	2020-2022	StockTwits	SPDR ETF(S&P500)	SVM	F1-skor
Ghanavati ve ark. [7]	2016	2014-2015	News	Hang Seng Composite Index (HSCI)	FuzzyML-SVM	F-ölçütü, Hata oranı
Kinyua ve ark.[2]	12.2020	2017-2019	X	S&P 500, Dow Jones	RF, DT, LR	Doğruluk, Kesinlik, F1-skor

Tablo 2.1. (Devam) Makine öğrenmesi yöntemleriyle yapılan çalışmalar

Bouktif ve ark. [8]	03.2020	2008-2018	X	Amazon, Apple, Microsoft, IBM, Facebook, Tesla	SVM, RF, ANN, LR, NB, XGBoost	Doğruluk
Saurabh ve Dey [9]	03.2020	2010-2014	X	FTSE 100 index	ANN, DA, SVM, DT	Doğruluk
Khan ve ark. [10]	03.2020	2016-2018	X, Business Insider	HP, IBM, Microsoft, Oracle, X, Nokia, Motorola, Red Hat	SVM, LR, ANN, DT, AB, RF, ET, GB, LDA, KNN	Doğruluk, Kesinlik, Duyarlılık, F1-skor
Qui ve ark. [11]	01.2022	2017-2019	Eastmoney	SSE 50	SVM, GBDT, DT, RF, LR, NB, KNN	Doğruluk, Kesinlik, Duyarlılık, F1-skor
Xiao ve Ihnaini [12]	02.2023	2021	X and News	Apple, Amazon, Microsoft, Netflix	KNN, DT, SVM, RF, LR, NB	Doğruluk, Kesinlik, Duyarlılık, F1-skor
Coyne ve ark. [13]	2018	2016-2017	StockTwits	Amerikan borsalarındaki 9 farklı hisse senedi	Linear Regression, MLP	Doğruluk
Nam ve Seong [14]	2019	2014-2016	News	Kore borsasındaki hisseler	MKL	Doğruluk, F1-skor
Chakraborty ve ark. [15]	2018	2016	Stocktwits	Apple, DJIA	Boosted Regression Tree	MAE, RMSE
Maqsood ve ark. [16]	07.2019	2000-2018	X	16 farklı hisse senedi	Linear Regression, SVM, CNN	RMSE, MAE
Mndawe ve ark. [17]	01.2022	2012-2019	X and News	Vodacom Group Limited	LDA, SVM, DT, RF, LSTM.	Doğruluk, MSE, RMSE

Derin öğrenme algoritmaları kullanılarak sınıflandırma analizi gerçekleştirilen çalışmalara bakıldığında; Li ve ark. [18] Çin menşeli Eastmoney sitesindeki paylaşımları duyarlılık analizi yöntemleri işledikten sonra LSTM algoritmasını kullanarak 2009-2014 yılları arasındaki CSI300 hisse senetlerinin ertesi gün açılış ve kapanış değerlerinin değişim yönünü tahmin etmişlerdir. 5000 adet sosyal medya paylaşımı manuel olarak el ile etiketlenmiş ve Naive Bayes algoritmasının eğitiminde kullanılmıştır. Eğitilen Naive Bayes algoritması kullanılarak sosyal medya platformundan toplanan tüm paylaşımların sınıflandırılması gerçekleştirilmiştir. Araştırma sonuçları incelendiğinde endeks açılış

fiyat yönünün tahmininde yatırımcı duygu değerlerinin dahil edildiği durumda LSTM algoritmasıyla %87,86 doğruluk oranına ulaşılmış ve yatırımcı duygu değerlerinin dahil edilmediği duruma göre %9.29 daha iyi sonuç elde edilmiştir. Kapanış fiyat yönünün tahmininde ise yatırımcı duygu değerlerinin dahil edildiği durumda LSTM algoritmasıyla %45.71 doğruluk oranına ulaşılabilmektedir. Verma ve ark. [19] Nse india haber sitesinden toplanan verileri paragraf vektör yöntemleri ile işledikten sonra LSTM algoritmasını kullanarak 2013-2017 yılları arasındaki 5 farklı NIFTY endeksinin 1,2 ve 5 günlük periyotlarda değişim yönünü tahmin etmişlerdir. Çalışma kapsamında toplanan haber verileri PESTEL analizinde kullanılan 6 başlığa göre sınıflandırılmaktadır. Bu 6 başlık; Politik (Political), Ekonomik (Economic), Sosyal (Social), Teknolojik (Technological), Çevresel (Environmental) ve Yasal'dır (Legal). Haber makalelerinin paragraf vektör yöntemi kullanılarak bu 6 sınıfa ait olma oranları belirlenmekte ve oranlar 6 boyutlu vektör haline getirilmektedir. Çalışma sonuçları incelendiğinde ulaşılan doğruluk oranlarının %50- %58.69 aralığında değiştiği görülmektedir.

Chen ve ark. [20], TSMC, Hon Hai ve Formosa Petrochemical şirketlerinin gün başı ve gün sonu arasındaki değişimin yönü LSTM algoritmasının kullanarak tahminlenmiştir. Araştırmacılar Google üzerinden seçilen şirkete ait haber verilerini toplamış ve metin madenciliği yöntemlerini kullanarak pozitif ve negatif olarak sınıflandırmışlardır. Analizler sonucunda TSMC için %76.32, Hon Hai için %80 ve Formosa Petrochemical %77.42 doğruluk oranına ulaşılmıştır. Wang ve Shirai [21], Sina Weibo sitesindeki yorumlardan elde edilen öznitelik ve LSTM algoritmalarını kullanarak SCI endeksinin değişim yönünü tahmin etmişlerdir. Araştırmacılar sosyal medya paylaşımlarını kullanarak Financial Attention Index (FAI) adını verdikleri bir öznitelik önermişlerdir. FAI endeksi finans ile alakalı yorumların sayısının, finans ile alakalı olmayan yorumların sayısına bölümüne eşittir ve 1 haftalık periyotlar halinde hesaplanmaktadır. Weibo yorumlarının finans ile alakalı olup olmadıklarının sınıflandırılması için SVM algoritması kullanılmıştır. SVM algoritmasının eğitilmesi için Tencent News ve THUC Project'e ait etiketlenmiş veriler kullanılmıştır. FAI endeksinin tahminleme performansına etkisini ölçmek amacıyla LSTM algoritması kullanılarak SCI endeksinin günlük değişim yönü tahminlenmiştir ve SCI endeksinin fiyatının değişiminin ± 0.02 'den küçük olduğu aralık sıfır olacak şekilde 3 sınıflı hale getirilmiş daha sonrasında sıfır ve pozitif olanlar 1, negatif olanlar ise 0 olacak şekilde 2 sınıflı versiyonu da oluşturulmuştur. FAI endeksi girdi olarak kullanılmış ve haftalık olarak hesaplandığından

hafta boyunca aynı değer günlük girdi olarak kullanılmıştır. Sonuçlar değerlendirildiğinde FAI endeksinin kullanıldığı durumda 3 sınıflı olarak %41.2, iki sınıflı olarak %54.6 doğruluk oranına ulaşılmıştır.

Li ve ark. [22] Çin'in en büyük sosyal medya platformlarından biri olan Eastmoney'deki paylaşımlara verileri ve endekse ait verileri kullanarak 2009-2014 yıllarında CSI 300 endeksinin açılış ve kapanış fiyatının yönü tahmin edilmiştir. Araştırmacılar Long-Short Term Memory (LSTM) algoritmasının performansını değerlendirmiş ve sonuçları SVM, LR ve Naive Bayes (NB) algoritmaları ile karşılaştırmışlardır. Eastmoney platformundan toplanan veriler sözlük tabanlı yaklaşım ve NB algoritması kullanarak pozitif, negatif ve nötr olmak üzere 3 sınıfa ayrılmış ve bu verilerin ağırlıklı toplamaları kullanılarak 3 farklı sosyal medya verilerine ait öznitelik oluşturulmuştur. Çalışma kapsamında endeks hareketinin pozitif yönlü ve negatif yönlü olma durumları iki farklı ikili sınıflandırma olarak tahmin edilmiştir. Girdi verisi olarak geçmiş 10 günlük açılış, kapanış, hacim ve duygu durum öznitelik verisi verilmiştir. 3 farklı duygu durum öznitelik verisi için 3 farklı alternatif karşılaştırılmıştır. Sonuçlar değerlendirildiğinde endeks açılış fiyat hareket yönünün tahmininde en yüksek doğru oranı LSTM algoritması ile %78.47 - %80.20 aralığında iken en düşük doğruluk oranı SVM algoritması ile %55 - %57.86 aralığında elde edilmiştir. İndeks kapanış fiyat hareket yönünün tahmininde ise birçok modelin doğruluk oranı %50'nin altında kalmış ve en yüksek doğruluk oranı olan %53.57 değerine NB algoritmasıyla ulaşılmıştır. Araştırmacılar açılış fiyatı tahmininde daha yüksek doğruluk oranına ulaşılmasının piyasa kapandıktan sonra yatırımcıların duyarlılığının etkileyen bilgilerin önemli ölçüde değişmemesinde kaynaklığını, kapanış fiyatı tahmininde düşük doğruluk oranları elde edilmesinin nedeninin ise işlem saatleri içerisinde yatırımcıların yeni bilgiler elde ederek beklentilerini güncellemelerinden kaynaklandığını belirtmişlerdir. Yapılan analizler sonucunda endeks açılış fiyat hareket yönü tahmininde duygu durum değerlerinin kullanılmasının doğruluk oranının %2.86 oranında arttırabildiği görülmüştür. Endeks kapanış fiyat hareket yönü tahmininde ise LSTM veya LR algoritmaları için duygu durum değerlerinin kullanılmasının doğruluk oranının %3.57 oranında arttırabildiği görülmüştür. Albahli ve ark. [23] sözlük tabanlı bir duygu analiz yöntemi olan SentiWordNet 3.0 yönteminin hisse senetleri için özelleştirerek StockSentiWordNet(SSWN) adını verdikleri yeni bir yöntem önermişlerdir. Araştırmacılar önerilen yöntemin performansını X platformundan Apple, Tesla, Microsoft, Walmart, Paypal, Nvidia, Intel, Facebook, X

ve Amazon şirketlerinin hisse senetlerine ait paylaşımları 21 gün boyunca toplayıp duygu durumlarını analiz ederek incelemişlerdir. Şirketlerin hisse senedi fiyatlarının hareket yönü Recurrent Neural Network (RNN) ve Extreme Learning Machine (ELM) algoritmalarıyla tahmin edilmiş ve Generalized Linear Model (GLM), Fast Large Margin (FLM), NB, DT, RF, GBDT ve SVM algoritmalarıyla karşılaştırılmıştır. Analizler sonucunda diğer algoritmalar ile %61-%71 aralığında doğruluk oranına ulaşılabilirken, RNN için %81,40 doğruluk oranına, ELM için ise %86.06 doğruluk oranına ulaşılmıştır.

Vargas ve ark. [24] hem RNN hem de CNN algoritmasının avantajlarından yararlanmak için iki katmanı birlikte kullanarak Recurrent Convolutional Neural Network (RCNN) algoritmasını geliştirmişlerdir. Bu algorithmada verilen öncelikle embedded katmanda sonrasında convolutional katmanda sonrasında da recurrent katmanda işlenmektedir. Reuters haber sitesinden toplanan haber başlıkları 8 farklı kelime kalıplama yöntemleriyle işlenip kelime vektörü haline getirildikten sonra 7 teknik indikatörle birlikte, 2006-2013 yılları arasındaki S&P500 endeksinin günlük değişim yönünün tahmin etmekte kullanılmıştır. Araştırmacılar önerdikleri RCNN algoritmasını CNN ve NN algoritmalarıyla karşılaştırmışlardır. Sonuçlar değerlendirildiğinden en yüksek değer olan %64,21 doğruluk oranına etkinlik kelime kalıplama yöntemi ve CNN algoritmasıyla ulaşılmıştır. Chen ve ark. [25] topladıkları haber verilerini metin madenciliği yöntemleri ile işleyip, 2015-2017 yılları arasındaki HS300 endeksinin değişim yönünü tahmin etmişlerdir. Araştırmacılar Adaboost ve RNN algoritmalarını birleştirerek hibrit bir algoritma olan RNN-Boost algoritmasını önermişlerdir. Çalışma kapsamında kullanılan haber verilerini içeren veri seti, Sina Weibo sosyal medya platformundaki haber kaynakları ile bağlantılı, onaylı hesapların paylaşımları toplanarak oluşturulmuştur. Çalışma sonuçları değerlendirildiğinde önerilen yöntem ile ortalama %66.54 doğruluk oranına ulaşıldığı görülmektedir.

Wang ve ark. [26] teknik göstergeler ile Sina Weibo'dan toplanan sosyal medya verilerini duyarlılık analizi yöntemleri ile işleyip Deep Random Subspace Ensembles (DRSE) algoritmasını kullanarak 2012-2015 yılları arasında SSE endeksinin günlük değişim yönü tahmin etmişlerdir. Bu çalışmada derin öğrenme ile kolektif öğrenme yöntemlerini birleştirerek DRSE algoritmasını geliştirmişlerdir. Duygu analizleri sözlük tabanlı yöntem ile gerçekleştirilmiştir. Tahminlerin gerçekleştirilmesinde duygu değerleriyle birlikte 14 farklı teknik indikatörde öznitelik olarak kullanılmıştır. Sonuçlar incelendiğinde en yüksek doğruluk değeri olan %91'e önerilen DRSE algoritmasıyla

ulaşılmıştır. Huynh ve ark. [27] Gated Recurrent Unit (GRU) algoritması üzerinde değişiklikler yapılarak ileriye ve geriye doğru, her ikisi de aynı çıktı katmanına bağlı olan iki ayrı tekrarlayan ağa sahip olan Bi-directional Gated Recurrent Unit (Bİ-GRU) algoritmasını kullanmışlardır. Reuters ve Bloomberg haber sitelerinden toplanan verileri kelime kalıplama yöntemleri ile işledikten sonra Bİ-GRU algoritmasını kullanarak 2006-2013 yılları arasındaki S&P500 endeksinin değişim yönünü tahmin etmişlerdir. Sonuçlar incelendiğinde en yüksek doğruluk oranı olan %59.98 değerine Bİ-GRU algoritmasıyla ulaşıldığı görülmektedir. Kraus ve Feuerriegel [28] Almanya borsası aracılığıyla yapılan şirket bildirimlerini işledikten sonra LSTM algoritmasını kullanarak 2010-2013 yılları arasındaki CDAX endeksinin 20 günlük fiyat değişim yönünü sınıflandırma analizi ile fiyat değişim miktarını ise regresyon analizi ile tahmin etmişlerdir. Araştırmacılar LSTM algoritmasını RNN ve 8 farklı makine öğrenmesi algoritmasıyla karşılaştırmışlardır. Sınıflandırma analizi sonuçları değerlendirildiğinde en yüksek doğruluk oranı olan %58 değerine LSTM algoritmasıyla ulaşılmıştır. Regresyon analizleri sonucunda ise en küçük RMSE değeri 6.029 ve en küçük MAE değeri 3.011'e yine LSTM algoritması ile ulaşıldığı görülmüştür.

Derin öğrenme algoritmaları kullanılarak regresyon analizi gerçekleştirilen çalışmalara bakıldığında; Zhuge ve ark. [29] SCI endeksinde yer alan 6 şirket için Çin menşeli Eastmoney sitesindeki paylaşımlar toplanmış ve Navie Bayes algoritması kullanılarak duygu durum değerleri sınıflandırılmıştır. Seçilen 6 hisse senedi için belirlenen duygu durum değerleri ve geçmiş hisse senedi verileri ile LSTM algoritması kullanılarak hisse senetlerinin ertesi gün açılış değerlerini tahmin etmişlerdir. Sonuçlar değerlendirildiğinde önerilen yöntem ile 0.000242 MSE değerine ulaşılmıştır. Wu ve ark. [30], geçmiş hisse senedi verileri, teknik indikatörleri ve Eastmoney platformundan topladıkları verilerden elde edilen duygu analiz endeksi değeri ile Çin'de yer alan 5 bankanın hisse senedi kapanış fiyatlarını tahmin etmişlerdir. Araştırmacılar geliştikleri S_I_LSTM algoritması ise analizleri gerçekleştirmişlerdir. Duygu analizlerinin gerçekleştirilmesi için gün içerisindeki getirinin pozitif ya da negatif olması durumuna göre toplanan sosyal medya ve haber verilerini etiketlemişlerdir ve CNN algoritmasını kullanarak duygu durum değerleri tahmin edilmiştir. 5 şirket için ortalama sonuçlara bakıldığında 2.38 olan en düşük MAE değeri tüm özniteliklerin kullandığı durumda elde edilmiştir.

Tablo 2.2. *Derin öğrenme yöntemleriyle yapılan çalışmalar*

Yazarlar	Yayın Tarihi	Veri Yılı	Veri Kaynağı 1	Veri Kaynağı 2	Yöntem	Performans Kriteri
Lİ ve ark. [18]	06.2017	2009-2014	Eastmoney	CSI300	LSTM + Naive Bayes	Doğruluk, Kesinlik, Duyarlılık, F1-skor
Verma ve ark. [19]	2017	2013-2017	News (nseindia.com)	NIFTY50 Index	LSTM	Doğruluk, MCC
Chen ve ark. [20]	03.2019	2016-2017	Google News	Taiwan 50 Index	LSTM	Doğruluk
Wang ve Shirai [21]	03.2018	2013-2016	Weibo	SCI index	LSTM	Doğruluk
Li ve ark. [22]	2020	2009-2014	Eastmoney	CSI 300 index	LSTM, SVM, LR, NB	Doğruluk
Albahli ve ark. [23]	10.2022	2021	X	Apple, Tesla, Microsoft, Walmart, Paypal, Nvidia, Intel, Facebook, X, Amazon	RNN, ELM	Doğruluk, Kesinlik, Duyarlılık, F1-skor
Vargas ve ark. [24]	2017	2006-2013	News (Reuters)	S&P500	RCNN, LSTM, CNN	Doğruluk
Chen ve ark. [25]	08.2018	2015-2017	Weibo	HS300 (Shangai)	RNN	Doğruluk, MAE, MAPE, RMSE
Wang ve ark. [26]	03.2018	2012-2015	Weibo	Çin borsası hisse senedi verileri	Deep Random Subspace Ensembles (DRSE)	Doğruluk, Kesinlik, Duyarlılık, F1-skor, AUROC
Huynh ve ark. [27]	12.2017	2006-2013	News (Reuters + Bloomberg)	S&P500 hisse senetleri	Bİ-GRU	Doğruluk
Kraus ve Feuerriegel [28]	10.2017	2010-2013	German ad hoc announcements	CDAX borsa verisi	LSTM	MSE, RMSE, MAE, Doğruluk, AUC
Zhuge ve ark. [29]	05.2017	2008-2015	Eastmoney	NetEase	LSTM	MSE
Wu ve ark. [30]	06.2021	2017-2020	EastMoney	Çin merkezli 5 banka	S_I_LSTM	MAE, MSE, RMSE

Literatürdeki çalışmalar incelendiğinde algoritmalar kullanılarak elde edilen tahminlerin market simülasyonu oluşturularak alım-satım işlemlerinde kullanıldığı ve böylece karlılık analizleri gerçekleştirildiği görülmektedir. Market simülasyonu ile karlılık hesaplayan çalışmalara bakıldığında; Minh ve ark. [31] GRU algoritması üzerinde değişiklikler yaparak Two-stream Gated Recurrent Unit (TGRU) algoritmasını geliştirmişlerdir. Bu algortmada Eğitim ileri ve geri iki ayrı tekrarlayan ağı bölünmekte ve her ikisi de çıktı katmanı olarak birleştirilmektedir. Reuters ve Bloomberg haber sitelerinden toplanan verileri kelime kalıplama yöntemleri ile işledikten sonra TGRU ve Stock2vec algoritmalarını kullanarak 2006-2013 yılları arasındaki Apple, Airbus ve Amazon hisse senedi fiyatlarının değişim yönünü tahmin etmişlerdir. Haber sitelerinden toplanıp işlenen verilerle birlikte Stochastic oscillator (%K), William (%R) ve Relative Strength Index (RSI) teknik indikatörleri de girdi olarak analizlere dahil edilmiştir. Analizler sonucunda Apple için %72.3, Airbus için %66.1 ve Amazon için %63.23 doğruluk oranına ulaşılmıştır. Araştırmacılar aynı zamanda Google şirketine ait GOOGL hisse senedi için 9 Ekim – 9 Kasım 2017 tarihleri arasındaki verileri kullanarak market simülasyonu gerçekleştirmişlerdir. Oluşturulan alım-satım stratejisi şu şekildedir;

- Eğer haber verisi mevcut değil ise işlem yapma
- Eğer hisse senedi alınmış ve tahmin negatif ise, hisse senetlerini sat
- Eğer hisse senedi alınmış ve tahmin pozitif ise, işlem yapma
- Eğer tahmin pozitif iken hisse senedi satılmış ise, işlem yapma
- Eğer tahmin negatif ve hisse senedi satılmış ise, eldeki tüm para ile hisse senedi satın al

Market simülasyonu kapsamında 1 aylık süre için 10.000\$ yatırım yapılmış ve %0.25 alım-satım işlem ücreti uygulanmıştır. 1 aylık süresinin sonunda yatırımın değeri 10.531\$ olmuştur. Araştırmacılar GOOGL hisse senedinde belirlenen süre için al-tut stratejisiyle elde edilebilecek kar miktarı hakkında bir veri paylaşmamışlardır. Ding ve ark. [32] Reuters ve Bloomberg haber sitelerinden toplanan verileri Neural Tensor Network yöntemiyle işlendikten sonra CNN algoritması kullanarak 2006-2013 yılları arasındaki S&P500 endeks fiyatının değişim yönünü tahmin etmişlerdir. Analizler sonucunda önerilen yöntem ile endeksin günlük değişim yönü %65.08 doğruluk oranıyla tahmin edilmiştir. S&P500 endeksinde yer alan 15 hisse senedi içinde analizler gerçekleştirilmiş ve hisse senetlerine bağlı olarak %56- %76 aralığında doğruluk oranlarına ulaşılmıştır. Araştırmacılar 3 Ocak 2000 tarihinden sonraki 40 günlük periyod

için market simülasyonu ile karlılık analizi gerçekleştirmişlerdir. Oluşturulan market simülasyonunda kullanılan alım-satım strateji şu şekildedir;

- Eğer hisse senedi fiyatının yukarı yönlü olacağı tahmin edilmiş ise 10,000\$ değerinde hisse senedi alınmaktadır. Hisse senedi alındıktan sonra gün içerisinde %2 oranında bir artış olması halinde anında alınan tüm hisse senetleri satılmaktadır, eğer istenen artış olmaz ise gün sonunda kapanış fiyatından tüm hisse senetleri satılmaktadır.
- Eğer hisse senedi fiyatının aşağı yönlü olacağı tahmin edilmiş ise açığa satış şeklinde 10,000\$ değerinde hisse senedi yatırımı yapılmaktadır. Hisse senedi yatırımı yapıldıktan sonra gün içerisinde %1 oranında bir fiyat düşüşü olması halinde anında açığa satışı kapatmak için hisse senedi alımı gerçekleştirilmektedir, eğer istenen düşüş olmaz ise gün sonunda kapanış fiyatından hisse senetleri alınarak açığa satış kapatılmaktadır.

Uygulanan alım-satım stratejisi incelendiğinde hem yükseliş hem de düşüş tahmin edilmesi durumlarında 10,000\$ değerinde yatırım yapıp, kar almak üzerine kurulduğu ve herhangi bir işlem ücreti uygulanmadığı görülmektedir. Gerçekleştirilen analizler sonucunda 15 şirket için ortalama 16,785\$ kar elde edilmiştir. Akita ve ark. [33] Nikkei Gazetesinde yer alan haber verileri paragraf vektör yöntemleri ile işledikten sonra 2001-2008 yılları arasındaki Nikkei 225 endeksinde yer alan 10 hisse senedinin kapanış fiyatlarını tahmin etmişlerdir. Çalışma kapsamında haber verisi vektörünü ve hisse senedi açılış fiyatı verisi kullanılarak, LSTM algoritmasıyla kapanış fiyatı tahmin edilmiştir. Araştırmacılar tahmin edilen kapanış fiyatlarını kullanarak 1 yıllık periyot için market simülasyonu ile karlılık analizi gerçekleştirmişlerdir. Oluşturulan market simülasyonunda kullanılan alım-satım strateji şu şekildedir;

- Eğer hisse senedi kapanış fiyatının açılış fiyatından yüksek olacağı tahmin edilmiş ise 1,000,000 Japon Yen'i değerinde hisse senedi alınmaktadır. Hisse senedi alındıktan sonra gün içerisinde %2 oranında bir artış olması halinde anında alınan tüm hisse senetleri satılmaktadır, eğer istenen artış olmaz ise gün sonunda kapanış fiyatından tüm hisse senetleri satılmaktadır.
- Eğer hisse senedi kapanış fiyatının açılış fiyatından daha düşük olacağı tahmin edilmiş ise açığa satış şeklinde 1,000,000 Japon Yen'i değerinde hisse senedi yatırımı yapılmaktadır. Hisse senedi yatırımı yapıldıktan sonra gün içerisinde %2 oranında bir fiyat düşüşü olması halinde anında açığa satışı kapatmak için hisse

senedi alımı gerçekleştirilmektedir, eğer istenen düşüş olmaz ise gün sonunda kapanış fiyatından hisse senetleri alınarak açığa satış kapatılmaktadır.

Uygulanan alım-satım stratejisi incelendiğinde hem yükseliş hem de düşüş tahmin edilmesi durumlarında 1,000,000 Japon Yen'i değerinde yatırım yapıp, kar almak üzerine kurulduğu ve herhangi bir işlem ücreti uygulanmadığı görülmektedir. Gerçekleştirilen analizler sonucunda 10 şirket için toplamda 284,190 Japon Yen'i kar elde edilmiştir. Algoritmanın başarısı market simülasyonu sonucu elde edilen kar miktarı üzerinden değerlendirilmiştir ve algoritmaya ait performans metriklerinin sonuçları paylaşılmamıştır. Hagenau ve ark. [34] DGAP (Deutsche Gesellschaft für Adhocc-Publizität) ve EuroAdhoc'dan topladıkları haber verilerini metin madenciliği yöntemleri ile analiz etmiş ve SVM algoritmasını kullanarak sınıflandırma analizi gerçekleştirilmiştir. Araştırmacılar dışsal geribildirim tabanlı öznelik seçimi yöntemi önermiş ve diğer yöntemler ile karşılaştırmışlardır. Sınıflandırma analizi sonucunda elde edilen pozitif ve negatif sinyaller simülasyonda kullanılmıştır. Alman borsasında yer alan en likit 110 hisse senedinden bir birim için alım-satım yapılarak ortalama getiriler hesaplanmış ve metin madenciliği yöntemleri için karşılaştırılmıştır. En yüksek değer olan %1.1'lik getiri oranına araştırmacılar tarafından önerilen ki-kare tabanlı yöntem ile ulaşılmıştır.

Dang ve Duong [35], Vietnam haber sitesinden topladıkları haber verilerini metin madenciliği yöntemleri ile değerlendirip SVM algoritmasını kullanarak sınıflandırmışlardır. Araştırmacılar VN30 endeksinin fiyat değişim yönüne bakarak, fiyat değişimi negatif ise o gün paylaşılan haberleri negatif, pozitif ise pozitif, değişim yok ise nötr olarak etiketlemişlerdir. Elde edilen sınıflandırma sonuçları VN30 endeksindeki şirketler arasından 5 tanesi seçilerek fiyat değişim yönleri tahmin edilmiş ve %64 doğruluk oranı elde edilmiştir. Araştırmacılar elde edilen tahmin sonuçlarının gerçek şartlar altında değerlendirmek için MSN hisse senedinden 100 milyon Vietnam Dong değerinde hisse senedi almışlar ve 2-13 Mart 2015 tarihleri arasındaki 10 işlem gününde market simülasyonu gerçekleştirmişlerdir. Market simülasyonu kapsamında 4 alım-satım işlemi gerçekleştirmiş ve 10 işlem günü sonunda 105.3 Milyon Vietnam Dong değerinde yatırım elde edilmiştir.

Das ve ark. [36] bu çalışmada X paylaşımlarının gerçek zamanlı olarak toplanıp, duygu analizlerinin gerçekleştirilmesinden sonra hisse senedi fiyat tahmininde kullanılması için gerekli sistem kurulumunu ve analizleri gerçekleştirmişlerdir. 2016-

2017 yılları arasındaki Google, Apple ve Microsoft hisse senetlerine ait X paylaşımları API hizmeti üzerinden toplandıktan sonra HDFS sistemi, Apache Spark hizmeti ve MLlib kütüphanesi kullanılarak işlenmiştir. İşlenen X paylaşımları Stanford Core NLP API hizmeti kullanılarak duygu durum değerleri belirlenmiştir. Duygu durum değeri ve geçmiş hisse senedi verileri kullanılarak hisse senetlerinin fiyat tahminleri gerçekleştirilmiştir. Araştırmacılar bu çalışma ile gerçek zamanlı sosyal medya paylaşımlarının hisse senedi fiyatı tahmininde kullanılabileceğini göstermiştir.

Literatürde hisse senedi tahmini yapılmaksızın sadece finansal sosyal medya paylaşımları toplanarak duygu durum analizinin gerçekleştirildiği çalışmalarda mevcuttur [37-43].

Tablo 2.3. *Market simülasyonu gerçekleştirilen çalışmalar*

Yazarlar	Yayın Tarihi	Veri Yılı	Veri Kaynağı 1	Veri Kaynağı 2	Yöntem	Kar Yüzdesi
Minh ve ark. [31]	10.2018	2006-2013	News (Reuters + Bloomberg)	Apple, Airbus, Amazon Stock Prices	T-GRU	%5,31
Ding ve ark. [32]	2015	2006-2013	News (Reuters + Bloomberg)	S&P500'den 5 hisse senedi	CNN	%67,85
Akita ve ark. [33]	06.2016	2001-2008	News	Nikkei 225'ten 10 hisse senedi	LSTM	%28,42
Hagenau ve ark. [34]	2012	1998-2010	DGAP, EuroAdhoc	Almanya borsa verileri	SVM	%1,1
Dang ve Duong [35]	2016	2014-2015	News	VN30 Endeksinde yer alan hisse senetleri	SVM	%5,3

3. YÖNTEM

Bu bölümde finansal zaman serilerinin analiz yöntemleri, yapay zeka ve makine öğrenmesi yöntemleri ve doğal dil işleme yöntemleri açıklanmaktadır.

3.1. Finansal Zaman Serilerinin Analizi

Finansal zaman serileri, finansal enstrümanların (hisse senedi, döviz, altın, petrol, tahvil senedi, fon vb.) belirli zaman periyotlarındaki değişimlerini gösteren sayısal verilerdir. Finansal zaman serisi analizlerinin temel amacı finansal enstrümanların fiyat değişimlerinin miktarı, değişimin yönü ve değişimin zamanın ölçülmesi ve gelecekte oluşacak hareketlerin tahmin edilmesidir.

Finansal zaman serisi verilerinin analizinde kullanılan yöntemler temel analiz, teknik analiz ve matematiksel analiz yöntemleri olarak sayılabilir fakat bu yöntemlerin yanı sıra gelişen teknolojik imkanlarla birlikte yapay zeka yöntemleri de sıklıkla kullanılır hale gelmiştir. Finansal verilerin analizinde yapay zeka yöntemlerinin altında yer alan makine öğrenmesi ve derin öğrenme yöntemleri kullanımı her geçen gün artış göstermektedir.

3.1.1. Temel analiz

Temel analiz, makro ekonomik göstergeler (para arz miktarı, döviz kurları, faiz oranları, enflasyon oranları, işsizlik oranları vb.) ve şirketin geçmiş mali verileri (temettü getirisi, kazanç getirisi, nakit akışı getirisi, defter-piyasa oranı, fiyat-kazanç oranı vb.) sektörel durumları da göz önüne alıp incelenmesiyle şirketin gerçek değerinin ölçülmeye çalışılması ve bu gerçek değer ile piyasa değerinin karşılaştırılmasıyla şirketin değerlendirildiği analiz yöntemidir [44].

Temel analiz yöntemi sırasıyla ekonomik analiz, sektörel analiz, şirket analizi ve değerlendirme olmak üzere 4 aşamadan oluşmaktadır. Temel analizin ilk aşaması olan ekonomik analizde finansal yatırım yapmak için genel ekonomik konjonktürün uygun olup olmadığı değerlendirilir. İkinci aşamadan ise sektörel analiz gerçekleştirilir ve hangi sektörlerle yatırım yapılabileceğinin değerlendirmesi yapılır. Sonraki aşamada belirlenen sektörlerdeki şirketlere ait hisse senetlerinin gerçek değerleri hesaplanır. Son aşamada ise toplanan verilerin değerlendirilmesiyle hangi hisse senetlerine yatırım yapıp yapılmayacağına karar verilir [45].

3.1.2. Teknik analiz

Teknik analiz karar vermede fiyatı ve hacmi kullanmaktadır ve genellikle verileri grafikler üzerinden incelemektedir. Teknik analiz kısa veya uzun dönem fark etmeksizin istenilen her zaman periyodu için uygulanabilmektedir. Teknik analizin altındaki mantık şu şekildedir; arz ve talep fiyatı belirler, arz ve talepteki değişimler fiyatta değişime neden olur ve fiyatlar grafiklerle ve teknik araçlarla tahmin edilebilir [46].

Teknik analiz, yatırımcı psikolojisi ve duyarlılığının incelenmesi olarak düşünülebilir. Piyasada fiyatlar, insanlar (veya onların alım-satım programları) tarafından belirlenir ve fiyat arz ve talep arasında dengededir. Temel analiz uzmanları piyasaların rasyonel olduğunu öne sürmektedirler fakat teknik analiz uzmanları, insanların çoğunlukla mantıksız ve duygusal olduğuna ve benzer durumlarda benzer şekilde davranma eğiliminde olduklarına inanmaktadır. İnsan davranışları hayatın birçok alanında, düzensizdir ve duygular tarafından yönlendirilmektedir bu nedenle teknik analiz uzmanları, yatırımın bir istisna olmayacağını düşünmektedirler. Dolayısıyla teknik analiz, piyasa eğilimlerinin veya desenlerinin incelenmesidir ve teknik analizciler piyasa eğilimlerinin ve desenlerin tekrarlama eğiliminde olduğunu ve bu nedenle tahmin edilebileceğine inanmaktadır [46].

Teknik analizin geçmişi temel analize kıyasla çok daha eskidir çünkü temel analiz için gerekli olan güvenilir verilere ulaşımın kolaylaşması çok yakın zamanda mümkün olmuştur. 1700’lerde Japonya’da pirinç piyasasının analizi için teknik analiz yöntemlerinin kullanıldığına dair kayıtlara ulaşılmıştır. Bu kayıtlarda grafik tasarımlarının ve desenlerin yer aldığı görülmüştür. Teknik analizin batıdaki ilk kullanımı ise 1890’larda Charles Dow öncülüğünde Amerika’da olmuştur [46].

Yatırımcılar yükseliş beklentisine sahip oldukları piyasa için “Boğa piyasası” terimini kullanmaktadırlar. Yatırımcılar bu beklentiye sahip olduklarında tüm imkanlarıyla, gerekirse kredi veya borç yoluyla, satım alım gerçekleştirmeye çalışmakta ve yükselişin ardından yüksek fiyattan satarak kar elde etmeyi hedeflemektedirler. Boğa piyasası teriminin ortaya çıkışında, boğanın saldırırken boynuzlarıyla her şeyi yukarı kaldırmasından esinlenildiği düşünülmektedir [47].

Yatırımcılar düşüş beklentisine sahip oldukları piyasa için ise “Ayı piyasası” terimini kullanmaktadırlar. Yatırımcılar düşüş beklentisinde olduklarında riski göz alarak ellerindeki yatırımları satmaya çalışırlar ve düşüşün ardından düşük fiyattan geri almayı hedeflemektedirler. Ayı piyasası terimin ortaya çıkışına dair ilk görüş “Ayıyı

yakalamadan derisini satmak” deyiminden esinlenildiği düşünülmektedir [47]. Diğer bir görüş ise ayıların saldırırken pençelerini aşağı savurmalarından esinlenildiği düşüncesidir.

3.1.3. Matematiksel analiz

Finansta matematiksel analiz, finansal zaman serisi verilerinin en büyük, en küçük, ortalama, hareketli ortalama, kovaryans, varyans, standart sapma gibi değerlerinin hesaplanmasıdır. Matematiksel analizde, matematiksel fonksiyonlar kullanılarak zaman serilerinin geçmiş değerlerinin incelenmesiyle gelecekte oluşabilecek değerler tahmin edilmeye çalışılmaktadır. Regresyon analizi gibi yöntemlerle verilerin birbirleriyle olan ilişkileri incelenerek matematiksel olarak ifade edilmektedir.

Otoregresyon Hareketli Ortalama (Autoregressive Moving Average -ARMA) ve Otoregresyon Entegreli Hareketli Ortalama (Autoregressive Integrated Moving Average -ARIMA- Box-Jenkins) modelleri, zaman serilerinin gelecek değerlerini tahmin etmekte kullanılan en önemli modellerdendir. ARMA modelinin genelleştirilmiş hali olan ARIMA modeli, George Box ve Gwilym Jenkins tarafından 1970 yılında yazılan “Time series analysis: forecasting and control” isimli kitapta tanımlanmıştır [48]. ARMA modeli ise Peter Whittle tarafından 1951 yılında yazılan “Hypothesis testing in time series analysis” isimli doktora tezinde tanımlanmıştır [49]. Bu modeller zaman serilerinin geçmiş verilerine ve olasılıksal hata terimlerine bakarak gelecekteki verilerin tahmin edilmesini sağlamaktadır [50].

3.2. Yapay Zeka, Makine Öğrenmesi ve Derin Öğrenme

Yapay zeka, insan beyninin, düşünce yapısını, öğrenme ve karar verme gibi yeteneklerini taklit ederek bunların makineler kullanarak modellenmesini amaçlamaktadır. “Yapay Zeka” terimi ilk kez 1956 yılında John McCarthy tarafından bir akademik konferansta kullanılmıştır. Fakat makinelerin gerçekten düşünebilir mi fikri bundan çok önce ortaya çıkmıştır. 1945’te Vannevar Bush [51] tarafından yayınlanan “As We May Think” isimli çalışmada insan zeka ve algısını kuvvetlendiren sistemler önermiştir. 1950 yılında Alan Turing [52] tarafından yayınlanan “Computing Machinery and Intelligence” makalesinde insan davranışlarını taklit eden ve satranç oynamak gibi zeka gerektiren şeyleri yapma yeteneğine sahip makinelerden bahsedilmiştir [53]. 1959 yılında Atatürk üniversitesinde Ord. Prof. Dr. Cahit Arf [54] tarafından sunulan “Makine

Düşünebilir Mi ve Nasıl Düşünebilir” başlıklı konferansta düşünen makineler kavramı açıklanmış ve makinelerin nasıl düşünebildiklerinden bahsedilmiştir.

Makine öğrenmesi, bir probleme ait verileri kullanarak o problemi modelleyen algoritmalar ve kendi kendine öğrenme yeteneğine sahip bu algoritmalar makinelerin veri kümelerinden öğrenmesini sağlar. Makine öğrenmesinin, yapay zekanın bir alt dalı olduğu söylenebilir. “Makine öğrenmesi” terimi ilk olarak IBM çalışanı olan Arthur Samuel [55] tarafından 1959’da icat edilmiştir. Makine öğrenimi ile ilgili çalışmalar 1965’te Nilsson [56], 1973’de Duda ve Hart [57] tarafından örüntü sınıflandırma ve tanımlamada kullanılmasıyla devam etmiştir. Makine öğrenmesi algoritmaları denetimli öğrenme (supervised learning), denetimsiz öğrenme (unsupervised learning) ve pekiştirmeli öğrenme (reinforcement learning) olmak üzere üç grupta değerlendirilmektedir.

Denetimli öğrenme yönteminde önceden belirlenmiş etiketler kullanılarak model oluşturulmaktadır. Oluşturulan model kullanılarak eldeki verilere sınıflandırma gerçekleştirilmektedir. Denetimli öğrenmede veriler eğitim, doğrulama ve test olarak 3 parçaya bölünmektedir. İlk aşamada eğitim için ayrılan veriler kullanılarak model oluşturulmakta ve eğitilmesi sağlanmaktadır. Doğrulama verisi kullanılarak model optimize edilmekte, test veri kullanılarak ise modelin başarısı ölçülmektedir. Karar ağaçları, rastgele orman, destek vektör makineleri, doğrusal regresyon ve lojistik regresyon gibi birçok denetimli öğrenme algoritması geliştirilmiştir.

Denetimsiz öğrenme algoritmasında herhangi bir eğitim aşaması bulunmamaktadır. Denetimsiz öğrenme, verilerin arasındaki bağlantıların bulunmasıyla birbirine benzer verilerin kümelenmesiyle sınıflandırma gerçekleştirilir. K-means, K-medoids ve Fuzzy C-Means gibi birçok denetimsiz öğrenme algoritması geliştirilmiştir.

Pekiştirmeli öğrenme yöntemlerinde deneme yanılmaya bağlı olarak tecrübe kazanımı ile öğrenme gerçekleştirilmektedir. Pekiştirmeli öğrenmede ödül ve ceza kavramları bulunmaktadır. Başarılı çıktılar artı puan ile ödüllendirilirken, başarısız çıktılar eksi puan ile cezalandırılmaktadır. Böylece algoritmanın başarılı sonuçlar elde etmesi hedeflenmektedir. Pekiştirmeli öğrenme yöntemleri oyun teorisi, kontrol teorisi, bilgi teorisi, simülasyon tabanlı optimizasyon ve çoklu ajan sistemleri gibi birçok alanda kullanılmaktadır. Q-learning ve R-learning gibi birçok pekiştirmeli öğrenme algoritması geliştirilmiştir.

Derin Öğrenme ise makine öğrenmesinin alt alanıdır. Çok katmanlı modellerle yapılan yapay öğrenmedir. Yapay sinir ağlarındaki gizli katmanın sayısının artırılmış ve geliştirilmiş hali derin öğrenmedir. Katmanlar, özellik çıkartma ve dönüştürme işlemleri için birçok doğrusal olmayan işlem birimi içerir ve her ardışık katman, bir önceki katmandaki çıktıyı girdi olarak kullanmaktadır [58]. Zaman içerisinde farklı problemlerin çözümleri için farklı yapıda derin sinir ağları geliştirilmiştir. Çok katmanlı derin yapay sinir ağları (Deep multilayer perceptron-DMLP), Yinelemeli sinir ağları (Recurrent neural networks-RNN), Uzun kısa vadeli hafıza (Long Short Term Memory-LSTM), Evrişimsel sinir ağları (Convolutional neural networks-CNN) derin sinir ağları modellerine örnek olarak sayılabilir.

Bu çalışma kapsamında uygulama analizlerinde dört farklı makine öğrenmesi ve bir adet derin öğrenme algoritması kullanılmıştır. Bunlar Karar ağaçları, Rastgele orman, Destek Vektör Makinesi, Lojistik regresyon ve Uzun Kısa Vadeli Hafıza Ağlarıdır.

3.2.1. Karar ağaçları

Bir karar ağacı (Decision Tree – DT) yapısı düğüm, dal ve yaprak olmak üzere 3 temel parçadan oluşmaktadır. Karar ağaçları, kök adı verilen ve girdisi olmayan ilk düğüm ile başlamakta ardından çıktıları bir başka düğümün girdisi olan iç düğümler ile devam etmektedir. Karar ağacı yapısında çıktısı bir başka düğüme girdi olmayan son düğümlere ise yaprak düğüm adı verilmektedir. Karar ağacı yapısında her iç düğüm en az iki parçaya ayrılmaktadır. Karar ağacı algoritmaları ile karar ağaçları oluşturulurken hata oranının en az olduğu ağaç yapıları oluşturulmaya çalışmaktadır [59]. Karar ağacı yapısındaki dallanma kriterlerinin belirlenmesi karar ağacı algoritmasının başarısının artırılmasında büyük öneme sahiptir. GINI indeksi, Ki-kare istatistiği ve Bilgi Kazancı gibi yaklaşımlar dallanma kriterlerinin belirlenmesinde kullanılan başlıca yaklaşımlardır. Karar ağacı algoritmasının performansının artırılmasında kullanılan yaklaşımlardan biri de budama yöntemidir. Budama yönteminde istatistiksel olarak geçerliliği düşük alt ağaçların ortadan kaldırılmasıyla karmaşıklığın azaltılması ve böyle ağaç yapısının basitleştirilmesi hedeflenmektedir [60].

Karar ağaçlarının ilk örneklerinden olan 1966 yılında Morgan ve Sonquist [61] adlı araştırmacılar tarafından yayınlanan Automatic Interaction Detector (AID) projesidir. Classification And Regression Tree (CART) algoritması ise AID algoritmasının devamı niteliğinde olup 1984 yılında Breiman ve diğer araştırmacılar [62] tarafından önerilmiştir.

CART algoritmasında öncelikle karar ağacının herhangi bir durma kuralı olmaksızın dallanmasına izin verilmekte ve ardından uçtan köke doğru budama işlemi gerçekleştirilmektedir. CART algoritması dallanma kriteri olarak GINI indeksi yaklaşımını kullanmaktadır. GINI indeksi, bir düğümdeki rastgele seçilen noktanın yanlış sınıflandırılma olasılığının ölçütüdür. Denklem 3.1.'de GINI indeksinin hesaplanma formülü gösterilmiştir. GINI indeks formülünde bulunan n ; sınıf sayısını, p_i ise düğümdeki veri noktalarının i . sınıfa ait olma olasılığını göstermektedir.

$$GINI = 1 - \sum_{i=1}^n p_i^2 \quad (3.1)$$

Bir diğer karar ağacı algoritması olan ID3 ise J.R. Quinlan [63] adlı araştırmacı tarafından 1986 yılında yayınlanmıştır. 1993 yılında J.R. Quilan [64] tarafından ID3 karar ağaçları algoritmasının eksiklikleri giderilmiş ve algoritma geliştirilerek C4.5 algoritması yayınlamıştır. C4.5 algoritması bilgi kazancı yaklaşımını dallama kriteri olarak kullanmaktadır. ID3 algoritması sadece kategorik veriler ile çalışabilirken C4.5 algoritması hem kategorik hem de numerik verilerle çalışabilmektedir. C4.5 algoritmasının çözüm hızı artırılıp, büyük veri setlerinde de çalışabilir hale getirilerek C5.0 algoritması yayınlanmıştır [65]. Bu çalışma kapsamında analizlerin gerçekleştirilmesinde CART algoritması kullanılmıştır.

3.2.2. Rastgele orman

Rastgele Orman (Random Forest -RF) algoritması, 2001 yılında Leo Breiman [66] tarafından geliştirilen bir denetimli makine öğrenmesi algoritmasıdır. Rastgele orman algoritmasında çok sayıda karar ağacının bir araya getirilmesi ile bir orman oluşturulur ve ardından tüm ağaçlardan gelen bilgiler birleştirilir. Ormandaki ağaçların birbirlerinden etkilenmeden yetiştirilirler ve ağaçlar karar ağaçlarındaki esaslara dayanılarak yetiştirilmektedir. Karar ağaçları, CART yöntemleri ile oluşturulmakta ve düğümlerdeki en iyi bölünmenin sağlanması için GINI indeksi kullanılmaktadır. Ormandaki her bir ağacın kararı bir oy olarak değerlendirilmektedir. Sınıflandırma problemleri için en çok oyu alan sınıf topluluğun kararı olarak belirlenirken, regresyon problemleri için kararların ortalaması topluluğun kararı olur. Rastgele orman algoritması Rastgele Alt Uzay (Random Subspace) ve Torbalama (Bagging) yöntemlerinin bir araya getirilmesi ile oluşturulmuştur ve karar ağaçlarından farklı olarak rastgele orman algoritmasında budama işlemi yapılmamaktadır. Hem kategorik hem de sürekli veri setleri için

kullanılabilen rastgele orman algoritması büyük veri setlerini de başarılı şekilde işleyebilmektedir [66].

3.2.3. Destek vektör makinesi

Destek vektör makinesi (Support Vector Machine – SVM) Cortes & Vapnik tarafından geliştirilen ve hem sınıflandırma hem de regresyon problemlerinde kullanılabilen bir denetimli makine öğrenmesi algoritmasıdır [67]. SVM hem iki sınıflı hem de üç sınıflı sınıflandırma problemlerinin çözümünde kullanılabilmektedir [68]. SVM farklı sınıfların sınırında yer alan destek vektörleri arasındaki uzaklığı maksimize eden ve destek vektörlerin ortasından geçip sınıfları birbirinden en uygun şekilde ayıran hiper düzleminin bulunmasını amaçlar [69].

SVM’ de veriler ayrılma durumuna göre doğrusal ve doğrusal olmayan olarak iki başlıkta incelenmektedir. Doğrusal SVM, veri setinin hatasız doğrusal ayrılma ve belirli bir hata ile doğrusal ayrılma durumuna göre ikiye ayrılmaktadır. Hatasız ayrılma durumu sert sınır (Hard Margin) olarak adlandırılırken, belirli bir hata ile ayrılma durumu gevşek sınır (Soft Margin) olarak adlandırılmaktadır. Doğrusal olmayan SVM algoritmasında, doğrusal olmayan haritalama yöntemi kullanılarak iki boyutlu uzayda doğrusal olarak ayrılamayan veri seti, üç boyutlu özellik uzayına taşınarak doğrusal olarak ayrılabilir duruma getirilmektedir [69]. Gerçek yaşam problemlerinin karmaşıklığından dolayı doğrusal olarak ayırmak genellikle mümkün olmamaktadır. Borsa verileri de doğrusal olmayan ve dinamik verilerdir bu nedenle hisse senedi tahminlerinde SVM algoritması yaygın olarak kullanılmaktadır [5].

SVM algoritmaları verileri çok boyutlu düzlemlerde haritalayabilmek için Çekirdek (Kernel) fonksiyonlarından yararlanmaktadır ve bu haritalama sayesinde sınıflar arasında ayrılma düzlemi oluşturulabilmektedir. SVM algoritmalarında haritalamanın gerçekleştirilmesi için farklı Çekirdek fonksiyonları kullanılabilmektedir ve doğru Çekirdek fonksiyonunun seçilmesi büyük öneme sahiptir [70].

3.2.4. Lojistik regresyon

Lojistik regresyon (Logistic Regression – LR) (Logit) algoritması, iki veya ikiden fazla kategoriye sahip bağımlı değişkenin, bağımsız değişken ile aralarındaki sebep sonuç ilişkisini belirlemeye yarayan denetimli bir makine öğrenmesi algoritmasıdır. Probit’e alternatif olarak Logit kavramı 1944 yılında J. Berkson [71] tarafından ortaya

konulmuştur. David Cox tarafından ikili lojistik regresyon üzerine 1958 [72] yılında yayınlanan "The regression analysis of binary sequences" isimli çalışma ve ardından çoklu lojistik regresyon üzerine 1966 [73] yılında yayınlanan "Some procedures connected with the logistic qualitative response curve" isimli çalışma ile lojistik regresyon ile alakalı çalışmalar hız kazanmıştır. Lojistik regresyon yöntemi, bağımlı değişkenin beklenen olasılık değerini hesaplayıp bu olasılık değerini sınıflandırma ve atama yapmak için kullanılmaktadır. Lojistik regresyon algoritması bağımlı değişkenin veri tipine bağlı olarak üçe ayrılmaktadır[74]. Bunlar;

- İkili (Binary) lojistik regresyon: Bağımlı değişkenin iki kategorili olduğu durumda kullanılır.
- Çoklu (Multinomial) lojistik regresyon: Bağımlı değişkenin ikiden fazla kategoriye sahip olduğu ve bu kategorilerin sırasız olması halinde kullanılır.
- Sıralı (Ordinal) lojistik regresyon: Bağımlı değişkenin ikiden fazla kategoriye sahip olduğu ve bu kategorilerin sıralı olması halinde kullanılır.

Olasılık oranı (Odds ratio) (OR), lojistik regresyonun temelini oluşturmaktadır. Olasılık oranı Denklem 2.2.'de görüldüğü gibi, bir olayın gerçekleşme olasılığının($P(x)$), gerçekleşmeme olasılığına oranı($1-P(x)$) ile hesaplanmaktadır. K tane bağımsız değişken için doğrusal regresyon modeli Denklem 3.3'teki gibi yazılmaktadır. Regresyon modelinde β_0 bağımsız değişkenler 0 değerini aldığı anda bağımlı değişkenin değerini ifade ederken, $\beta_1, \beta_2, \dots, \beta_k$ bağımsız değişkenlerin regresyon katsayılarını, ε ise gözlemdeki Rastgele hata oranını göstermektedir.

$$Odds\ Ratio = \frac{P(x)}{1 - P(x)} \quad (3.2)$$

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon \quad (3.3)$$

Olasılık oranının doğal logaritması alınarak lojistik regresyon modeli elde edilmektedir. Olasılık oranının hesaplanması Denklem 2.4.'de görülmektedir.

$$\ln\left(\frac{P(x)}{1 - P(x)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon \quad (3.4)$$

Modelin parametrelerinin tahmininde maksimum olabilirlik kestirimi (Maximum likelihood Estimation) (MLE) yöntemi sıklıkla kullanılır. Parametrelerin belirlenmesi ise Denklem 2.5'te görülen lojistik regresyon denklemi elde edilmiş olur.

$$\text{Logit} = \ln\left(\frac{P(x)}{1 - P(x)}\right) = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k \quad (3.5)$$

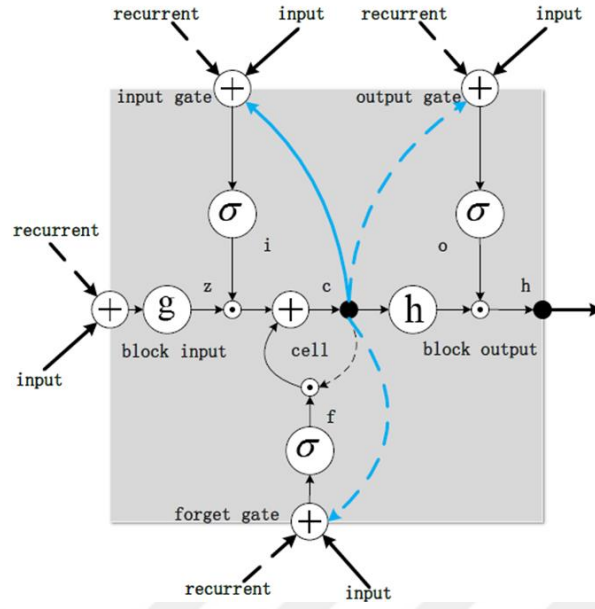
Buradan denklemden yola çıkarak incelenen olayın gerçekleştirme olasılığı $P(x)$ Denklem 2.6.'daki gibi hesaplanır [75].

$$P(x) = \frac{e^{(b_0+b_1X_1+b_2X_2+\dots+b_kX_k)}}{1 + e^{(b_0+b_1X_1+b_2X_2+\dots+b_kX_k)}} \quad (3.6)$$

3.2.5. Uzun kısa vadeli hafıza

Uzun Kısa Vadeli Hafıza (Long Short Term Memory- LSTM) ağları, Yinelemeli Sinir Ağının (Recurrent Neural Networks- RNN) uzun vadeli bağları öğrenebilen özel bir türüdür. LSTM, Hochreiter ve Schmidhuber tarafından 1997 yılında tanıtılmıştır [76]. LSTM ağlarının temel davranışı, bilginin uzun süreler boyunca ağda hatırlanmasıdır. RNN ağlarının yaşadığı Kaybolan Gradyan Sorunu, LSTM ağları ile çözülmüştür. LSTM ağlarının RNN ağlarından farkı, düğümlerinin farklı yapıdan oluşmasıdır [77].

Bir LSTM bloğu, girdi (input gate), çıktı (output gate) ve unut kapıları (forget gate) olmak üzere 3 farklı kapı içermektedir. LSTM, bu 3 kapıyı kullanarak hücreye bilgi ekleme ya da çıkarma yeteneğine sahiptir ve bu kapıları kullanarak durumunu kontrol etmektedir. Sigmoid katman 0-1 aralığında çıktı verir ve bileşenlerin ne kadarına izin verileceğini belirler. 0 değeri hiçbir şeyin geçmesine izin vermezken, 1 değeri her şeyin geçmesine izin verileceği anlamına gelir. Giriş kapısının rolü, bellek hücresinin durumunu etkileyebilecek olan giriş sinyalini düzenlemek ya da engellemektir. Çıkış kapısının rolü, bellek hücresinin durumunun diğer birimler üzerinde etkili olmasına izin vermek ya da bunu engellemektir. Unutma kapısının rolü ise hücrenin kendi kendisine bağlanmasını sağlayarak önceki durumunu hatırlamasını sağlamak ya da unutmasına izin vermektir. Bu kapılar sayesinde LSTM, RNN'den daha fazla bağlamsal bilgiye erişebilir ve böylece kaybolan gradyan sorununu hafifletir [29].



Şekil 3.1. LSTM hücre yapısı [29]

Şekil 3.1.'de t . zamanındaki LSTM hafıza hücresini görülmektedir. LSTM'in ilk adımı hücreden hangi bilginin atılacağına karar vermektir. Bu karar sigmoid fonksiyonu içeren unutma kapısı katmanında (f_t) gerçekleşmektedir. Unutma kapısı f_t , $t-1$. zaman dilimindeki çıkış vektörüne ve t . zaman dilimindeki giriş vektörüne bakarak 0-1 arasında bir değer verir. 1 değeri bilgiyi tamamen saklamayı temsil ederken, 0 değeri bilgiden tamamen kurtulmayı ifade eder. Bir sonraki adım, hücre durumuna hangi yeni bilginin depolanacağına karar vermektir. Bu iki bölümden oluşur. İlk olarak, giriş kapısı katmanı (i_t) olarak adlandırılan, sigmoid katman hangi değerlerin güncelleneceğine veya güncellenmeyeceğine karar verir. Sonraki adımda, hücre durumu güncellenir. Son olarak, hücrenin hangi bölümlerinin çıktı olarak verileceğini veya unutulacağını kararlaştıran bir çıkış kapısı katmanı (o_t) çalıştırılır. Kararlaştırılan parçalar çıktı olarak verilir [77].

3.3. Duygu Analizi

Duygu analiz yöntemlerinin başlangıcı 1960'lı yıllardaki metin sınıflandırma çalışmalarına dayanmaktadır [78]. 1990'lı yıllarda gelişen teknoloji ile bilgi sistemlerinin işlem kapasitelerinin artması duygu analizine olan ilgiyi arttırmıştır [79]. Sosyal medya platformlarının kullanımının her geçen gün artış göstermesiyle bu platformlar duygu analizleri için büyük veri kaynakları haline gelmişlerdir. Erişilebilir veri kaynaklarındaki

artış duygu analizine büyük önem kazandırmış ve çalışmaların hız kazanmasını sağlamıştır [80].

Duygu Analizi, metinde kullanılan ifadelerin ve kelimelerin özelliklerinden yola çıkarak metnin duygu içeriğini belirlemeye çalışmaktadır. Duygu analizi, metindeki duyguların anlatım biçimlerini ve bu anlatımlardaki olumlu, olumsuz veya tarafsız durumları belirlemeye çalışan analiz yöntemleridir [81]. İnsanlar paylaşımda bulunduklarında 2 türde ifade kullanılmaktadırlar. Bunlar gerçekleri içeren ifadeler ve görüşlerini bildiren ifadelerdir. Gerçekler nesnel ifadelerden oluşurken, görüşler öznel ifadelerden oluşmaktadırlar. Görüşler anlamsal açıdan daha kapsamlı olmakta ve olumlu, olumsuz veya tarafsız ifadeler içerebilmektedir [82].

Duygu analizi yöntemleri sözlük temelli yaklaşımlar ve makine öğrenmesi temelli yaklaşımlar olmak üzere 2 gruba ayrılmaktadır [83]. Sözlük temelli yaklaşımda öncelikle kelimeler veya kelime grupları içerdikleri duyguya (pozitif, negatif, nötr) göre sınıflandırılır, ardından bu kelimeler ve sınıfları kullanılarak duygu sözlükleri oluşturulur. Oluşturulan bu sözlükler kullanılarak metinlerin duygu durumlarının tespit edilmesi sağlanmaktadır. Metinlerin içerdikleri pozitif ve negatif kelimeler veya kelime gruplarının sayıları belirlenmekte, yüksek sayıda olan duygu ortaya çıkarılmaktadır. Metin sınıflandırılırken, daha çok pozitif kelime içeriyorsa duygusu pozitif, eğer negatif kelime içeriyorsa duygusu negatif olmaktadır [84].

Makine öğrenmesi temelli yaklaşımlarında, her kelimenin birer özellik olarak Karar Ağacı, Destek Vektör Makineleri, Yapay Sinir Ağları ve Naive Bayes gibi sınıflandırma algoritmalarında girdi olarak kullanılır ve metinler sınıflandırılarak duygu tahminleri gerçekleştirilir [85]. Metinler sınıflandırma modellerine girdi olarak kullanılmadan önce metin ön işleme aşamalarından geçirilerek dilbilimsel açıdan anlamsız olan terimlerin ve sembollerin temizlenmesi gerekmektedir. Temizlenmiş veri setindeki kelimeler ve kelime gruplarının terim frekans- ters doküman frekansı yöntemi gibi vektörizasyon yöntemleri ile vektör haline getirilmesi gerekmektedir. Ardından denetimli öğrenme modelleri için veri setinin bir kısmının etiketlenerek eğitim seti oluşturulur ve sınıflandırma algoritmaları uygulanarak metinlerin duygu durumları belirlenir.

4. VERİLER

Bu çalışma kapsamında NASDAQ borsasında yer alan ve Most Active by dolar Volume listesinde yer alan firmalar incelenmiş ve 5 adet teknoloji firması seçilmiştir. Seçilen bu firmalar Tesla, Apple, Microsoft, Nvidia ve Netflix'tir. Yapılacak çalışma kapsamında finansal hisse senedi verilerine ve sosyal medya verilerine ihtiyaç duyulmaktadır. Sosyal medya verilerinin günümüzde en çok kullanılan platformlardan biri olan X platformu aracılığıyla toplanmıştır. Hisse senedi verileri ise finansal kaynaklardan elde edilmiştir. Toplanan hisse senedi verileri ve sosyal medya verileri yeni öznitelikler oluşturmak üzere kullanılmıştır. Bu bölümde sosyal medya verileri, hisse senedi verileri ve bu veriler kullanılarak elde edilen öznitelikler açıklanmaktadır.

4.1. Sosyal Medya Verileri

Bu çalışma kapsamında gerekli olan sosyal medya verilerinin toplanması için günümüzde en çok kullanılan sosyal medya platformlarından biri olan X tercih edilmiştir. Mart 2006'da San Francisco'da Twitter adıyla kurulan X platformu, 27 Ekim 2022'de iş adamı Elon Musk tarafından satın alınmıştır ve 17 Mayıs 2024 tarihinde Twitter adı X olarak değiştirilmiştir. X'in sahip olduğu popülerite, platformu akademik çalışmalar için de önemli bir veri kaynağı haline getirmektedir.

Gönderi, X kullanıcılarının paylaştıkları durum mesajlarıdır. Bu paylaşımlar metnin yanı sıra emoji, fotoğraf, video, e-posta adresi ve web adresleri de içerebilmektedir. X'deki kullanıcılar paylaşımlarında, başına "@" karakteri ekleyen diğer kullanıcılardan bahsedebilmektedirler. Bir konu başlığından bahsetmek istediklerinde ise paylaşmak istedikleri konu başlığının başına "#" karakterini ekleyebilmektedirler. X kullanıcıları diğer sosyal medya platformlarından farklı olarak paylaşımlarında hisse senedinden bahsetmek istediklerinde başına "\$" karakterini ekleyerek istedikleri hisse senedinden bahsedebilmektedirler (cashtag).

X, sahip olduğu veri kaynağından geliştiricilerin, araştırmacıların ve şirketlerin faydalanabilmelerini sağlamak amacıyla X Uygulama Programlama Arayüzü (X API) hizmetini sunmuştur. X API hizmetinden yararlanabilmek için X geliştirici hesabına sahip olunması gerekmektedir. X geliştirici hesabı sahiplerine verilen kullanıcı adı ve şifreler sayesinde X API hizmetine bağlanılarak geçmiş gönderileri arama işlemleri gerçekleştirilebilmektedir.

Bu çalışma kapsamında X API arayüzü aracılığıyla hisse senedine ait cashtag kodunun kullanıldığı gönderiler, şirket ile bağlantılı hesaplarından bahsedilen gönderiler (mention), şirket ile bağlantılı hashtag kullanılarak paylaşılan gönderiler, şirketin kendi hesabından atmış olduğu gönderiler ve bu gönderilere verilen cevaplar toplanarak hisse senedine ait sosyal medya veri seti oluşturulmaktadır.

Bu çalışma kapsamında 5 farklı şirket için X API hizmeti kullanılarak 01.01.2020 – 14.03.2022 tarihleri arasındaki sadece İngilizce dilindeki X paylaşımları 05.04.2022 tarihinde toplanmıştır. X API hizmeti kullanılarak hisse senedine ait paylaşımlar \$ cashtag kullanılarak, şirket hakkında yapılan paylaşımlar @ mention ve # hashtag kullanılarak toplanmıştır. Toplanan gönderilerin sorgulara göre dağılımı Tablo 4.1.'de verilmiştir. Toplanan tüm veriler kullanılarak 1. veri seti hazırlanırken ayrıca sadece cashtag ile toplanan verilerden 2. veri seti hazırlanmıştır.

Tablo 4.1. Şirketler için toplanan gönderi sayıları

Tesla	Gönderi Sayısı	Microsoft	Gönderi Sayısı	Apple	Gönderi Sayısı	Netflix	Gönderi Sayısı	Nvidia	Gönderi Sayısı
\$TSLA	2,229,597	\$MSFT	419,806	\$AAPL	999,135	\$NFLX	340,524	\$NVDA	338,679
@Tesla	2,499,239	@Microsoft	1,208,756	@Apple	2,285,927	@Netflix	3,477,654	@Nvidia	149,348
#Tesla	899,833	#Microsoft	927,701	#Apple	961,869	#Netflix	1,425,907	#Nvidia	135,665
Tesla Hesabı	229,074	Microsoft Hesabı	35,919	Apple Hesabı	4,741,466	Netflix Hesabı	51,437	Nvidia Hesabı	439
Toplam	5,857,743	Toplam	2,592,182	Toplam	8,988,397	Toplam	5,295,522	Toplam	624,131

Toplanan gönderi verilerinin zaman dilimleri ile NASDAQ borsasının zaman dilimlerinin farklı olmaları nedeniyle X verilerinin zaman dilimleri NASDAQ borsasının zaman dilimiyle eşitlenmiştir. X veri setindeki NASDAQ borsasının kapalı olduğu gün ve zamanlardaki gönderiler çıkarılmıştır. Bu işlem sonrasında tüm gönderileri içeren veri seti için 4,848,509 olan Gönderi sayısı 1,616,761 adete, sadece cashtag ile paylaşılan gönderileri içeren veri seti için 1,817,368 olan Gönderi sayısı 775,450 adete azaltılmıştır.

X API V2.0 arayüzü kullanılarak yapılan sorgularda elde edilen bilgiler 3 ana grupta toplanabilir.

1. Paylaşılan gönderinin kendisi
2. Yazar istatistikleri
 - a. Yazarın kullanıcı adı
 - b. Yazarı takip eden hesap sayısı
 - c. Yazarın takip ettiği hesap sayısı

- d. Yazarın paylaştığı gönderi sayısı
 - e. Yazarın hesap onay durumu
3. Gönderi istatistikleri
- a. Paylaşılan gönderinin tekrar paylaşılma sayısı
 - b. Paylaşılan gönderinin aldığı beğeni sayısı
 - c. Paylaşılan gönderinin aldığı cevap sayısı
 - d. Paylaşılan gönderinin aldığı alıntı sayısı

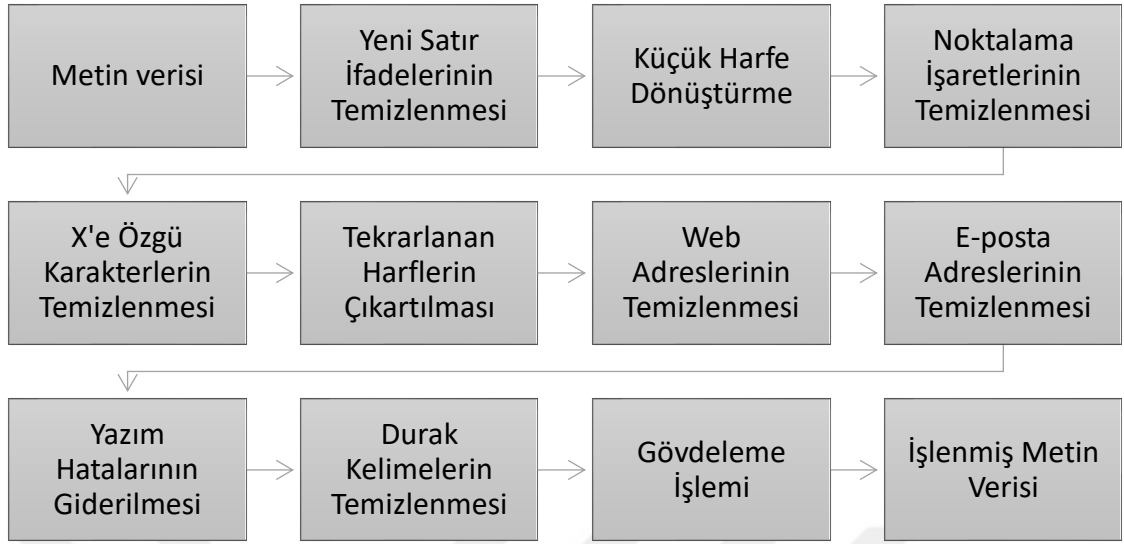
Her bir gönderi için toplanan bu veriler kullanılarak çalışma kapsamında analizlere dahil edilecek öznitelikler oluşturulmuştur.

4.2. Sosyal Medya Verilerinden Öznitelik Oluşturma

Sosyal medya verilerinden elde edilebilecek birçok öznitelik bulunmaktadır. Metin verilerinden özniteliklerin oluşturulabilmesi için öncelikle metin ön işleme süreçlerinin tamamlanması gerekmektedir. Metin ön işleme süreçlerinin tamamlanmasının ardından metin vektörel olarak ifade edilebilir hale gelmektedir. Metin ön işleme süreçlerini tamamlayan sosyal medya verileri duygu analizinde kullanılarak paylaşımların duygu değerleri elde edilebilmektedir. Sosyal medya paylaşımlarına ait istatistiksel veriler kullanılarak ise yazar etki değeri, post etki değeri ve paylaşım yapan yazar sayısı gibi öznitelikler hesaplanabilmektedir.

4.2.1. Metin ön işleme

Duygu analizi (Sentiment Analysis) sürecine başlamadan önce sosyal medya verilerinin metin ön işleme sürecinden geçirilmesi gerekmektedir. Metin ön işleme süreci, metinde bulunan fakat anlam ifade etmeyen karakter, noktalama işareti ve sembollerin metinden temizlenmesi sürecidir. Gönderiler dilbilimsel olarak anlamsız sembol ve kelimeleri yoğun olarak içermesiyle kitap, dergi ve makale gibi metinlerden ayrılmaktadır. Dilbilimsel açıdan anlamı olmayan karakter, işaret ve sembollerin temizlenmesi duygu analizi için büyük önem arz etmektedir. Çalışma kapsamında gerçekleştirilen metin ön işleme süreçleri Şekil 4.1.'de gösterilmiştir.



Şekil 4.1. Metin ön işleme akış şeması

Yeni Satır İfadelerinin Temizlenmesi: Duygu analizi süreçlerine başlamadan önce dilbilimsel açıdan anlam ifade etmeyen yeni satır ifadelerinin temizlenmesi gerekmektedir. Bazı kullanıcılar, paylaştıkları gönderileri başka kullanıcıların dikkatini çekmek ya da içeriklerini listelemek amacıyla birden fazla satırda yazmaktadırlar. X API ile toplanan gönderilerde birden fazla satır olması halinde terimler arasında “/n” sembolü görülmektedir ve bunların temizlenmesi gerekmektedir.

Küçük Harfe Dönüştürme: Analizler sırasında aynı kelimenin büyük ve küçük harflerinin algoritma tarafından 2 farklı kelime gibi algılanmasının önüne geçmek için büyük kelimeler küçük harfli hale dönüştürülmesi gerekmektedir.

Noktalama İşaretlerinin Temizlenmesi: Dilbilimsel açıdan anlam ifade etmeyen noktalama işaretlerinin de duygu analiz süreçlerine başlamadan önce temizlenmesi gerekmektedir. Gönderi içerisinde anlam bütünlüğünü sağlamak için “,”, soru ifade etmek için “?” ve cümle sonunu ifade etmek için “.” gibi noktalama işaretlerine sıklıkla başvurulmaktadır.

X'e Özgü Karakterlerin Temizlenmesi: X'e özgü karakter ve ifadelerinde dilbilimsel açıdan bir ifade etmemesi nedeniyle temizlenmesi gerekmektedir. X'e özgü karakterlere örnek olarak hashtag (#), cashtag (\$) ve mention (@) verilebilir. Kullanıcılar spesifik bir konu hakkında konuşmak istediklerinde hashtag sembolünü, bir hisse senedi hakkında

konuşmak istediklerinde cashtag sembolünü ve bir başka kullanıcı hakkında konuşmak istediklerinde mention sembolünü sıklıkla kullanmaktadır.

Tekrarlanan harflerin çıkarılması: Sosyal medya kullanıcılarının herhangi bir duyguyu güçlü bir şekilde ifade etmek istediklerinde gönderilerindeki kelimelerde bazı harfleri tekrarlayarak yazmaları sık karşılaşılan bir durumdur. Metnin vektörel ifade edilmesi sırasında harfleri tekrarlanmış kelimeler farklı kelimeler olarak analize katılacaklarından bu kelimelerdeki tekrarlanan harflerin tespit edilerek düzeltilmesi gerekmedir.

Web adreslerinin temizlenmesi: Kullanıcılar X üzerinden yaptıkları paylaşımlarda metin içeriğinin yanı sıra fotoğraf, video ve web adresleri de paylaşabilmektedirler. Yapılan bu paylaşımlar X API ile toplanan verilerde metin içeriğine dahil edildiğinden metin içerisinde dilbilimsel olarak anlam ifade etmeyen bu içeriklerin duygu analizine başlamadan önce temizlenmesi gerekmektedir.

E-posta adreslerinin temizlenmesi: Kullanıcılar X üzerinden yaptıkları paylaşımlarda e-posta adresleri de paylaşabilmektedirler. Duygu analizine başlamadan önce dilbilimsel açıdan anlam ifade etmeyen bu içeriğin temizlenmesi gerekmektedir.

Yazım Hatalarının Düzeltilmesi: Sosyal medyada sıklıkla karşılaşılan yazım hatalarının duygu analizine başlamadan önce düzeltilmeye çalışılması gerekmektedir. X kullanıcıları gerek karakter sınırından kaynaklı kelimelerin kısaltılmasından gerekse sosyal medya paylaşımlarında resmi dil kullanılmamasından kaynaklı olarak birçok yazım hatası yapmaktadırlar. Bir gönderi içerisinde duygu belirlen bir kelimede meydana gelebilecek bir hata tüm gönderinin duygu değerinin yanlış hesaplanmasına yol açacağından duygu analizine başlamadan önce bu yazım hatalarının mümkün olduğunda düzenlenmesi gerekmektedir.

Durak Kelimelerin Temizlenmesi: Durak kelimelerin (Stop Words) çıkarılması işlemi, tek başına anlamı olmayan kelimelerin temizlenmesi işlemidir. Durak kelimeler; edat, bağlaç, zamir gibi tek başına anlamlı olmayan, ancak yazım dilinde çok sık kullanılan kelimelerdir. İngilizce dili için bu kelimelere örnek olarak “to”, “be”, “do”, “the”, “and” ve “is” gibi çok sık kullanılan kelimeler verilebilir. Durak kelimelerin çıkarılması duygu analizinin veriminin artırılmasında büyük öneme sahiptir.

Gövdeleme (Lemmatization) işlemi: Gövdeleme, kelimelerin basit hallerine çevrilmesi işlemidir. Kelimedeki çekim ve yapım ekleri atılması ile kelimenin kök formuna ulaşmak istenmektedir. Bu işlemde fiiller mastar hallerine çevrilir ya da çoğul kelimeler tekil

duruma dönüştürülür. Kelimelerinin kök hallerinin bilinmesi ile kelimelerin biçimsel olarak benzer halleri bulunarak kök hallerine çevrilebilir. Böylece benzer kelimelerin metin vektörüne dahil edilmemesi sayesinde duygu analiz işlemlerinin verimi artırılmış olur.

4.2.2. Metnin vektörel olarak ifade edilmesi

Metin önileme süreci sonucunda elde edilen gönderi verilerinin analizlere dahil edilebilmesi için sayısal vektör haline dönüştürülmesi gerekmektedir. Metin verilerinin vektör haline getirilmesi için en sık kullanılan yöntemlerden biri TF-IDF (Term Frequency- Inverse Document Frequency) yöntemidir. TF-IDF terim frekansı ve ters doküman frekansı olmak üzere iki istatistiğin çarpımı ile hesaplanmaktadır.

Terim Frekansı (Term Frequency-TF) kavramı ilk olarak H. P. Luhn [86] tarafından 1957 yılında önerilmiştir. TF, bir terimin belgedeki görülme sıklığıdır. Bir belgedeki diğerlerine göre daha sık görülen terimin öneminin daha yüksek olacağı esasına dayanmaktadır [87]. Denklem 4.1’de görüldüğü gibi terim frekansı bir terimin belgede görülme sayısının, belgedeki toplam kelime sayısına bölünmesi ile hesaplanmaktadır.

$$TF(t) = \frac{t \text{ teriminin belgede kaç kez geçtiği}}{\text{Belgedeki toplam kelime sayısı}} \quad (4.1)$$

Ters Doküman Frekansı (Inverse Document Frequency-IDF) kavramı ilk olarak Karen S. Jones [88] tarafından 1972 yılında "Term Specificity" olarak önerilmiştir. Ters Doküman Frekansı, Terim Frekansından farklı olarak bir belgede az sayıda bulunan terimin ayırt edici özelliğe sahip olacağı esasına dayanan ve terimin önemini gösteren bir ölçüttür [89]. Terim frekansı (TF) hesaplanırken her terim eşit öneme sahiptir fakat bazı sık görülen terimlerin çok az öneme sahip olduğu bilinmektedir. Her dilde bu tarz düşük öneme sahip terimler bulunmaktadır. İngilizce dilinde sık kullanılan “is”, “and” “of” ve “that” gibi kelimeler bu duruma örnektir. Ters Doküman Frekansı değeri Denklem 4.2.’de görüldüğü gibi toplam doküman sayısının terimi içeren doküman sayısına bölümünün logaritması ile hesaplanmaktadır.

$$IDF(t) = \log \frac{\text{Toplam Döküman Sayısı}}{\text{Terimi İçeren Döküman Sayısı}} \quad (4.2)$$

Ters doküman frekansında metinde az geçen terimler için yüksek, metinde çok geçen kelimeler için düşük değer vermektedir. TF-IDF hesaplanırken Terim Frekansı,

Ters Doküman Frekansı ile ağırlıklandırılarak metinlerde çok sıklıkla kullanılan kelimelerin etkisinin azaltılması sağlanmış olmaktadır. Denklem 4.3.'de görüldüğü gibi i. kelimenin j. dokümandaki TF-IDF değeri olan x_{ij} , TF ve IDF değerlerinin çarpımı ile hesaplanmaktadır.

$$x_{ij} = TF_{ij} * IDF_j \quad (4.3)$$

4.2.3. Emoji analizi

Emojiler, insanların duygularını daha kolay ve otomatik bir şekilde ifade etmesini sağlayan görsel sembollerdir [90]. Emoji kelimesi, Japonca'da “resim” ve “karakter” kelimelerinin birleşiminden oluşmaktadır. Akıllı cihazlardaki klavyelere emoji desteğinin gelmesiyle birlikte anlık mesajlaşmadaki kullanımı her geçen gün artış göstermiştir. Emojilerin en önemli işlevleri hem dilden bağımsız olarak insanların iletişime geçmesini kolaylaştırması hem de iletilmek istenen mesajın duygusal yönünü kuvvetlendirmesidir. Emoji, duygusal pazarlama platformu tarafından yayınlanan rapora göre, çevrimiçi kullanıcılarının %92'si, duygularını ifade ederken emojilerden faydalanmaktadır (<http://5>).

Emojiler, gönderi içerisinde Unicode formatında kodlanmaktadır ve bu emoji kodları 2009 yılında standartlaşmıştır [91]. Kralj Novak ve ark. [92, 93] Emojilerin birçoğunu için [-1,1] aralığında duygu durum değeri belirlemişlerdir.

4.2.4. Paylaşım yapan yazar sayısı

X API kullanılarak toplanan gönderiler Borsanın açık olduğu zamanlar için 5 dakika, 15 dakika, 30 dakika, 60 dakika ve 120 dakikalık zaman periyodlarına bölünmüştür. SPAM paylaşım yaparak belirlenen zaman periyodu içerisinde hisse senedinden bağımsız şekilde periyod içerisindeki gönderi sayılarını arttıran hesaplar bulunmaktadır. Bu çalışma kapsamında SPAM paylaşım yapan hesapların analiz üzerindeki etkisini ortadan kaldırmak amacıyla belirlenen zaman periyodları içerisinde paylaşılan toplam gönderi sayısı yerine paylaşım yapan toplam hesap sayısı belirlenmiş ve analize yazar sayısı değişkeni olarak dahil edilmiştir.

4.2.5. Yazar ve post etki deęerleri

Bu alıřma kapsamında X API kullanılarak her bir Gnderi iin elde edilen yazar istatistik veriler ve Gnderi istatistik verileri yazar etki deęeri ve post etki deęeri olmak zere 2 deęer haline dnřtrlmřtr.

Yazar etki deęeri, yazarın paylařımlarını ulařtırabildięi kiři sayısını yani etki alanını ifade etmektedir. Yazar etkileme deęeri, yazarın takipi sayısından takip ettięi kiři sayısının ıkarılması ile elde edilmektedir. Takip eden kiři sayısından, takip edilen kiři sayısı ıkarılarak, ok sayıda kiřiye takip ederek takip eden kiři sayısını arttırmayı hedefleyen SPAM paylařım hesaplarının deęerinin azaltılmasını hedeflenmektedir.

Post etki deęeri, paylařılan gnderinin almıř olduęu etkileřim ve paylařım sayılarını gz nne alarak etkiledięi kiři sayısını yani etki alanını ifade etmektedir. Post etkileme deęeri, paylařılan gnderinin repost sayısı, beęeni sayısı, cevap sayısı ve alıntı sayısının toplamı ile elde edilmektedir.

4.2.6. Metin ve emoji duygu durum deęerleri

Bu alıřma kapsamında metinlerin duygu durum deęerlerinin elde edilmesi iin kullanılabilecek duygu analizi yntemleri arařtırılmıřtır. Yapılan arařtırmalar sonucunda duygu analizlerinin gerekleřtirilmesi iin Python programlama dilindeki VADER [94], Stanza [95] ve Textblob ktphanelerinde bulunan fonksiyonların kullanılmasına karar verilmiřtir. Metin niřleme sreci sonucunda elde edilen iřlenmiř gnderi verisi VADER, Stanza ve Textblob ktphanelerindeki fonksiyonlar kullanılarak duygu analizleri gerekleřtirilmiř ve olumlu, olumsuz ve tarafsız olarak 3 sınıflı olarak sınıflandırılmıřtır. Metinlerden ayrılan emojilerin duygu durum deęerlerinin analiz edilmesi iin Knař Novak ve ark. [93] tarafından hazırlanan Emoji Duygu Sıralaması 1.0 alıřmasındaki duygu durum deęerleri kullanılmıřtır. Gnderilerin ierinde yer alan her bir emoji Knař Novak ve ark. tarafından hazırlanan emoji duygu durum deęerleri ile eřleřtirilerek analiz edilmiřtir. Her bir gnderinin ierdięi emojilerin duygu durum deęerlerinin ortalamaları alınarak o gnderinin emoji duygu durum skoru elde edilmiřtir. Metin ve Emoji duygu durum deęerlerinin birleřtirilmesi bu iki deęerin ortalaması alınarak yapılmıřtır. Hesaplanan bu duygu durum deęerlerine paylařım yapan yazarın etkisinin ve gnderinin aldıęı etkileřimin etkisinin dahil edilebilmesi iin Tablo 4.2.'de grlen 6 farklı duygu durum deęeri hesaplanarak analize dahil edilmiřtir.

Tablo 4.2. *Duygu değer değişkenleri*

Duygu Değer No	Duygu değeri değişkenleri
Duygu Değeri 1	Duygu değerleri toplamı
Duygu Değeri 2	Duygu değerleri toplamı / Periyoddaki toplam gönderi sayısı
Duygu Değeri 3	(Yazar etki değeri x Duygu değeri) toplamı
Duygu Değeri 4	(Yazar etki değeri x Duygu değeri) toplamı/ Periyoddaki toplam gönderi sayısı
Duygu Değeri 5	(Post etki değeri x Duygu değeri) toplamı
Duygu Değeri 6	(Post etki değeri x Duygu değeri) toplamı/ Periyoddaki toplam gönderi sayısı

Duygu değeri 1, paylaşılan gönderilen duygu durum değerleri toplamıdır. Duygu değeri 2, paylaşılan gönderilerin duygu durum değerleri ortalamasını göstermektedir. Duygu değeri 3, paylaşılan gönderilerin duygu durum değerleri ile yazar etki değeri çarpımlarının toplamıdır. Duygu değeri 4, paylaşılan gönderilerin duygu durum değerleri ile yazar etki değeri çarpımlarının ortalamasıdır. Duygu değeri 5, paylaşılan gönderilerin duygu durum değerleri ile gönderi etki değeri çarpımının toplamıdır. Duygu değeri 6, paylaşılan gönderilerin duygu durum değerleri ile post etki değeri çarpımının ortalamasıdır.

4.3. Hisse Senedi Verileri

Bu çalışma kapsamında seçilen 5 firmaya ait 01.01.2020 – 14.03.2022 tarihleri arasındaki 5 dakika, 15 dakika, 30 dakika, 60 dakika ve 120 dakika frekanslara sahip hisse senedi verileri ücretsiz kaynaklardan temin edilmiştir. Fiyat verilerinde her bir hisse senedi için belirlenen frekanslarda 5 adet değer mevcuttur. Bunlar, belirlenen frekans aralığında en yüksek fiyat, en düşük fiyat, açılış fiyatı, kapanış fiyatı, işlem hacmidir. Toplanan bu veriler teknik göstergelerin hesaplanmasında kullanılmıştır.

4.4. Hisse Senedi Verilerinden Öznitelik Oluşturma

Hisse senedi verilerinden öznitelik çıkarmak için literatürde en sık başvuru alan yöntemin teknik göstergelerden yararlanmak olduğu görülmektedir. Teknik göstergeler, hisse senetlerinin en yüksek, en düşük, açılış ve kapanış fiyatlarının yanı sıra işlem hacmi gibi zaman içerisinde değişen verileri temel alan ve matematiksel ve istatistiksel hesaplamalar sonucunda elde edilen teknik araçlardır. Bu göstergeler, yatırımcıların ve analistlerin piyasa hareketlerini doğru anlamasına ve yorumlamasına olanak tanır [96]. Yatırımcılar ve analistler, farklı teknik göstergeleri birlikte kullanarak daha doğru

tahminler yapmaya çalışmaktadırlar. Teknik göstergelerin girdi olarak kullandıkları parametreler birbirinden farklılık göstermektedir. Bir gösterge tüm hisse senedi parametrelerini girdi olarak alıp hesaplanabilirken, başka bir gösterge sadece kapanış fiyatını esas alarak hesaplanabilmektedir. Teknik göstergeler, 4 temel grup altında sınıflandırılabilir;

1. Momentum göstergeleri
2. Trend göstergeleri
3. Volatilite (Oynaklık) göstergeleri
4. Hacim göstergeleri

Momentum temelli teknik göstergeler, belirli bir zaman periyodundaki fiyatın değişim hızını ölçmekte kullanılan göstergelerdir. Momentum temelli teknik göstergeler fiyatın zaman ile karşılaştırılması yoluyla fiyatın değişim hızını belirlemeye çalışmaktadır. Momentum temelli teknik göstergelere örnek olarak Relative Strength Index (RSI), Commodity channel Index (CCI), Williams %R ve Rate of Change (ROC) verilebilir.

Trend temelli göstergeler fiyat oluşumlarının kullanılması ise fiyatların aşağı yönlü yoksa yukarı yönlü mü hareket ettiğinin anlaşılmasını sağlarken aynı zamanda hareketin gücünü de ölçmektedir. Trend temelli teknik göstergeler örnek olarak Simple Moving Average (SMA), Exponential Moving Average (EMA) ve Moving Average Convergence Divergence (MACD) verilebilir.

Volatilite (Oynaklık) temelli teknik göstergeler, fiyatlardaki hareketlerin oranlarını ölçmektedir. Bu göstergeler genelde en düşük ve en yüksek fiyat verilerini temel alarak hesaplanırlar ve fiyatın yönünü dikkate almazlar. Volatilite temelli göstergelere örnek olarak Average True Range (ATR), Bollinger Bands ve Keltner Channels verilebilir.

Hacim temelli teknik göstergeler, trendin gücünü ölçmekte kullanılmaktadır. Hacim temelli teknik göstergelere örnek olarak On-balance volume (OBV), Money Flow Index (MFI) ve Ease of movement (EMV) verilebilir.

Bu çalışmada kapsamında 4 farklı teknik gösterge seçilerek analize dahil edilmiştir. Bunlar; Keltner Channel Percentage (KCP), Ease of Movement (EMV), Commodity channel index (CCI) ve Stochastic relative strenght index (Stoch RSI)'dir.

4.4.1. Keltner channel (KC) ve keltner channel percentage (KCP)

Keltner Channel (KC) (Keltner Kanalı), hisse senedinin trend yönünü ve aşırı alım-satım seviyelerini belirlemede kullanılan volatilité (fiyat oynaklığı) osilatörüdür. Keltner kanalları göstergesi, Chester Keltner [97] tarafından 1960 yılında ortaya atılmıştır. Keltner kanalları 3 ayrı hattan oluşmaktadır ve genel yapısı itibarıyla Bollinger bantlarına benzemektedir. Keltner kanalı indikatörü farklı olarak, Üstel Hareketli Ortalama (EMA) değerine ek olarak fiyat oynaklığını ölçmek için kullanılan Ortalama Gerçek Aralık (ATR) değerini de kullanmaktadır. ATR değerinin hesaplanması için öncelikle şu 3 değer hesaplanır; (En yüksek – En düşük), Mutlak (En yüksek – Bir önceki Kapanış Fiyatı) ve Mutlak (En düşük – Bir önceki kapanış). 10 periyod boyunca basit bu 3 değerden en büyük olanının hareketli ortalaması alınarak ATR değeri hesaplanmış olur. Bu çalışma kapsamında 20 periyodluk EMA değeri keltner kanalının temeli olan orta hattı oluşturmaktadır. EMA değerine ATR değerinin 2 katı eklenerek üst hat, 2 katı çıkarılarak ise alt hat elde edilmektedir. Bu 3 hattın değerleri kullanılarak Keltner Channel Percentage (KCP) değeri hesaplanabilmektedir. KCP değeri orta hat değerinden alt hat değerinin çıkarılıp üst hattan alt hattın çıkarılmasıyla elde eden değere bölünün 100 ile çarpılması ile elde edilmektedir.

$$KCP = \frac{(Orta\ Hat - Alt\ Hat)}{(\text{Üst}\ Hat - Alt\ Hat)} * 100 \quad (4.4)$$

4.4.2. Ease of movement (EMV)

Ease of Movement Value (EMV) (Hareket kolaylığı) fiyat ile hacim arasındaki ilişkinin kullanılmasıyla hisse senedinin ne kadar kolay hareket ettiğini göstermek için tasarlanan bir hacim osilatörüdür. EMV, Richard W. Arms Jr. [98] Tarafından 1990 yılında tanıtılmıştır. EMV göstergesi fiyat ile hacim arasındaki ilişkiyi göstermek üzere pozitif ve negatif değerler arasında dalgalanacak şekilde tasarlanmıştır. EMV göstergesi pozitif iken hisse senedinin yukarı yönlü hareket ettiği, negatif iken ise aşağı yönlü hareket ettiği anlamına gelmektedir. EMV göstergesinin mutlak değerinin yüksek olması, hisse senedinin büyük hacimle kolay hareket ettiğini gösterirken, düşük olması ise düşük hacimde zor hareket ettiğini göstermektedir. EMV göstergesinin hesaplanabilmesi için Denklem 4.5'te formül ile hesaplanan taşınan mesafe değerinin ve Denklem 4.6'da görülen formül ile hesaplanan kutu oranı değerlerinin elde edilmesi gerekmektedir. 1 tane

periyodu EMV değeri (Taşınan mesafe/Kutu oranı) şeklinde hesaplanmaktadır. EMV göstergesi bu 1 periyotluk EMV değerinin 14 periyod için hesaplanarak basit hareketli ortalamasının alınması sonucunda elde edilmektedir.

$$Taşınan\ mesafe = \frac{(En\ yüksek + En\ Düşük)}{2} - \frac{(Bir\ önceki\ en\ yüksek + Bir\ önceki\ en\ düşük)}{2} \quad (4.5)$$

$$Kutu\ oranı = \frac{\left(\frac{Hacim}{100,000,000}\right)}{En\ yüksek - En\ Düşük} \quad (4.6)$$

4.4.3. Commodity channel index (CCI)

Commodity channel index (CCI)(Emtia Kanal Endeksi), belirlenen periyottaki fiyatın ortalama fiyattan ne kadar saptığını gösteren bir Moment osilatörüdür ve Donald Lambert tarafından 1980’de “Commodities Magazine” dergisinde tanıtılmıştır [99]. CCI göstergesi genellikle -100 ve +100 arasında değer almaktadır fakat bu aralığın dışına da çıkabilmektedir. 100 ve üstü değerler aşırı alıma bağlı fiyat yükselişini gösterirken, -100 ve altındaki değerler aşırı satış ve fiyat düşüşünü göstermektedir. CCI göstergesinin hesaplanması için üç farklı değere ihtiyaç vardır. Tipik fiyat, hisse senedinin en yüksek, en düşük ve kapanış fiyatlarının ortalamasıdır. Bir diğer değer, tipik fiyatın belirlenen periyotta basit hareketli ortalamasının hesaplanması ile elde edilmektedir. Bu çalışma kapsamında periyod sayısı 20 olarak alınmıştır. Ortalama sapma değeri ise hesaplanan basit hareketli ortalamadan sapmayı ifade etmektedir. Tipik fiyattan basit hareketli ortalama değerinin çıkarılması ve çıkan değerın ortalama sapmanın 0.015 ile çarpımına bölünmesi ile CCI göstergesi hesaplanmaktadır.

$$CCI = \frac{Tipik\ Fiyat - BHO(Tipik\ Fiyat)}{0,015 * Ortalama\ sapma} \quad (4.7)$$

4.4.4. Stochastic relative strength index (Stoch RSI)

Stochastic Relative Strength Index (Stoch RSI) (Stokastik Bağlı Güç İndeksi), Relative Strength Index (RSI) göstergesinden türetilmiş bir göstergedir. Stokastik RSI (Stoch RSI) göstergesi Tushard Chande ve Stanley Kroll [100] tarafından 1994 yılında “The New Technical Trader” adlı kitapta tanıtılmıştır. Stoch RSI 0 ile 1 arasında

dalgalandan bir momentum osilatörüdür. Stoch RSI'in 0.8'in üzerinde olması aşırı alım olduğunu, 0.2'nin altında olması ise aşırı satım olduğunu göstermektedir.

Stoch RSI göstergesinin hesaplanabilmesi için öncelikle RSI göstergesinin hesaplanması gerekmektedir. RSI, Welles Wilder [101] tarafından geliştirilmiş ve 1978'de "New Concepts in Technical Trading System" adlı kitapta tanıtılmıştır. RSI, 0 ile 100 arasında değer almaktadır ve 30'un altında olması aşırı satış olduğunu, 70'in üstünde olması ise aşırı alım olduğunu göstermektedir. RSI değerinin hesaplanabilmesi için öncelikle RS değerinin hesaplanması gerekmektedir. RS değeri Denklem 4.8'de görülen formül ile hesaplanmaktadır. Bu çalışma kapsamında RS değeri hesaplanırken 14 periyod değerlendirilmiştir. RSI göstergesi ise Denklem 4.9'da görülen formül ile hesaplanmaktadır. Bulunan RSI değeri Stoch RSI göstergesinin hesaplanması için Denklem 4.10'da görülen formülde yerine konulur. Stoch RSI göstergesinin hesaplanmasında da zaman aralığı 14 periyod olarak belirlenmiştir.

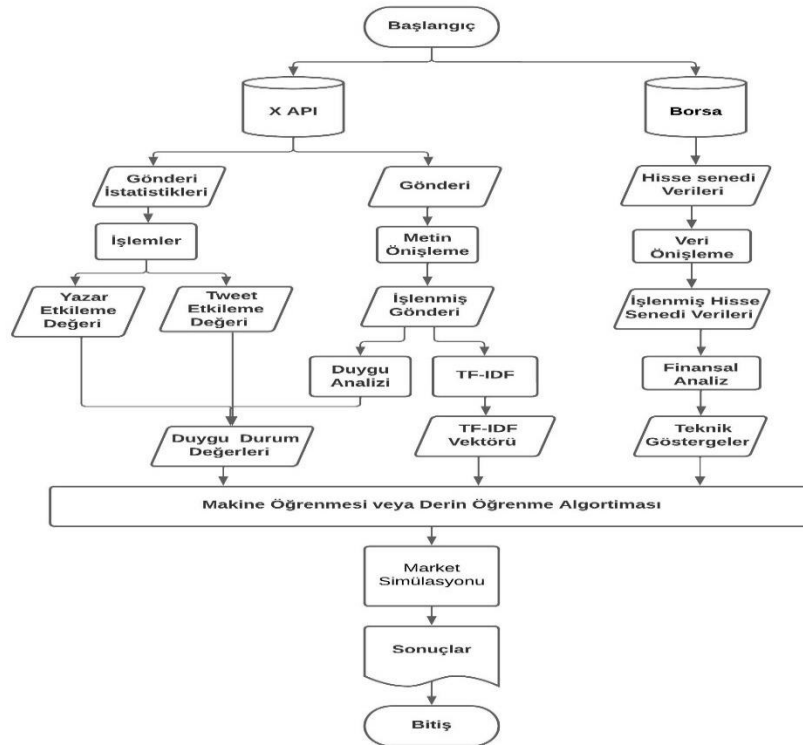
$$RS = \frac{t \text{ periyodda pozitif fiyat değişimlerinin ortalaması}}{t \text{ periyodda negatif fiyat değişimlerinin ortalaması}} \quad (4.8)$$

$$RSI = 100 - \frac{100}{1+RS} \quad (4.9)$$

$$Stoch RSI = \frac{Mevcut RSI - En Düşük RSI}{En Yüksek RSI - En Düşük RSI} \quad (4.10)$$

5. UYGULAMA

Bu çalışma kapsamında sosyal medya verileri ve teknik indikatörler kullanılarak hisse senetlerinin gelecek hareketleri tahmin edilmekte ve bu tahminler kullanılarak market simülasyonu oluşturulmaktadır. Bu çalışmanın akış şeması Şekil 5.1’de görüldüğü gibidir. X platformundan toplanan veriler gönderi istatistikleri ve metin verisi olmak üzere ikiye ayrılmaktadır. Gönderi istatistikleri kullanılarak yazar etki değeri ve post etki değeri hesaplanmaktadır. X platformundan toplanan gönderilere ait metin verilerine metin önileme adımları uygulanmakta ve hem TF-IDF yöntemi kullanılarak kelime vektörü elde edilmekte hem de duygu analiz yöntemleri kullanılarak duygu durumları bulunmaktadır. Bu duygu durum değeri, yazar etki değeri ve post etki değerleri ile birleştirilerek duygu durum değeri alternatifleri oluşturulmaktadır. Hisse senetlerine ait veriler ise ver önileme adımlarının ardından finansal analiz yöntemleri kullanılarak teknik indikatörlerin hesabında kullanılmaktadır. Hesaplanan tüm bu veriler makine öğrenmesi algoritmalarının kullanıldığı sınıflandırma analizleri ile hisse senedi hareketlerinin tahmini için kullanılmaktadır. Elde edilen tahmin sonuçları kullanılarak market simülasyonu oluşturulmaktadır.



Şekil 5.1. Algoritma akış şeması

Bu çalışmada analizlerin gerçekleştirilmesi için yapay zeka alanında en çok tercih edilen programlama dillerinin başında gelen Python programlama dili tercih edilmiştir. Yapay zeka ve makine öğrenmesi alanındaki akademik çalışmalarda Python dili çokça tercih edilmektedir. Bunun en önemli nedeni yapay zeka, makine öğrenmesi ve derin öğrenme alanında pek çok kütüphaneye sahip olması ve geliştirmelerin Python dili ile yapılmasıdır.

Python nesne yüksek seviyeli bir dildir, yorumsal ve modülerdir. Bir programlama dili; insanların kolay anlayabileceği bir haldeyse yüksek seviye, daha karmaşık bir haldeyse yani makine diline yaklaşırsa düşük seviye olarak anılmaktadır. Python programlama dili ise insan mantığına yakın olmasından dolayı yüksek seviye programlama dili olarak adlandırılmaktadır. Python geliştirme arayüzü olarak da PyCharm IDE kullanılmıştır.

Çalışmalarda Rastgele orman, Karar ağacı, Destek vektör makinesi ve Lojistik regresyon algoritmalarının analizlere dahil edilmesinde Scikit-Learn [102] Kütüphanesinden yararlanılmıştır. Scikit-Learn kütüphanesi makine öğrenmesi algoritmalarının uygulanmasında sıklıkla tercih edilmektedir.

Bu çalışmada derin öğrenme algoritmalarının uygulanmasında Tensorflow [103] kütüphanesinden yararlanılmıştır. Derin öğrenme alanında kullanılan teknolojilere bakıldığında en çok kullanılan kütüphanelerden birisinin TensorFlow kütüphanesi olduğu görülmektedir. Son zamanlarda yapay zekâ alanına yaptığı yatırımlarla bilinen Google tarafından geliştirilmektedir.

Çalışma kapsamında ARIMA analiz sonuçlarının elde edilmesinde Statsmodels[104] kütüphanesinden yararlanılmıştır. Pmdarima [105] kütüphanesindeki Auto ARIMA fonksiyonu kullanılarak hiper-parametre optimizasyonu gerçekleştirilmiş ve her analiz için en uygun parametreler kullanılarak analizler gerçekleştirilmiştir.

Bu çalışma kapsamında makine öğrenmesi algoritmaları kullanılarak ikili ve üç sınıflı olmak üzere 2 farklı sınıflandırma analizi gerçekleştirilmiştir. Oluşturulan modellerin, daha önce gösterilmeyen yeni verileri daha öncesinde belirlenen sınıflardan hangisine ait olduğunu belirlemesine sınıflandırma denilmektedir. Sınıflandırma, denetimli öğrenme tekniklerinden birisidir ve amacı, eğitim seti ile tanımlanan sınıflar arasına yeni gelen verileri doğru şekilde yerleştirmektedir. Sınıflandırma 2 türe ayrılmaktadır. Bunlar:

- İkili Sınıflandırma (Binary Classification): Yapılan analizlerde 2 farklı sonucu olan yani 2 sınıfa ayrılan denetimli öğrenme algoritmalarıdır. Örneğin, X'da duygu analizi yapıldığı zaman sınıflar olumlu gönderiler için pozitif veya olumsuz gönderiler için negatif olarak ayrılabilir.
- Çok Sınıflı Sınıflandırma (Multi Class Classification): Yapılan analizlerde ikiden fazla sınıfa sahip olan denetimli öğrenme algoritmalarıdır ve her veriye bir etiket atanmaktadır. Örneğin, Hisse senedi zaman serilerinde artış olması 1, azalış olması -1 ve değişim olmaması durumu 0 olarak etiketlenebilir.

İkili sınıflandırma analizinde getirinin negatif olduğu durumlar -1, sıfır veya pozitif olduğu durumlar ise 1 olarak etiketlenerek ikili sınıflandırmaya uygun hale getirilmiştir. Üç sınıflı sınıflandırma analizinde ise getirinin <0.001 olduğu durumlar -1, >0.001 olduğu durumlar +1 ve getirinin ± 0.001 aralığında olduğu durumlar ise 0 olarak etkilenecek üç sınıflı sınıflandırma analizine uygun hale getirilmiştir. Üç sınıflı sınıflandırma analizinde getirinin ± 0.001 aralığında olduğu durumların 0 etiketlenmesi ile düşük getirinin olduğu durumlarda alım-satım işlemi yapılmasının önüne geçmeye çalışarak sadece işlem maliyetini karşılayacak durumlarda alım-satım işlemi yapılması hedeflenmektedir.

Sınıflandırma analizlerinin performanslarının değerlendirilmesinde Şekil 5.2'de yer alan karmaşıklık matrisinden yararlanılarak oluşturulan kriterler kullanılmaktadır.

	Tahmini Pozitif	Tahmini Negatif
Gerçek Pozitif	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
Gerçek Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)

Şekil 5.2. Karmaşıklık matrisi

Doğruluk (Accuracy): Sınıflandırma analizindeki doğru tahminlerin, tüm tahminlere göre yüzdesini hesaplayan temel metriktir. Denklem 5.1'deki formül ile hesaplanmaktadır.

$$Doğruluk = \frac{DP + DN}{DP + YP + DN + YN} \quad (5.1)$$

Kesinlik (Precision): Sınıflandırma analizindeki pozitif tahminlerinin ne kadarının gerçek pozitif olduğunu ölçer. Sınıflandırma sonucunun kesinlik derecesini gösterir. Denklem 5.2'deki formül ile hesaplanır.

$$Kesinlik = \frac{DP}{DP + YP} \quad (5.2)$$

Duyarlılık (Recall): Sınıflandırma analizindeki tahminlerin gerçek pozitifleri ne kadar yakaladığını gösterir. Duyarlılık pozitif olarak etiketlenmiş örneklerin gerçekten pozitif olan örneklerin toplamına oranıdır. Denklem 5.3'teki formül ile hesaplanır.

$$Duyarlılık = \frac{DP}{DP + YN} \quad (5.3)$$

F1 Skoru (F1 Score): Kesinlik ve Duyarlılık değerlerinin harmonik ortalamasıdır. Hem Kesinlik hem de Duyarlılık değerinin etkisini içermektedir. Denklem 5.4'teki formül ile hesaplanmaktadır.

$$F1 \text{ Skoru} = 2x \frac{Kesinlik \times Duyarlılık}{Kesinlik + Duyarlılık} \quad (5.4)$$

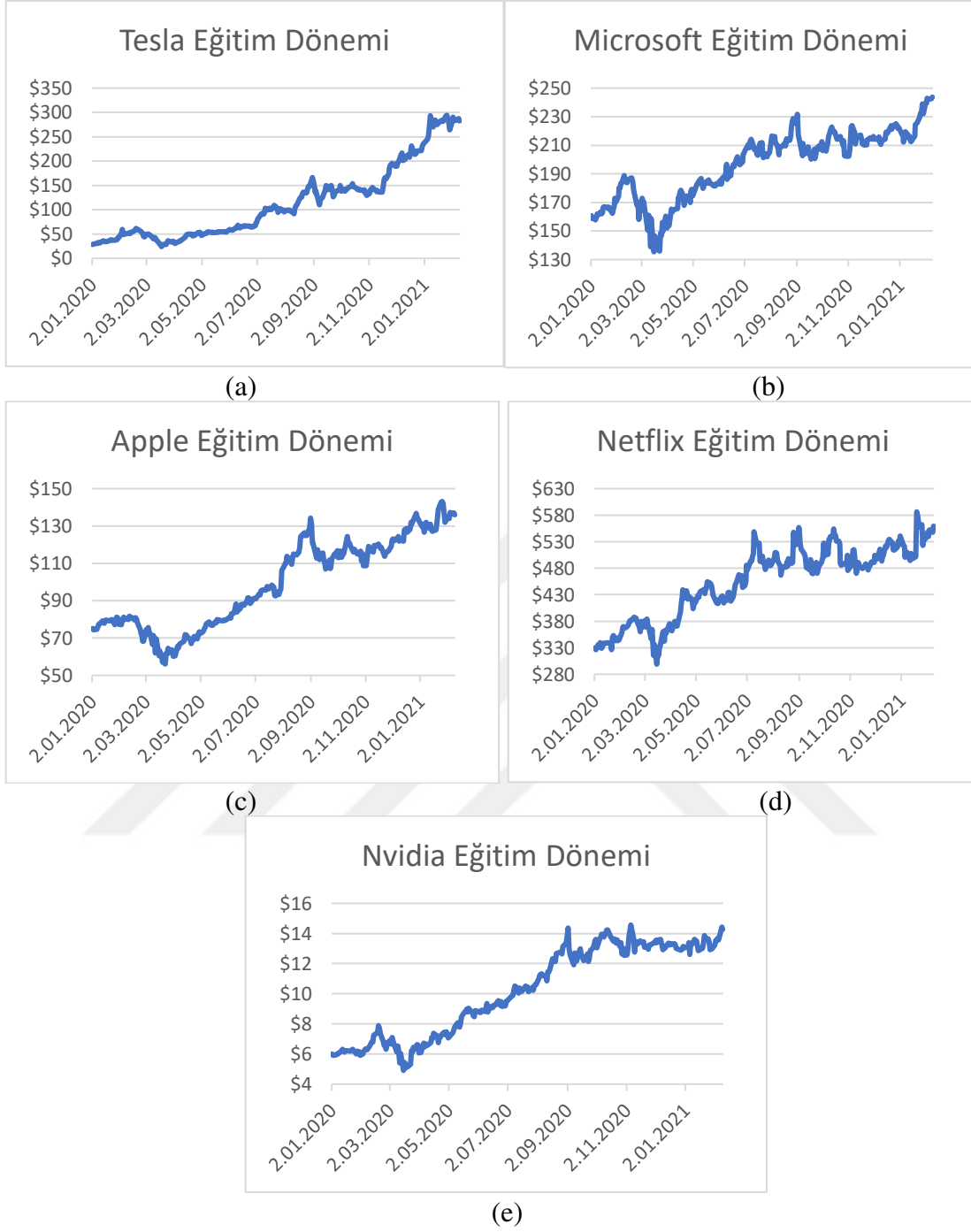
Çalışma kapsamındaki sayısal değerlerin normalize edilmeleri için min-max normalizasyon yöntemi uygulanmıştır. Bu yöntemde, bir grup verinin içerisindeki en büyük ve en küçük değerler ele alınır. Diğer bütün veriler, bu değerlere göre normalleştirilir. Buradaki amaç en küçük değeri 0 ve en büyük değeri 1 olacak şekilde normalleştirmek ve diğer bütün verileri bu 0-1 aralığına yaymaktır. Denklem 5.5'teki formül ile hesaplanır.

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (5.5)$$

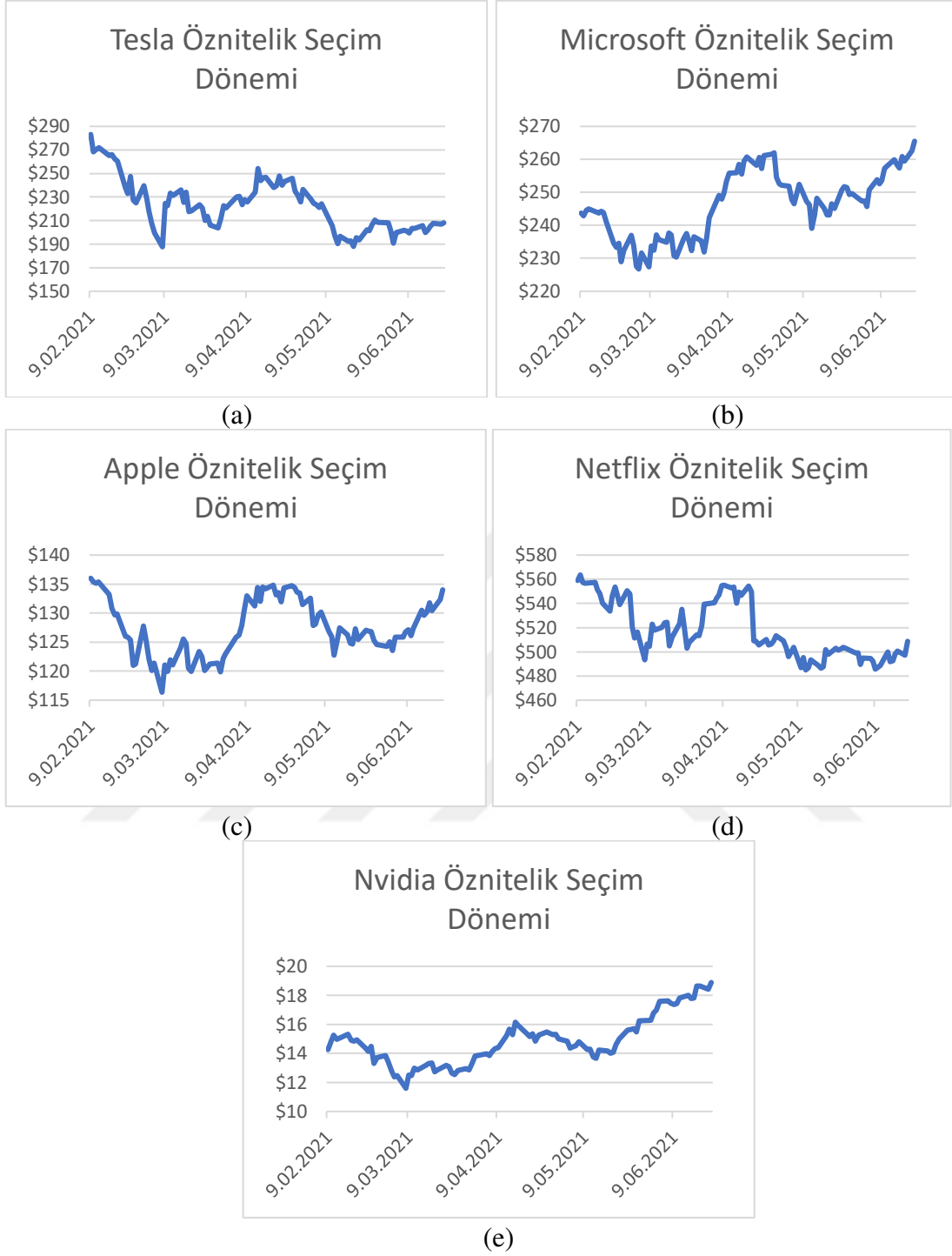
Toplanan veriler Eğitim, Öznitelik seçimi ve Test olmak üzere üç farklı döneme ayrılmıştır ve Test dönemi içinde üç farklı alternatif oluşturulmuştur. 01.01.2020 – 09.02.2021 tarihleri arasında kalan dönem makine öğrenmesi algoritmalarının eğitiminde kullanılmak üzere eğitim verisi olarak ayrılmıştır. Eğitim dönemindeki veriler kullanılarak eğitilen makine öğrenmesi algoritmaları, en yüksek karlılığı sağlayacak öznitelik kombinasyonunun belirlenebilmesi için 09.02.2021-21.06.2021 tarihleri arasında kalan yaklaşık dört aylık dönem öznitelik seçim dönemi olarak belirlenmiştir. Eğitim dönemin için beş şirketin fiyat değişim grafikleri Şekil 5.4’te, Öznitelik seçim dönemi için ise Şekil 5.5’te gösterilmiştir. Belirlenen en iyi öznitelik kombinasyonunun test edilmesi amacıyla da üç farklı test dönemi oluşturulmuştur. Oluşturulan eğitim, öznitelik seçim ve test dönemlerinin zaman çizelgesi Şekil 5.3.’de görülebilmektedir. Birinci test dönemi 22.06.2021–31.10.2021 tarihleri arasındaki yaklaşık dört aylık dönemi kapsamaktadır ve borsanın yükseliş dönemini temsil etmektedir. Birinci test döneminde beş şirketin fiyat değişimlerini gösteren grafikler Şekil 5.6.’de görülebilmektedir. İkinci test dönemi 01.11.2021–14.03.2022 tarihleri arasındaki yaklaşık dört aylık dönemi kapsamaktadır ve borsanın düşüş dönemini temsil etmektedir. İkinci test döneminde 4 şirket ciddi değer kaybı yaşamıştır. İkinci test döneminde beş şirketin fiyat değişimlerini gösteren grafikler ise Şekil 5.7.’de görülebilmektedir.



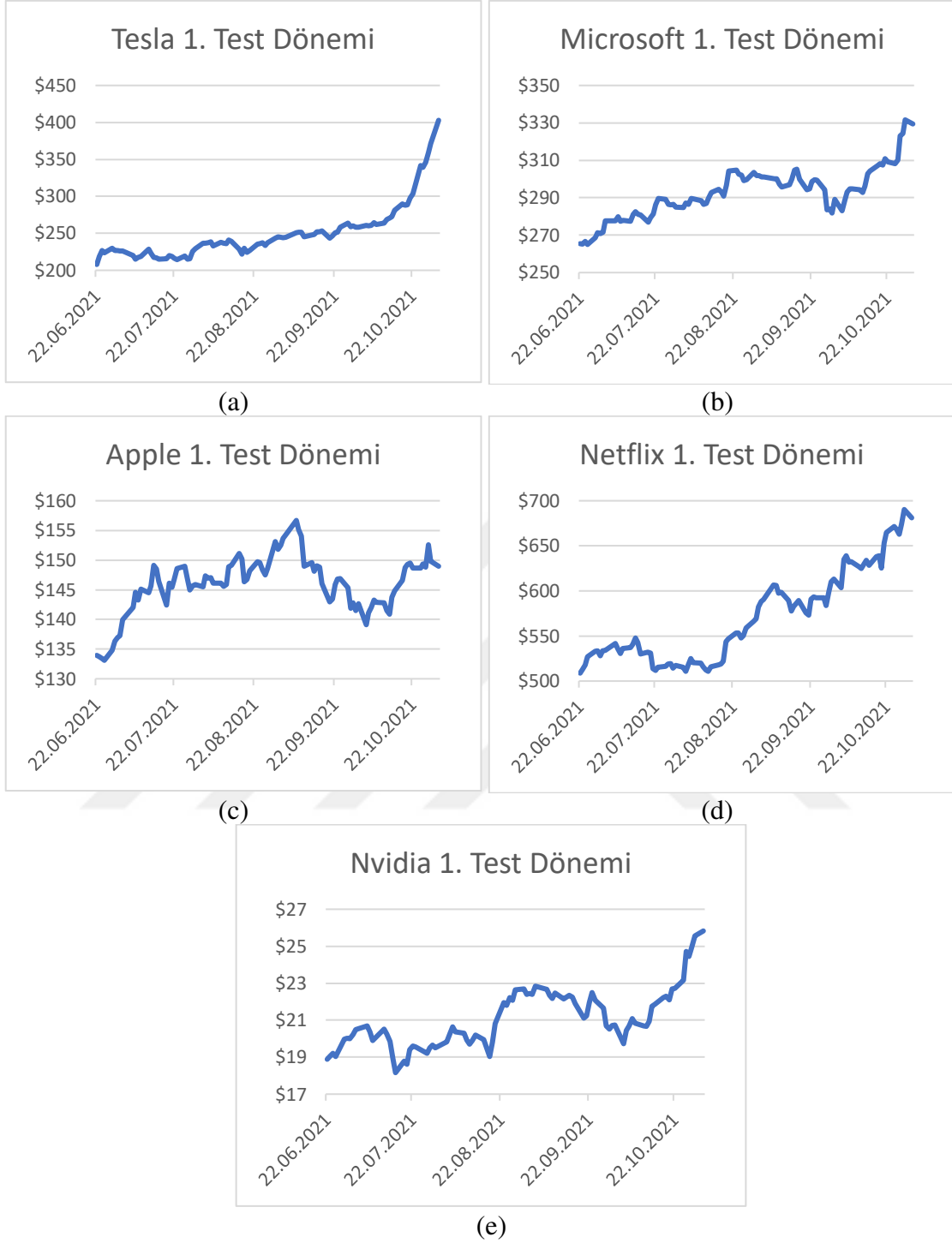
Şekil 5.3. Algoritma verilerinin zaman çizelgesi



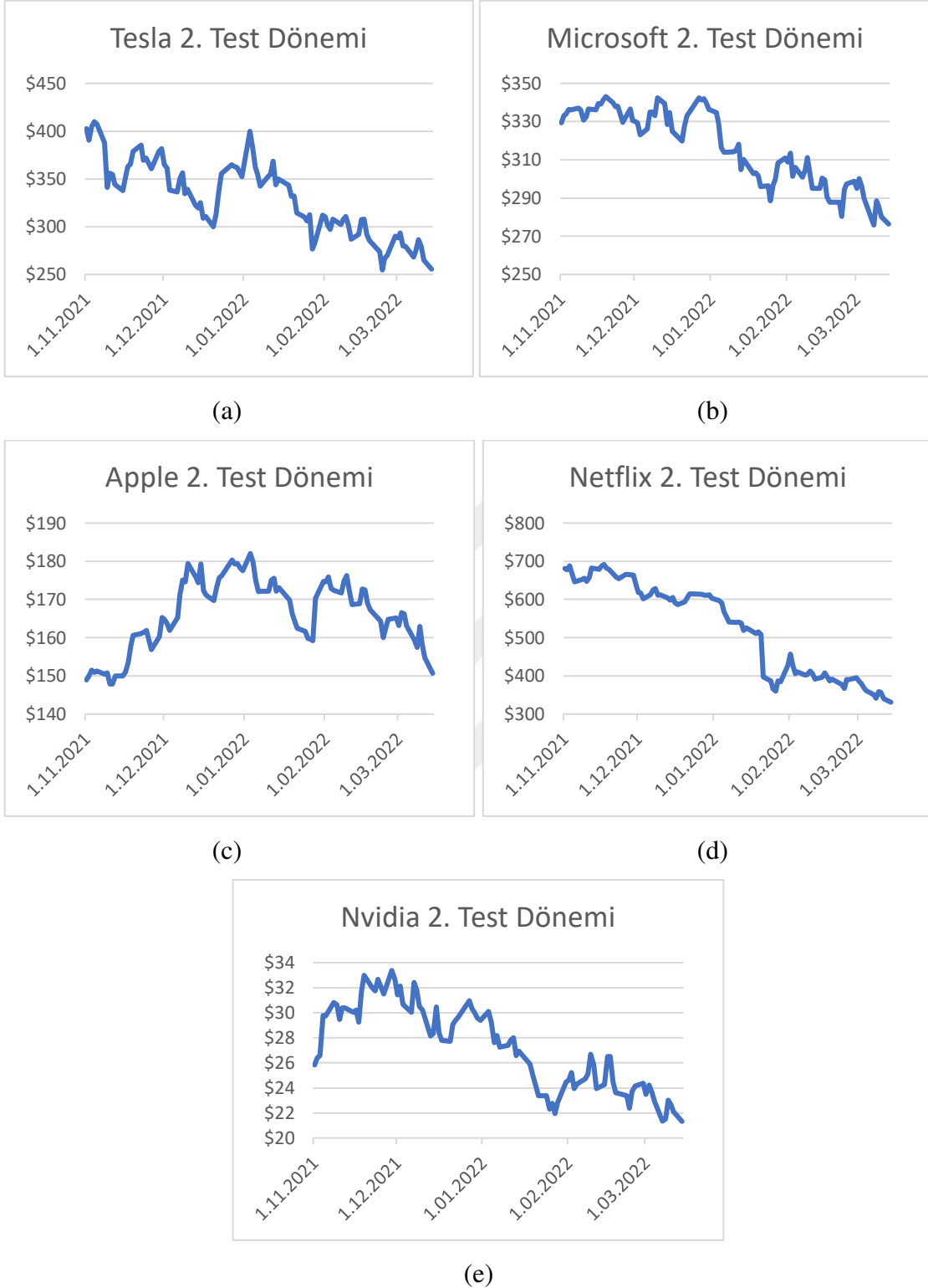
Şekil 5.4. Eğitim dönemi beş şirketin fiyat grafikleri a)Tesla b)Microsoft c)Apple d)Netflix e)Nvidia



Şekil 5.5. Öznitelik dönemi beş şirketin fiyat grafikleri a)Tesla b)Microsoft c)Apple d)Netflix e)Nvidia



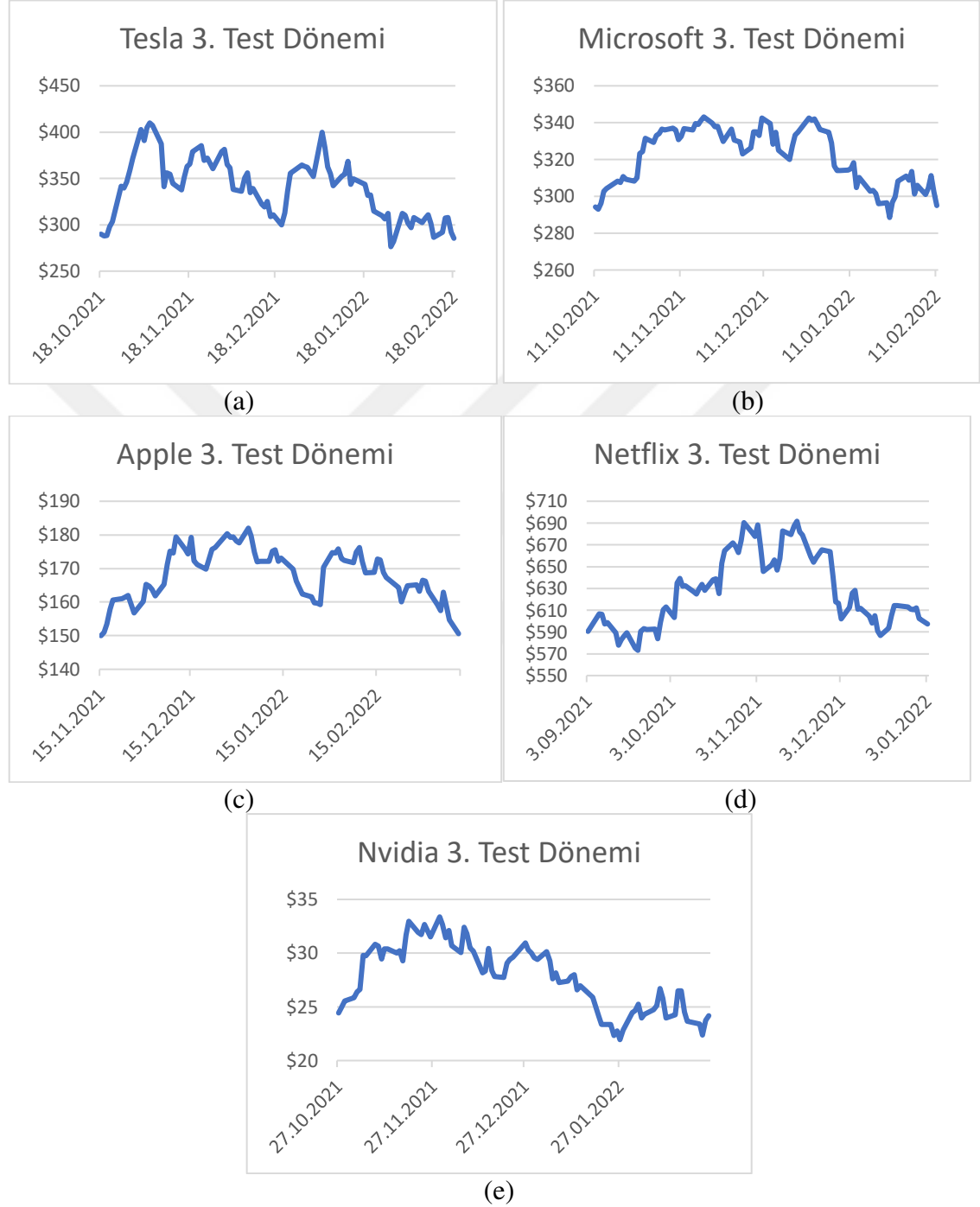
Şekil 5.6. Birinci test dönemi beş şirketin fiyat grafikleri a)Tesla b)Microsoft c)Apple d)Netflix e)Nvidia



Şekil 5.7. İkinci test dönemi beş şirketin fiyat grafikleri a)Tesla b)Microsoft c)Apple d)Netflix e)Nvidia

Üçüncü test döneminde ise yaklaşık dört aylık sürede başlangıç ve bitiş anlarındaki fiyat farkının minimum olduğu durum temsil edilmek istenmiştir. Her şirket için farklı tarih aralıkları belirlenmiştir ve belirlenen tarih aralıkları Tablo 5.2’de

görülmektedir. Üçüncü test döneminde beş şirketin fiyat değişimlerini gösteren grafikler Şekil 5.8.'de görülebilmektedir.



Şekil 5.8. Üçüncü test dönemi beş şirketin fiyat grafikleri a)Tesla b)Microsoft c)Apple d)Netflix e)Nvidia

Tablo 5.1. Beş şirketin üçüncü test dönemi başlangıç ve bitiş tarihleri

	Başlangıç Tarihi	Bitiş Tarihi
Tesla	18.10.2021	18.02.2022
Microsoft	11.10.2021	11.02.2022
Apple	14.11.2021	14.03.2022
Netflix	3.09.2021	3.01.2022
Nvidia	27.10.2021	27.02.2022

5 şirkete ait hisse senetlerinin 5 farklı zaman periyodundaki fiyat değişim yönleri ikili sınıflandırma analizi için pozitif ve negatif, üçlü sınıflandırma için ise pozitif, nötr ve negatif olarak etiketlenmiştir. Üçlü sınıflandırmada getirinin ± 0.001 aralığında olduğu durumlar nötr olarak etiketlenmiştir. 3 farklı Test döneminde hisse senetlerinin dağılımlarının yüzdelik oranları ikili sınıflandırma durumu için Tablo 5.2.'de, üçlü sınıflandırma durumu için ise Tablo 5.3.'de verilmiştir.

Tablo 5.2. Beş şirketin pozitif ve negatif dağılım yüzdeleri

		Tesla		Microsoft		Apple		Netflix		Nvidia	
	Periyot (dk)	Pozitif (%)	Negatif (%)	Pozitif (%)	Negatif (%)	Pozitif (%)	Negatif (%)	Pozitif (%)	Negatif (%)	Pozitif (%)	Negatif (%)
1.Test Dönem	5	50,94	49,06	51,62	48,38	52,32	47,68	50,56	49,44	51,9	48,1
	15	51,53	48,47	52,19	47,81	52,53	47,47	51,42	48,58	51,57	48,43
	30	51,63	48,37	53,38	46,62	53,13	46,87	50,96	49,04	51,71	48,29
	60	53,88	46,12	55,12	44,88	54,18	45,82	52,02	47,98	53,42	46,58
	120	54,35	45,65	57,34	42,66	52,97	47,03	53,26	46,74	52,17	47,83
2.Test Dönem	5	48,52	51,48	49,8	50,2	50,8	49,2	48,33	51,67	50,15	49,85
	15	47,94	52,06	49,98	50,02	51,43	48,57	47,94	52,06	50,59	49,41
	30	48,32	51,68	48,7	51,3	49,83	50,17	47,1	52,9	50,08	49,92
	60	46,72	53,28	47,74	52,26	50,47	49,53	43,21	56,79	49,22	50,78
	120	46,72	53,28	46,07	53,93	51,36	48,64	40,87	59,13	47,83	52,17
3.Test Dönem	5	49,16	50,84	50,87	49,13	50,87	49,13	49,49	50,51	50,42	49,58
	15	49,29	50,71	51,04	48,96	51,72	48,28	50,05	49,95	51,36	48,64
	30	48,89	51,11	49,56	50,44	50,24	49,76	49,59	50,41	50,77	49,23
	60	48,6	51,4	49,18	50,82	50,52	49,48	48,73	51,27	50,25	49,75
	120	49,42	50,58	48,14	51,86	51,22	48,78	47,93	52,07	48,82	51,18

Tablo 5.3. Beş şirketin pozitif, nötr ve negatif dağılım yüzdeleri

		Tesla				Microsoft			Apple		
		Periyot	Pozitif	Nötr	Negatif	Pozitif	Nötr	Negatif	Pozitif	Nötr	Negatif
		(dk)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)
1.Test	5	26,53	47,62	25,85	12,08	76,86	11,06	15,7	69,69	14,61	
Dönemi	15	35,39	31,13	33,47	23,19	56,12	20,69	26,1	49,6	24,3	
	30	40,85	21,22	37,93	31,41	40,6	27,99	35,61	33,53	30,86	
	60	45,5	17,24	37,27	41,15	27,64	31,21	39,47	26,63	33,9	
	120	49,18	8,7	42,12	46,74	20,65	32,61	45,41	16,76	37,84	
2.Test	5	35,66	26,11	38,24	21,2	56,43	22,37	23,28	53,86	22,85	
Dönemi	15	40,34	16,01	43,66	31,1	36,68	32,23	32,75	34,63	32,62	
	30	41,93	12,18	45,88	35,88	26,91	37,22	38,09	24,5	37,42	
	60	43,75	6,41	49,84	37,64	18,97	43,39	40,81	16,67	42,52	
	120	44,26	6,01	49,73	40,65	11,11	48,24	44,29	12,5	43,21	
3.Test	5	34,87	28,64	36,49	18,66	62,52	18,82	24,71	50,9	24,38	
Dönemi	15	40,44	17,91	41,64	29,43	41,86	28,72	33,71	32,3	33,99	
	30	41,96	12,8	45,24	34,13	32,27	33,6	39,02	22,24	38,74	
	60	44,46	8,26	47,27	37,99	21,88	40,13	42,13	14,86	43,01	
	120	47,11	5,2	47,69	41,83	13,47	44,7	45,73	10,06	44,21	
		Netflix				Nvidia					
	Periyot	Pozitif	Nötr	Negatif	Pozitif	Nötr	Negatif				
	(dk)	(%)	(%)	(%)	(%)	(%)	(%)				
1.Test	5	21,02	59	19,98	25,55	50,07	24,38				
Dönemi	15	31,66	38,53	29,82	35,19	31,34	33,47				
	30	39,01	27,74	33,25	40,02	23,81	36,17				
	60	43,48	19,57	36,96	44,41	18,17	37,42				
	120	46,47	15,22	38,32	47,83	9,24	42,93				
2.Test	5	27,73	42	30,27	36,1	27,84	36,07				
Dönemi	15	33,89	26,75	39,36	42,15	16,29	41,56				
	30	36,52	18,81	44,67	43,96	10,82	45,22				
	60	37,91	12,17	49,92	45,48	8,41	46,11				
	120	38,15	8,72	53,13	45,65	5,43	48,91				
3.Test	5	22,8	54,18	23,02	35,72	29,09	35,19				
Dönemi	15	32,25	34,81	32,94	42,41	17,27	40,32				
	30	37,67	24,48	37,85	44,32	11,17	44,5				
	60	41,46	16,41	42,13	46,21	9,11	44,69				
	120	43,2	10,65	46,15	47,06	5,29	47,65				

Öznitelik seçim döneminde 5 farklı şirket (Tesla, Microsoft, Apple, Netflix ve Nvidia) için ikili sınıflandırma ve üç sınıflı sınıflandırma analizleri dört farklı makine öğrenmesi algoritması (Rastgele orman, Karar ağacı, Destek vektör makinesi ve Lojistik regresyon) kullanarak gerçekleştirilmiştir. Bu sınıflandırma analizlerinde oluşturulan 2

farklı veri seti için (tüm gönderiler ve sadece cashtag içeren gönderiler), üç farklı duygu analiz algoritması kullanılarak (Stanza, Textblob, Vader), 5 farklı zaman periyodu için (5 - 15 - 30 - 60 - 120), kelime vektörünün dahil edildiği ve edilmediği 2 farklı durum , sadece metin duygu durum değeri olduğu ve hem metin hem de emoji duygu durum değerlerinin olduğu 2 farklı durum , 6 farklı duygu durum değeri değişkeni ve Tablo 5.4.'de gösterilen 45 farklı bağımsız değişken kombinasyonu kullanılarak analizler gerçekleştirilmiştir. 13 - 27 arasındaki değişken kombinasyonlarında “Önceki Hareket Yönü” adıyla bir önceki periyottaki hisse senedi değişim yönü analize dahil edilmiştir. Çalışma kapsamında değişken kombinasyonlarını oluşturmak üzere 4 adet teknik indikatör seçilmiş ve analizlere dahil edilmiştir. Bu teknik indikatörlerin seçiminde Scikit-Learn Kütüphanesinin SelectKbest fonksiyonundan yararlanılmıştır. Seçilen teknik indikatörlerin farklı indikatör gruplarını temsil etmesi hedeflenmiştir. Öznitelik belirleme dönemindeki analizlerde, Rastgele orman, Karar ağacı, Destek vektör makinesi ve Lojistik regresyon algoritmalarının analizlere dahil edilmesinde Scikit-Learn Kütüphanesinin varsayılan hiper-parametreleri kullanılmıştır.

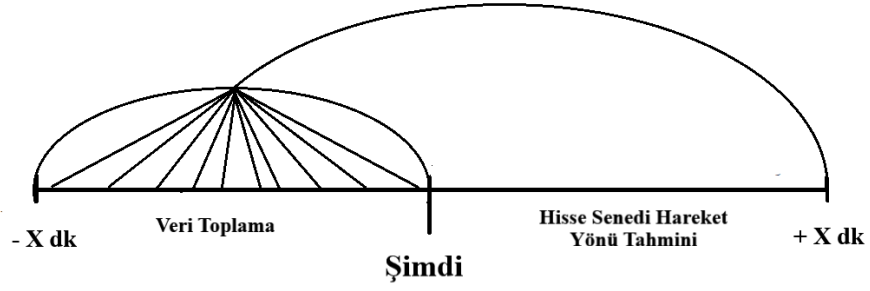
Tablo 5.4. Sınıflandırma değişken kombinasyonları

No	Değişkenler
1	KCP
2	KCP, Duygu Değeri
3	KCP, Yazar Sayısı, Duygu Değeri
4	EOM
5	EOM, Duygu Değeri
6	EOM, Yazar Sayısı, Duygu Değeri
7	CCI
8	CCI, Duygu Değeri
9	CCI, Yazar Sayısı, Duygu Değeri
10	Stokastik RSI
11	Stokastik RSI, Duygu Değeri
12	Stokastik RSI, Yazar Sayısı, Duygu Değeri
13	Önceki Hareket Yönü
14	Önceki Hareket Yönü, Duygu Değeri
15	Önceki Hareket Yönü, Yazar Sayısı, Duygu Değeri
16	KCP, Önceki Hareket Yönü
17	KCP, Önceki Hareket Yönü, Duygu Değeri

Tablo 5.4. (Devam) Sınıflandırma değişken kombinasyonları

18	KCP, Önceki Hareket Yönü, Yazar Sayısı, Duygu Değeri
19	EOM, Önceki Hareket Yönü
20	EOM, Önceki Hareket Yönü, Duygu Değeri
21	EOM, Önceki Hareket Yönü, Yazar Sayısı, Duygu Değeri
22	CCI, Önceki Hareket Yönü
23	CCI, Önceki Hareket Yönü, Duygu Değeri
24	CCI, Önceki Hareket Yönü, Yazar Sayısı, Duygu Değeri
25	Stokastik RSI, Önceki Hareket Yönü
26	Stokastik RSI, Önceki Hareket Yönü, Duygu Değeri
27	Stokastik RSI, Önceki Hareket Yönü, Yazar Sayısı, Duygu Değeri
28	KCP, EOM
29	KCP, EOM, Duygu Değeri
30	KCP, EOM, Yazar Sayısı, Duygu Değeri
31	KCP, Stokastik RSI
32	KCP, Stokastik RSI, Duygu Değeri
33	KCP, Stokastik RSI, Yazar Sayısı, Duygu Değeri
34	KCP, CCI
35	KCP, CCI, Duygu Değeri
36	KCP, CCI, Yazar Sayısı, Duygu Değeri
37	EOM, Stokastik RSI
38	EOM, Stokastik RSI, Duygu Değeri
39	EOM, Stokastik RSI, Yazar Sayısı, Duygu Değeri
40	EOM, CCI
41	EOM, CCI, Duygu Değeri
42	EOM, CCI, Yazar Sayısı, Duygu Değeri
43	CCI, Stokastik RSI
44	CCI, Stokastik RSI, Duygu Değeri
45	CCI, Stokastik RSI, Yazar Sayısı, Duygu Değeri

Çalışma kapsamında belirlenen zaman periyotları için sosyal medya verileri ile finansal verilerin ilişkilendirilme mantığı şekil 5.9.'da görülmektedir. Çalışma kapsamında bir önceki zaman periyodunda toplanan sosyal medya paylaşımlarından elde edilen veriler bir sonraki zaman periyodu sonundaki fiyat değişiminin yönünü tahmin etmek üzere kullanılmıştır.



Şekil 5.9. Sosyal medya ile finansal verilerin ilişkilendirilmesi

5.1. Market Simülasyonu ile Kazanç Analizi

Bu çalışmada, bir market simülasyonu oluşturularak karlılık analizi gerçekleştirilmiştir. Tahmin edilen hareket yönü kullanılarak bir market simülasyonu oluşturulmuş ve karlılık analizi gerçekleştirilmiştir. Market simülasyonu şu kurallarla çalışmaktadır;

- Eğer hisse senedi alındıysa ve tahmin negatif ise, hisse senedi satılır.
- Eğer hisse senedi alındıysa ve tahmin pozitif ise, beklenir.
- Eğer hisse senedi satıldıysa ve tahmin pozitif ise, eldeki tüm bakiye ile hisse senedi alınır.
- Eğer hisse senedi satıldıysa ve tahmin negatif ise, beklenir.

Başlangıç sermayesi her bir şirket için 10,000 \$ olarak belirlenmiş ve 5 şirket için toplam 50,000\$ yatırım gerçekleştirilmiştir. Yapılan analiz sonucunda her bir adımda hisse senedi değerinin yükseleceği tahmin ediliyorsa eldeki para ile hisse senedi alınmış, eğer hisse senedi değerinin azalacağı tahmin ediliyorsa eldeki hisse senetleri satılmıştır. Hisse senedi alım-satımında işlem ücreti olduğu durumda hem hisse senedi alımında hem de hisse senedi satımında %0,1 oranında işlem ücreti kesilmiştir. Analizler şirketler için birbirlerinden bağımsız olacak şekilde gerçekleştirilmiştir.

5.2. İki Sınıflı Analiz için Sonuçlar

Öznitelik belirleme dönemi için yapılan tahminlerin market simülasyonunda kullanılması sonrasında en yüksek karlılığı veren parametre kombinasyonları Tablo 5.5'te görülebilmektedir.

Tablo 5.5. İkili sınıflandırmada öznitelik seçim dönemi için en yüksek toplam kar değerleri

TF-IDF	Duygu Analiz	Veri Seti	Algoritma	Değişkenler	Duygu Değeri	Zaman Periyod (dk)	Toplam Kar-Zarar (\$)	Doğruluk (%)
Yok	Vader	Tüm Gönderiler	SVM	KCP, Değeri 5	Duygu Metin	30	2152,89	51,19
Yok	Vader	Tüm Gönderiler	SVM	KCP, Değeri 5	Duygu Metin ve Emoji	30	2142,48	51,23
Yok	Vader	Tüm Gönderiler	SVM	EOM, CCI, Duygu Değeri 3	Metin	120	2131,61	51,67
Yok	Textblob	Tüm Gönderiler	SVM	EOM, CCI, Duygu Değeri 5	Metin	120	2003,08	51,84
Yok	Vader	Sadece Cashtag	SVM	KCP, CCI, Yazar Sayısı, Duygu Değeri 3	Metin	120	1991,67	51,51
Yok	Textblob	Sadece Cashtag	SVM	EOM, CCI, Duygu Değeri 3	Metin	120	1973,02	51,89
Yok	Textblob	Sadece Cashtag	SVM	EOM, CCI, Duygu Değeri 3	Metin ve Emoji	120	1973,02	51,89
Yok	Vader	Tüm Gönderiler	SVM	CCI, Değeri 3	Duygu Metin	120	1943,51	51,29
Yok	Vader	Tüm Gönderiler	SVM	CCI, Değeri 3	Duygu Metin ve Emoji	120	1921,36	51,23
Yok	Textblob	Tüm Gönderiler	SVM	CCI, Değeri 4	Duygu Metin	120	1902,43	51,29

En yüksek karlılık değerine, SVM algoritması kullanılarak, 30-30 zaman periyodunda ulaşılmıştır. KCP teknik indikatörleri ile Duygu değeri 5'in kullanıldığı bu durum için dört aylık sürede 5 şirket için toplam 2152 \$ kazanç sağlanmıştır. Duygu durum değeri, Vader kütüphanesi ile tüm gönderilerden oluşan veri seti için sadece sosyal medya paylaşımlarının metin verileri kullanılarak elde edilmiştir. Her bir Paylaşımın duygu durum değeri, Post etki değeri ile çarpılıp periyod boyunca toplanarak Duygu değeri 5 elde edilmiştir. Sosyal medya paylaşımlarının metin verisine ait kelime vektörünün dahil edilmediği durumda en yüksek karlılık değerine ulaşılmıştır.

SPSS programı kullanılarak en yüksek toplam kar olan 2152.89\$ değerine Tek Örneklem T-test'i uygulanarak ortalama anlamlı farklılığı incelenmiştir. Bu analiz için H_0 ve H_1 hipotezleri aşağıdaki gibidir:

H_0 = En iyi öznitelik kombinasyonunun toplam kar miktarı, tüm öznitelik kombinasyonlarının toplam karlılık değerlerinin ortalamasından farklı değildir

H_1 = En iyi öznitelik kombinasyonunun toplam kar miktarı, tüm öznitelik kombinasyonlarının toplam karlılık değerlerinin ortalamasından farklıdır.

Tablo 5.6'da gösterilen sonuçlar incelendiğinde P(sig) değerinin 0.05 değerinden küçük olması nedeniyle H_0 hipotezi reddedilir ve belirlenen öznitelik kombinasyonuna

ait toplam kar değerinin ortalamadan istatistiksel olarak anlamlı derecede farklı olduğu söylenebilir.

Tablo 5.6. İkili sınıflandırmada toplam kar-zarar değerinin T-testi değerleri

Tek-Orneklem Test						
	Test Değeri = 2152.89					
	t	df	Sig. (2-kuyruklu)	Ortalama Fark	Farkın %95'lik Güven Aralığı	
					Alt	Üst
Toplam Kar-Zarar	-419,754	89999	,000	-21894.98	-21997.21	-21792.74

Öznitelik belirleme dönemi için elde edilen en iyi değerler karşılaştırıldığında en iyi sonuçların duygu değerinin dahil olduğu parametre kombinasyonlarında elde edildiği görülmektedir. Bu durumdan yola çıkarak duygu durum değerinin analizlere dahil edilmesinin karlılığı arttığı söylenebilir.

Belirlenen en iyi öznitelikler kullanılarak üç farklı test dönemi için analizler gerçekleştirilmiştir. Seçilen şirketler için market simülasyonu ile test dönemi sonunda ulaşılan kar veya zarar, bu kar ve zarar sonucunda dönem sonunda ulaşılan toplam yatırım değeri, market simülasyonu ile alım-satım işlemlerinin sayısı ve nakitte kalma oranı, algoritma tahminlerinin doğruluk yüzdesi, algoritma tahminlerinin F1-skor değeri, Şirket hisse senetleri ile Al-Tut stratejisi uygulanması halinde hisse senetlerinin Kar-Zarar miktarları ve Al-Tut stratejisi sonucunda hisse senetlerinin dönem sonu toplam yatırım değeri hesaplanmıştır.

Her bir test dönemi için 2 farklı durum oluşturulmuş ve karşılaştırılmıştır. İlk durumda en iyi sonucu veren değişken kombinasyonu için sonuçlar elde edilmiştir. İkinci durumda en iyi değişken kombinasyonunun sosyal medya verilerinden elde edilen özniteliklerin çıkarılmış hali için sonuçlar elde edilmiştir. İlk durumda belirlenen KCP teknik indikatörü ve Duygu değeri 5 değeri ile SVM algoritması kullanılarak 30 dk zaman periyodu için analizler gerçekleştirilmiştir. İkinci durumda ise Duygu değeri 5 analizden çıkarılmış ve sadece KCP teknik indikatörü kullanılarak SVM algoritması ile 30 dk zaman periyodu için analiz gerçekleştirilmiştir. Borsanın yükseliş durumunu temsil eden 1. Test dönemi için karşılaştırmalı sonuçlar Tablo 5.7’de görülebilmektedir. En iyi değere SVM algoritması ile ulaşıldığı bilinmektedir. Tablo 5.8’de ise birinci test döneminde SVM algoritması ile elde edilen sonuçlar ile ARIMA yöntemi kullanılarak elde edilen

sonuçların karşılaştırması gösterilmiştir. ARIMA yöntemi ile öncelikle hisse senetlerinin belirlenen öznitelikteki zaman periyodu için getiri tahmini gerçekleştirilmiş ve daha sonrasında ikili sınıflandırma probleminde kullanmak üzere pozitif getiri değerleri 1, negatif getiri değerleri 0 olarak etiketlenmiştir. Bu 2 sınıflı olarak etiketlenen ARIMA tahminleri, market simülasyonunda kullanılmış böylece diğer tablo değerlerinin hesaplaması gerçekleştirilmiştir. Tablolardaki kar-zarar miktarları ve toplam yatırım değerleri 5 şirketin değerlerinin toplamı alınarak gösterilmiştir. Tablolardaki dönem boyunca nakitte kalma oranı, algoritma ile alım-satım işlem sayısı, doğruluk yüzdesi ve F1-skor değeri, 5 şirketin ortalaması alınarak gösterilmiştir.

Tablo 5.7. İkili sınıflandırmada birinci test dönemi analiz sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Algoritma ile Elde Edilen Kazanç (\$)	12304,97	3866,21
Dönem Sonu Toplam Yatırım Değeri (\$)	62304,97	53866,21
Dönem Boyunca Nakitte Kalma Oranı (%)	10	42
Algoritma ile Alım-Satım İşlem Sayısı	41	170
Doğruluk Oranı (%)	51,48	51,34
F1-Skor Değeri	0,65	0,50
Al-Tut Stratejisi ile Kazanç (\$)	19320,93	19320,93
AL-Tut Dönem Sonu Toplam Yatırım Değer (\$)	69320,93	69320,93

Tablo 5.8. İkili sınıflandırmada birinci test dönemi yöntem karşılaştırması

	Belirlenen Öznitelikler ile Analiz Sonucu	ARIMA Yöntemi ile Ulaşılan Sonuç
Algoritma ile Elde Edilen Kazanç (\$)	12304,97	-8655,36
Dönem Sonu Toplam Yatırım Değeri (\$)	62304,97	41344,64
Dönem Boyunca Nakitte Kalma Oranı (%)	10	46
Algoritma ile Alım-Satım İşlem Sayısı	41	413
Doğruluk Oranı (%)	51,48	49,18
F1-Skor Değeri	0,65	0,52
Al-Tut Stratejisi ile Kazanç (\$)	19320,93	19320,93
AL-Tut Dönem Sonu Toplam Yatırım Değer (\$)	69320,93	69320,93

Belirlenen Öznitelikler ve SVM algoritması kullanılarak 30dk zaman periyodu için elde edilen tahminler sonucunda %51,48 doğruluk yüzdesi ve 0,65 F1-skor değerine

ulaşılmıştır. Analizlerin toplam işlem süresi 18 saniyedir. Elde edilen bu tahminlerin market simülasyonunda kullanılması sonucunda ise 5 Şirket için 12304\$'lık kar elde edilmiş ve dönem sonunda 62304\$ toplam yatırım değerine ulaşılmıştır. 5 şirket için Al-Tut stratejisi ile elde edilen kazanç miktarı 19320\$ ve dönem sonundaki toplam yatırım değeri 69320\$ olarak hesaplanmıştır. Borsanın yükseliş dönemini temsil eden 1. Test döneminde 12304\$ kar elde edilmesine rağmen hisse senetlerinin Al-Tut strateji sonucunda kazandığı değer altında kazançlar elde edilebilmiştir.

Tablo 5.7'deki 2 farklı durum incelendiğinde Belirlenen özniteliklerden sosyal medya verilerinin çıkarılması durumunda elde edilen kar miktarının 12304\$'dan 3866\$'a, toplam yatırım miktarı ise 62304\$'dan 53866\$'a gerilediği görülmektedir. Sonuçlar değerlendirildiğinde alım-satım işlem sayısı sosyal medya verilerinin çıkarıldığı durumda ortalama 173 adete artış göstermiştir. Alım-satım işlem sayısının yüksek olması her işlem başına %0.1 oranındaki işlem ücreti nedeniyle karlılık miktarını düşürmektedir. Sonuçlar değerlendirildiğinde 1. Test dönemi için sosyal medya verilerinin kullanılmasının karlılık miktarını arttırdığı söylenebilir.

Tablo 5.8'deki sonuçlar değerlendirildiğinde belirlenen öznitelik değerleri ile SVM algoritması kullanılarak 12304\$ kar elde edilirken ARIMA yöntemi tahminlerinin market simülasyonunda kullanılması sonucunda 8655\$ zarar edildiği görülmektedir. SVM algoritması ile 62304\$'lık toplam yatırım değerine ulaşılırken, ARIMA yöntemiyle 41344\$ toplam yatırım değerine ulaşılabilmiştir. Alım-satım işlem sayılarına bakıldığında SVM algoritmasıyla ortalama 41 adet işlem gerçekleştirilirken, ARIMA yöntemiyle ortalama 413 adet işlem gerçekleştirilmiştir. İşlem adetinin bu kadar yüksek olması %0.1'lik işlem ücreti nedeniyle ciddi bir kar kaybına yol açmaktadır.

Borsanın düşüş dönemini temsil eden İkinci test dönemi için aynı öznitelikler ve parametreler kullanılarak analizler tekrarlanmıştır. Sosyal medya verilerinin olduğu ve sosyal medya verilerinin çıkarıldığı iki durum için karşılaştırmalı sonuçlar Tablo 5.9'da gösterilmiştir. İkinci test dönemi için SVM ile elde edilen sonuçların ARIMA yöntemi sonuçlarıyla karşılaştırılması ise Tablo 5.10'da görülebilmektedir.

Tablo 5.9. İkili sınıflandırmada ikinci test dönemi analiz sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Algoritma ile Elde Edilen Kazanç (\$)	-4575,88	-6401,22
Dönem Sonu Toplam Yatırım Değeri (\$)	45424,12	43598,78
Dönem Boyunca Nakitte Kalma Oranı (%)	34	46
Algoritma ile Alım-Satım İşlem Sayısı	27	110
Doğruluk Oranı (%)	49,55	49,77
F1-Skor Değeri	0,47	0,48
Al-Tut Stratejisi ile Kazanç (\$)	-11200,3	-11200,3
AL-Tut Dönem Sonu Toplam Yatırım Değer (\$)	38799,74	38799,74

Tablo 5.10. İkili sınıflandırmada ikinci test dönemi yöntem karşılaştırması

	Belirlenen Öznitelikler ile Analiz Sonucu	ARIMA Yöntemi ile Ulaşılan Sonuç
Algoritma ile Elde Edilen Kazanç (\$)	-4575,88	-18647,5
Dönem Sonu Toplam Yatırım Değeri (\$)	45424,12	31352,51
Dönem Boyunca Nakitte Kalma Oranı (%)	34	53
Algoritma ile Alım-Satım İşlem Sayısı	27	395
Doğruluk Oranı (%)	49,55	51,55
F1-Skor Değeri	0,47	0,50
Al-Tut Stratejisi ile Kazanç (\$)	-11200,3	-11200,3
AL-Tut Dönem Sonu Toplam Yatırım Değer (\$)	38799,74	38799,74

Belirlenen Öznitelikler ve SVM algoritması kullanılarak 30dk zaman periyodu için elde edilen tahminler sonucunda ikinci test döneminde %49,55 doğruluk yüzdesi ve 0,47 F1-skor değerine ulaşılmıştır. Analizlerin toplam işlem süresi 17 saniyedir. Elde edilen bu tahminlerin market simülasyonunda kullanılması sonucunda ise 5 Şirket için 4576\$ zarar edilmiş ve dönem sonunda 45424\$ toplam yatırım değeri elde edilebilmiştir. 5 şirket için Al-Tut stratejisi ile elde edilen zarar miktarı ise 11200\$ ve dönem sonundaki toplam yatırım değeri 38800\$ olarak hesaplanmıştır. Borsanın düşüş dönemini temsil eden 2. Test döneminde toplam 4576\$ zarar edilmesine rağmen gerçekleşen zarar hisse senetlerinin Al-Tut strateji sonucunda kaybettiği değer çok altında kalmıştır.

Tablo 5.9'daki 2 farklı durum incelendiğinde Belirlenen özniteliklerden sosyal medya verilerinin çıkarılması durumunda ulaşılan zarar miktarının 4576\$'dan 6401\$'a

çıkıldığı ve toplam yatırım miktarının 45424\$'dan 43599\$'a gerilediği görülmektedir. Sonuçlar incelendiğinde alım-satım işlem sayısı ve nakitte kalma oranının, sosyal medya verilerinin çıkarıldığı durumda daha yüksek olduğu görülmektedir. Alım-satım işlem sayısının yüksek olması her işlem başına %0.1 oranındaki işlem ücreti nedeniyle karlılık miktarını düşürmektedir. Sonuçlar değerlendirildiğinde 2. Test dönemi için sosyal medya verilerinin kullanılmasının karlılık miktarını arttırdığı söylenebilir.

Tablo 5.10'daki sonuçlar değerlendirildiğinde belirlenen öznitelik değerleri ile SVM algoritması kullanılarak 4576\$ zarar edilirken ARIMA yöntemi tahminlerinin market simülasyonunda kullanılması sonucunda 18648\$ zarar edildiği görülmektedir. SVM algoritması ile 45424\$'lık toplam yatırım değeri elde edilirken, ARIMA yöntemiyle 31553\$ toplam yatırım değerine ulaşılabilmiştir. Alım-satım işlem sayılarına bakıldığında SVM algoritmasıyla ortalama 27 adet işlem gerçekleştirilirken, ARIMA yöntemiyle ortalama 395 adet işlem gerçekleştirilmiştir. İşlem adetinin bu kadar yüksek olması %0.1'lik işlem ücreti nedeniyle ciddi bir kar kaybına yol açmaktadır.

Üçüncü test dönemi içinde aynı öznitelikler ve parametreler kullanılarak analizler tekrarlanmıştır. Sosyal medya verilerinin olduğu ve sosyal medya verilerinin çıkarıldığı iki durum için karşılaştırmalı sonuçlar Tablo 5.11'de gösterilmiştir. İkinci test dönemi için SVM ile elde edilen sonuçların ARIMA yöntemi sonuçlarıyla karşılaştırılması ise Tablo 5.12'de görülebilmektedir.

Tablo 5.11. *İkili sınıflandırmada üçüncü test dönemi analiz sonuçları*

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Algoritma ile Elde Edilen Kazanç (\$)	4292,92	1386,98
Dönem Sonu Toplam Yatırım Değeri (\$)	54292,92	51386,98
Dönem Boyunca Nakitte Kalma Oranı (%)	34	50
Algoritma ile Alım-Satım İşlem Sayısı	25	134
Doğruluk Oranı (%)	50,57	50,22
F1-Skor Değeri	0,49	0,48
Al-Tut Stratejisi ile Kazanç (\$)	-476,23	-476,23
AL-Tut Dönem Sonu Toplam Yatırım Değer (\$)	49523,77	49523,77

Tablo 5.12. İkili sınıflandırmada üçüncü test dönemi yöntem karşılaştırması

	Belirlenen Öznitelikler ile Analiz Sonucu	ARIMA Yöntemi ile Ulaşılan Sonuç
Algoritma ile Elde Edilen Kazanç (\$)	4292,92	-12383,3
Dönem Sonu Toplam Yatırım Değeri (\$)	54292,92	37616,7
Dönem Boyunca Nakitte Kalma Oranı (%)	34	51
Algoritma ile Alım-Satım İşlem Sayısı	25	368
Doğruluk Oranı (%)	50,57	51,28
F1-Skor Değeri	0,49	0,51
Al-Tut Stratejisi ile Kazanç (\$)	-476,23	-476,23
AL-Tut Dönem Sonu Toplam Yatırım Değer (\$)	49523,77	49523,77

Belirlenen Öznitelikler ile gerçekleştirilen tahminler sonucunda %50,57 doğruluk yüzdesi ve 0,49 F1-skor değerine ulaşılmıştır. Analizlerin toplam işlem süresi 17 saniyedir. Elde edilen bu tahminlerin market simülasyonunda kullanılması sonucunda ise 5 Şirket için 4293\$'lık kar elde edilmiş ve dönem sonunda 54293\$ toplam yatırım değerine ulaşılmıştır. 5 şirket için Al-Tut stratejisi ile 476\$ zarar edilmiş ve dönem sonundaki toplam yatırım değeri 49524\$ olarak hesaplanmıştır. 3. Test döneminde Al-Tut stratejisi ile hisse senedi 476 dolar zarar etmesine rağmen Algoritmanın tahminleri kullanılarak gerçekleştirilen market simülasyonu sonucunda 4293\$ kar elde edilebilmiştir ve toplam yatırım değerinde 54293\$ değerine ulaşmıştır.

Tablo 5.11'deki 2 farklı durum incelendiğinde Belirlenen özniteliklerden sosyal medya verilerinin çıkarılması durumunda elde edilen kar miktarının 4293\$'dan 1387\$'a, toplam yatırım miktarı ise 54293\$'dan 51386\$'a gerilediği görülmektedir. Sonuçlar değerlendirildiğinde alım-satım işlem sayısının sosyal medya verilerinin çıkarıldığı durumda ortalama 134 değerine yükseldiği görülmektedir. Alım-satım işlem sayısının yüksek olması her işlem başına %0.1 oranındaki işlem ücreti nedeniyle karlılık miktarını düşürmektedir. Sonuçlar değerlendirildiğinde 3. Test dönemi için sosyal medya verilerinin kullanılmasının karlılık miktarını arttırdığı söylenebilir.

Tablo 5.12'deki sonuçlar değerlendirildiğinde belirlenen öznitelik değerleri ile SVM algoritması kullanılarak 4293\$ kar elde edilirken ARIMA yöntemi tahminlerinin market simülasyonunda kullanılması sonucunda 12383\$ zarar edildiği görülmektedir. SVM algoritması ile 54293\$'lık toplam yatırım değerine ulaşılırken, ARIMA yöntemiyle 37617\$ toplam yatırım değerine ulaşılabilmiştir. Alım-satım işlem sayılarına bakıldığında SVM algoritmasıyla ortalama 25 adet işlem gerçekleştirilirken, ARIMA

yöntemiyle ortalama 368 adet işlem gerçekleştirilmiştir. İşlem adetinin bu kadar yüksek olması %0.1'lik işlem ücreti nedeniyle ciddi bir kar kaybına yol açmaktadır.

Öznitelik belirleme dönemi için yapılan tahminlerin market simülasyonunda kullanılması sonrasında en yüksek doğruluk yüzdesine ulaşan parametre kombinasyonları Tablo 5.13'te görülebilmektedir.

Tablo 5.13. İkili sınıflandırmada öznitelik seçim dönemi için en yüksek doğruluk yüzdeleri

TF-IDF	Duygu Analiz	Veri Seti	Algoritma	Değişkenler	Duygu Değeri	Zaman Periyod (dk)	Doğruluk (%)	Toplam Kar-Zarar (\$)
Yok	Stanza	Tüm Gönderiler	RF	KCP, Önceki Hareket Yönü, Duygu Değeri 3	Metin ve Emoji	120	54,48	-5434,73
Yok	Stanza	Tüm Gönderiler	DT	KCP, CCI, Duygu Değeri 4	Metin	120	54,48	-4199,97
Yok	Textblob	Sadece Cashtag	RF	KCP, Önceki Hareket Yönü, Duygu Değeri 5	Metin	120	54,04	-6831,37
Yok	Vader	Tüm Gönderiler	DT	KCP, EOM, Duygu Değeri 6	Metin	120	54,04	-4996,24
Var	Textblob	Sadece Cashtag	RF	Stoch RSI, Önceki Hareket Yönü, Duygu Değeri 4	Metin ve Emoji	120	54,03	-6223,18
Yok	Stanza	Sadece Cashtag	DT	Önceki Hareket Yönü, Yazar Sayısı, Duygu Değeri 3	Metin ve Emoji	120	53,93	-7633,39
Yok	Stanza	Tüm Gönderiler	DT	EOM, Duygu Değeri 4	Metin	120	53,92	-10282,6
Var	Vader	Tüm Gönderiler	RF	Stoch RSI, Önceki Hareket Yönü, Duygu Değeri 5	Metin ve Emoji	120	53,88	-2816,92
Var	Stanza	Tüm Gönderiler	RF	CCI, Önceki Hareket Yönü, Duygu Değeri 5	Metin ve Emoji	120	53,82	-2489,79
Var	Textblob	Tüm Gönderiler	RF	EOM, CCI, Yazar Sayısı, Duygu Değeri 3	Metin ve Emoji	120	53,82	-3852,53

En yüksek doğruluk değerine, RF algoritması kullanılarak, 120-120 zaman periyodunda ulaşılmıştır. KCP teknik indikatörü ve önceki hareket yönü ile Duygu değeri 3'ün kullanıldığı bu durum için dört aylık sürede 5 şirket için ortalama %54.48 doğruluk yüzdesine ulaşılabilmektedir. Duygu durum değeri, Stanza kütüphanesi ile tüm gönderilerden oluşan veri seti için hem sosyal medya paylaşımlarının metin duygu değerleri hem de Emojilerin duygu değerlerini ortalaması alınarak kullanılmasıyla elde edilmiştir. Her bir Paylaşımın duygu değeri, yazar etki değeri ile çarpılıp periyod boyunca toplanarak Duygu değeri 3 elde edilmiştir. Sosyal medya paylaşımlarının metin verisine ait kelime vektörünün dahil edilmediği durumda en yüksek doğruluk yüzdesine ulaşılmıştır.

SPSS programı kullanılarak en yüksek ortalama doğruluk yüzdesi olan 54,48 değerine Tek Örneklem T-test'i uygulanarak ortalama anlamlı farklılığı incelenmiştir.

Bu analiz için H_0 ve H_1 hipotezleri aşağıdaki gibidir:

H_0 = En iyi öznitelik kombinasyonunun doğruluk yüzdesi, tüm öznitelik kombinasyonlarının doğruluk yüzdesi ortalamasından farklı değildir

H_1 = En iyi öznitelik kombinasyonunun doğruluk yüzdesi, tüm öznitelik kombinasyonlarının doğruluk yüzdesi ortalamasından farklıdır.

Tablo 5.14'te gösterilen sonuçlar incelendiğinde P(sig) değerinin 0.05 değerinden küçük olması nedeniyle H_0 hipotezi reddedilir ve belirlenen öznitelik kombinasyonuna ait doğruluk yüzdesi değerinin ortalama anlamlı derecede farklı olduğu söylenebilir.

Tablo 5.14. İkili sınıflandırmada ortalama doğruluk yüzdesi değerinin T-testi değerleri

Tek-Örneklem Test						
	Test Değeri = 54.48					
	t	df	Sig. (2-kuyruklu)	Ortalama Fark	Farkın %95'lik Güven Aralığı	
					Alt	Üst
Doğruluk Yüzdesi	-1669,791	89999	,000	-4,093	-4,098	-4,089

Öznitelik belirleme dönemi için elde edilen en iyi değerler karşılaştırıldığında en iyi sonuçların duygu değerinin dahil olduğu öznitelik kombinasyonlarında elde edildiği görülmektedir. Bu durumdan yola çıkarak duygu durum değerinin analizlere dahil edilmesinin doğruluk yüzdesini arttığı söylenebilir. Tablo 5.5'teki değerler ile karşılaştırıldığında yüksek doğruluk yüzdelерinin elde edildiği sonuçlarda toplam kar miktarlarının oldukça düşük olduğu görülmektedir. %54,48 doğruluk yüzdesine ulaşıldığı durumda algoritma tahminleri kullanılarak gerçekleştirilen market simülasyonu sonucunda 5435\$ zarar edilmiştir. Bu sonuçlardan yola çıkarak doğruluk yüzdesinin artmasının karlılığın artmasını sağlamadığı söylenebilir.

Belirlenen en iyi parametreler kullanılarak üç farklı test dönemi için analizler gerçekleştirilmeden önce algoritmaların hiper-parametrelerinin belirlenmesi ve optimize edilmesi için çalışma yürütülmüştür. Her bir algoritma için hiper-parametre alternatifleri belirlenmiştir. Belirlenen hiper-parametre alternatifleri Tablo 5.15'te görülebilmektedir. Daha sonrasında algoritmalar için belirlenen parametreler ait bütün kombinasyonlar

Scikit-Learn kütüphanesindeki GridSearchCV fonksiyonu kullanılarak denenmiş ve en iyi parametre kombinasyonları tespit edilmiştir.

Tablo 5.15. *Algoritma Hiper-parametre kombinasyonları*

Algoritma	Hiper-Parametreler
Random Forest	
Max depth	3, 4, 5, 6, 8, 10, 20, 30
Max leaf nodes	3, 6, 9, 12, 15
Min samples split	2, 4, 6, 8, 10, 20, 30
Min samples leaf	1, 2, 4, 6, 8
N estimators	25, 50, 100
Max features	sqrt, log2, None
Decision Tree	
Max depth	3, 4, 5, 6, 8, 10, 20, 30
Min samples split	2, 4, 6, 8, 10, 20, 30
Min samples leaf	1, 2, 4, 6, 8
Criterion	gini, entropy
Support Vector Machine	
C	0.001, 0.01, 0.1, 1, 10, 100
Solver	newton-cg, lbfgs, sag, saga
Logistic regression	
C	1,10,100,1000
Gamma	1,0.1,0.001,0.0001

Gerçekleştirilen hiper-parametre optimizasyon analizi sonucunda RF algoritması için hiper parametreler Max Depth = 3, Max leaf nodes=3, Min Samples Split=2 ve Max features= Sqrt olarak belirlenmiştir. Belirlenen en iyi hiper-parametreler ve öznitelikler kullanılarak üç farklı test dönemi için yeniden analizler gerçekleştirilmiştir. Seçilen şirketler için algoritma tahminlerinin eğitim ve test dönemleri için doğruluk yüzdeleri, aynı şekilde eğitim ve test dönemleri için F1-skor değerleri, market simülasyonu sonucunda dönem sonunda ulaşılan kar veya zarar miktarı ve market simülasyonu sırasında gerçekleşen alım-satım işlem sayısı hesaplanmıştır.

Her bir test dönemi için 2 farklı durum oluşturulmuş ve karşılaştırılmıştır. İlk durumda en iyi sonucu veren değişken kombinasyonu için sonuçlar elde edilmiştir. İkinci durumda en iyi değişken kombinasyonunun sosyal medya verilerinden elde edilen özniteliklerin çıkarılmış hali için sonuçlar elde edilmiştir. İlk durumda belirlenen KCP

teknik indikatörü, önceki hareket yönü ve Duygu değeri 3 ile RF algoritması kullanılarak 120 dk zaman periyodu için analizler gerçekleştirilmiştir. İkinci durumda ise Duygu değeri 3 analizden çıkarılmış ve sadece KCP teknik indikatörü ve önceki hareket yönü kullanılarak, RF algoritması ile 30 dk zaman periyodu için analiz gerçekleştirilmiştir. Borsanın yükseliş durumunu temsil eden 1. Test dönemi için karşılaştırmalı sonuçlar Tablo 5.17’de görülebilmektedir. Tablo 5.18’de ise birinci test döneminde RF algoritması ile elde edilen sonuçlar ile ARIMA yöntemi kullanılarak elde edilen sonuçların karşılaştırması gösterilmiştir. Tablo 5.17’deki iki farklı durum için RF algoritmasına alternatif olarak derin öğrenme yöntemlerinden LSTM ağları kullanılarak da analizler gerçekleştirilmiş ve sonuçlar Tablo 5.19’da gösterilmiştir. Derin öğrenme ile gerçekleştirilen analizlerde hiper-parametrelerin belirlenmesi için denemeler gerçekleştirilmiştir. Yapılan denemeler sonucunda Tablo 5.16’daki hiper-parametrelere karar verilmiştir. LSTM ağ yapısı 32 nörona sahip 2 adet LSTM ağı katmanından oluşturulmuştur.

Tablo 5.16. *LSTM hiper parametreleri*

Hiper Parametre	Değer
Learning rate	0.01
Optimizer	Adam
Loss	Mean Squared Error
Validation Split	0.1
Batch size	32
Epoch	100
Window size	5

ARIMA yöntemi ve LSTM algoritması ile öncelikle hisse senetlerinin belirlenen öznitelikteki zaman periyodu için getiri tahmini gerçekleştirilmiş ve daha sonrasında ikili sınıflandırma probleminde kullanmak üzere pozitif getiri değerleri 1, negatif getiri değerleri 0 olarak etiketlenmiştir. Bu 2 sınıflı olarak etiketlenen tahminler, market simülasyonunda kullanılmış böylece diğer tablo değerlerinin hesaplaması gerçekleştirilmiştir. Tablolardaki doğruluk yüzdesi, F1-skor değeri, dönem boyunca nakitte kalma oranı ve algoritma ile alım-satım işlem sayısı 5 şirketin ortalaması alınarak gösterilmiştir. Tablolardaki kar-zarar miktarları 5 şirketin değerlerinin toplamı alınarak gösterilmiştir.

Tablo 5.17. İkili sınıflandırmada birinci test dönemi RF algoritması sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Eğitim Doğruluk Yüzdesi (%)	55,44	55,07
Eğitim F1-skor Değeri	0,69	0,68
Test Doğruluk Yüzdesi (%)	53,37	53,15
Test F1-skor Değeri	0,53	0,66
Algoritma ile Elde Edilen Kazanç (\$)	11901,42	16347,45
Algoritma ile Alım-Satım İşlem Sayısı	44	38

Tablo 5.18. İkili sınıflandırmada birinci test dönemi RF- ARIMA yöntem karşılaştırması

	Belirlenen Öznitelikler ile Analiz Sonucu	ARIMA Yöntemi ile Ulaşılan Sonuç
Test Doğruluk Yüzdesi (%)	53,37	49,37
Test F1-skor Değeri	0,53	0,54
Algoritma ile Elde Edilen Kazanç (\$)	11901,42	1649,09
Algoritma ile Alım-Satım İşlem Sayısı	44	129

Tablo 5.19. İkili sınıflandırmada birinci test dönemi LSTM algoritması sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Eğitim Doğruluk Yüzdesi (%)	95,81	91,01
Eğitim F1-skor Değeri	0,96	0,91
Test Doğruluk Yüzdesi (%)	86,89	83,91
Test F1-skor Değeri	0,90	0,87
Algoritma ile Elde Edilen Kazanç (\$)	6686,49	5242,88
Algoritma ile Alım-Satım İşlem Sayısı	139	144

Belirlenen öznitelikler ve hiper-parametre değerleri için RF algoritmasıyla 1. Test dönemi için gerçekleştirilen analizler sonucunda test verisinde %53,37 doğruluk yüzdesine ulaşıldığı Tablo 5.17’de görülmektedir. Test verisinde 0,53 F1-skor değeri ve algoritma tahminleri kullanılarak 11901\$ kar elde edilmiştir. Analizlerin toplam işlem süresi 9 saniyedir. Belirlenen özniteliklerden sosyal medya verilerinin çıkartıldığı ikinci durumda doğruluk yüzdesinin %53,15’e gerilediği ve F1-skor değerinin 0,66’ya yükseldiği görülmektedir. Tablo 5.18’deki iki durum incelendiğinde Belirlenen öznitelikler ve RF algoritması kullanılarak %53.15 test doğruluk yüzdesine ulaşılırken,

ARIMA yöntemiyle %49,37 doğruluk yüzdesine ulaşılabilmiştir. RF algoritması tahminlerinin market simülasyonunda kullanılması sonucunda 38 adet alım-satım işlemi sonucunda 11576\$ kar elde edilebilirken, ARIMA yöntemi sonucunda 129 adet alım-satım işlemi sonucunda 1649\$ kar elde edilebilmiştir.

Tablo 5.19'daki sonuçlar değerlendirildiğinde 1. Test dönemi için belirlen öznitelikler ve LSTM algoritması kullanılarak %86,89 test doğruluk yüzdesine ulaşılmıştır. LSTM algoritmasının kullanılması durumunda, RF algoritmasına kıyasla doğruluk yüzdesinde %33,52 artış sağlanmıştır. LSTM algoritması ile toplam işlem süresi 2 dakika 33 saniyedir. LSTM algoritması ile elde edilen tahminlerin market simülasyonunda kullanılması halinde 139 alım-satım işlemi sonrasında 6686\$ kar elde edilmiştir. Bu sonuçlarda doğruluk oranının artmasının kar miktarında net bir artış sağlamadığını göstermektedir. Belirlenen öznitelikler kullanılarak %86,89 doğruluk yüzdesine ulaşılabilirken, sosyal medya verilerinin kullanılmadığı durumda doğruluk yüzdesi %83,91'e gerilemektedir. Bu sonuçtan yola çıkarak 1. Test dönemi için sosyal medya verilerinin kullanılması halinde daha yüksek doğruluk yüzdesi değerlerine ulaşılabilirdi söylenebilir. LSTM algoritması ile elde edilen tahminleri market simülasyonunda kullanırken alım-satım işlem ücreti olmaması halinde elde edilen kar miktarı 14987\$ olarak gerçekleşmektedir.

Borsanın düşüş dönemini temsil eden İkinci test dönemi için aynı öznitelikler ve parametreler kullanılarak analizler tekrarlanmıştır. Sosyal medya verilerinin olduğu ve sosyal medya verilerinin çıkarıldığı iki durum için RF algoritması ile elde edilen karşılaştırmalı sonuçlar Tablo 5.20'de gösterilmiştir. İkinci test dönemi için RF ile elde edilen sonuçların ARIMA yöntemi sonuçlarıyla karşılaştırılması ise Tablo 5.21'de görülebilmektedir. Sosyal medya verilerinin olduğu ve sosyal medya verilerinin çıkarıldığı iki durum için LSTM algoritmasıyla elde edilen karşılaştırmalı sonuçlar Tablo 5.22'de gösterilmiştir.

Tablo 5.20. İkili sınıflandırmada ikinci test dönemi RF algoritması sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Eğitim Doğruluk Yüzdesi (%)	55,44	55,07
Eğitim F1-skor Değeri	0,69	0,68
Test Doğruluk Yüzdesi (%)	49,73	47,77
Test F1-skor Değeri	0,58	0,62
Algoritma ile Elde Edilen Kazanç (\$)	-12923,64	-11423,81
Algoritma ile Alım-Satım İşlem Sayısı	75	15

Tablo 5.21. İkili sınıflandırmada ikinci test dönemi RF- ARIMA yöntem karşılaştırması

	Belirlenen Öznitelikler ile Analiz Sonucu	ARIMA Yöntemi ile Ulaşılan Sonuç
Test Doğruluk Yüzdesi (%)	49,73	51,01
Test F1-skor Değeri	0,58	0,48
Algoritma ile Elde Edilen Kazanç (\$)	-12923,6	-13775,1
Algoritma ile Alım-Satım İşlem Sayısı	75	118

Tablo 5.22. İkili sınıflandırmada ikinci test dönemi LSTM algoritması sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Eğitim Doğruluk Yüzdesi (%)	91,33	90,9
Eğitim F1-skor Değeri	0,92	0,92
Test Doğruluk Yüzdesi (%)	80,53	74,51
Test F1-skor Değeri	0,75	0,70
Algoritma ile Elde Edilen Kazanç (\$)	-15393,3	-14092,1
Algoritma ile Alım-Satım İşlem Sayısı	146	147

Belirlenen öznitelikler ve hiper-parametre değerleri için RF algoritmasıyla 2. Test dönemi için gerçekleştirilen analizler sonucunda test verisinde %49,73 doğruluk yüzdesine ulaşıldığı Tablo 5.20’de görülmektedir. Test verisinde 0,58 F1-skor değerine ulaşılmış ve algoritma tahminleri kullanılarak 75 alım-satım işlemi sonucunda 12924\$ zarar edilmiştir. Analizlerin toplam işlem süresi 11 saniyedir. Belirlenen özniteliklerden sosyal medya verilerinin çıkartıldığı ikinci durumda doğruluk yüzdesinin %47,77’ya gerilediği ve F1-skor değerinin 0,62’ye yükseldiği görülmektedir. Tablo 5.21’deki iki durum incelendiğinde Belirlenen öznitelikler ve RF algoritması kullanılarak %57,69 test doğruluk yüzdesine ulaşılırken, ARIMA yöntemiyle %49,96 doğruluk yüzdesine

ulaşılabilmektedir. RF algoritması tahminlerinin market simülasyonunda kullanılması sonucunda 75 adet alım-satım işlemi sonucunda 129424\$ zarar edilirken, ARIMA yöntemi sonucunda 118 adet alım-satım işlemi sonucunda 13775\$ zarar edilmiştir.

Tablo 5.22'deki sonuçlar değerlendirildiğinde 2. Test dönemi için belirlen öznitelikler ve LSTM algoritması kullanılarak %80,53 test doğruluk yüzdesine ulaşılmıştır. LSTM algoritmasının kullanılması durumunda, RF algoritmasına kıyasla doğruluk yüzdesinde %30,8 artış sağlanmıştır. LSTM algoritmasıyla toplam işlem süresi 2 dakika 27 saniyedir. LSTM algoritması ile elde edilen tahminlerin market simülasyonunda kullanılması halinde 146 alım-satım işlemi sonrasında 15393\$ zarar edilmiştir. Belirlenen öznitelikler kullanılarak %80,53 doğruluk yüzdesine ulaşılabılırken, sosyal medya verilerinin kullanılmadığı durumda doğruluk yüzdesi %74,51'e gerilemektedir. Bu sonuçtan yola çıkarak 2. Test dönemi için sosyal medya verilerinin kullanılması halinde daha yüksek doğruluk yüzdesi değerlerine ulaşılabildiği söylenebilir. LSTM algoritması ile elde edilen tahminleri market simülasyonunda kullanırken alım-satım işlem ücreti olmaması halinde ulaşılan zarar 10079\$'a gerilemektedir.

Üçüncü test dönemi için aynı öznitelikler ve parametreler kullanılarak analizler tekrarlanmıştır. Sosyal medya verilerinin olduğu ve sosyal medya verilerinin çıkarıldığı iki durum için RF algoritması ile elde edilen karşılaştırmalı sonuçlar Tablo 5.23'te gösterilmiştir. Üçüncü test dönemi için RF ile elde edilen sonuçların ARIMA yöntemi sonuçlarıyla karşılaştırılması ise Tablo 5.24'te görülebilmektedir. Sosyal medya verilerinin olduğu ve sosyal medya verilerinin çıkarıldığı iki durum için LSTM algoritmasıyla elde edilen karşılaştırmalı sonuçlar Tablo 5.25'te gösterilmiştir.

Tablo 5.23. İkili sınıflandırmada üçüncü test dönemi RF algoritması sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Eğitim Doğruluk Yüzdesi (%)	55,44	55,07
Eğitim F1-skör Değeri	0,69	0,68
Test Doğruluk Yüzdesi (%)	50,05	49,4
Test F1-skör Değeri	0,56	0,61
Algoritma ile Elde Edilen Kazanç (\$)	-4389,9	-2540,05
Algoritma ile Alım-Satım İşlem Sayısı	62	25

Tablo 5.24. İkili sınıflandırmada üçüncü test dönemi RF- ARIMA yöntem karşılaştırması

	Belirlenen Öznitelikler ile Analiz Sonucu	ARIMA Yöntemi ile Ulaşılan Sonuç
Test Doğruluk Yüzdesi (%)	50,05	50,67
Test F1-skör Değeri	0,56	0,51
Algoritma ile Elde Edilen Kazanç (\$)	-4389,9	-6248,84
Algoritma ile Alım-Satım İşlem Sayısı	62	108

Tablo 5.25. İkili sınıflandırmada üçüncü test dönemi LSTM algoritması sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Eğitim Doğruluk Yüzdesi (%)	92,15	93,61
Eğitim F1-skör Değeri	0,93	0,94
Test Doğruluk Yüzdesi (%)	86,74	82,9
Test F1-skör Değeri	0,87	0,79
Algoritma ile Elde Edilen Kazanç (\$)	-8851,77	-7653,49
Algoritma ile Alım-Satım İşlem Sayısı	152	142

Belirlenen öznitelikler ve hiper-parametre değerleri için RF algoritmasıyla 3. Test dönemi için gerçekleştirilen analizler sonucunda test verisinde %50,05 doğruluk yüzdesine ulaşıldığı Tablo 5.23'te görülmektedir. Test verisinde 0,56 F1-skör değerine ulaşılmış ve algoritma tahminleri kullanılarak 62 alım-satım işlemi sonucunda 4390\$ zarar edilmiştir. Analizlerin toplam işlem süresi 9 saniyedir. Belirlenen özniteliklerden sosyal medya verilerinin çıkartıldığı ikinci durumda doğruluk yüzdesinin %49,4'e gerilediği ve F1-skör değerinin 0,61'e yükseldiği görülmektedir. Tablo 5.24'teki iki durum incelendiğinde belirlenen öznitelikler ve RF algoritması kullanılarak %55,05 test doğruluk yüzdesine ulaşılırken, ARIMA yöntemiyle %50,78 doğruluk yüzdesine ulaşılabilmiştir. RF algoritması tahminlerinin market simülasyonunda kullanılması sonucunda 62 adet alım-satım işlemi sonucunda 4390\$ zarar edilirken, ARIMA yöntemi sonucunda 108 adet alım-satım işlemi sonucunda 6249\$ zarar edilmiştir.

Tablo 5.25'teki sonuçlar değerlendirildiğinde 3. Test dönemi için belirlen öznitelikler ve LSTM algoritması kullanılarak %86,74 test doğruluk yüzdesine ulaşılmıştır. LSTM algoritmasının kullanılması durumunda, RF algoritmasına kıyasla doğruluk yüzdesinde %31,69 artış sağlanmıştır. LSTM algoritmasıyla toplam işlem süresi 2 dakika 29 saniyedir. LSTM algoritması ile elde edilen tahminlerin market

simülasyonunda kullanılması halinde 152 alım-satım işlemi sonrasında 8852\$ zarar edilmiştir. Belirlenen öznitelikler kullanılarak %86,74 doğruluk yüzdesine ulaşılabilirken, sosyal medya verilerinin kullanılmadığı durumda doğruluk yüzdesi %82,9'a gerilemektedir. Bu sonuçtan yola çıkarak 3. Test dönemi için sosyal medya verilerinin kullanılması halinde daha yüksek doğruluk yüzdesi değerlerine ulaşılabilirdi söylenebilir. LSTM algoritması ile elde edilen tahminleri market simülasyonunda kullanırken alım-satım işlem ücreti olmaması halinde ulaşılan zarar 2122\$'a gerilemektedir.

5.3. Üç Sınıflı Analiz için Sonuçlar

Üç sınıflı analizde öznitelik belirleme dönemi için yapılan tahminlerin market simülasyonunda kullanılması sonrasında en yüksek karlılığı veren parametre kombinasyonları Tablo 5.26'da görülebilmektedir.

Tablo 5.26. Üç sınıflı sınıflandırmada öznitelik seçim dönemi için en yüksek toplam kar değerleri

TF-IDF	Duygu Analiz	Veri Seti	Algo ritma	Değişkenler	Duygu Değeri	Zaman Periyod (dk)	Toplam Kar-Zarar (\$)	Doğruluk (%)
Var	Vader	Tüm Gönderiler	RF	KCP, Önceki Hareket Yönü, Yazar Sayısı, Duygu Değeri 5	Metin ve Emoji	120-120	6161,77	24,53
Var	Vader	Sadece Cashtag	RF	KCP, Duygu Değeri 6	Metin	60-60	5415,77	12,32
Yok	Vader	Sadece Cashtag	SVM	EOM, Duygu Değeri 3	Metin ve Emoji	60-60	4609,20	3,22
Var	Vader	Tüm Gönderiler	RF	Stoch RSI', Duygu Değeri 5	Metin ve Emoji	120-120	4505,72	23,22
Yok	Vader	Tüm Gönderiler	SVM	EOM, Yazar Sayısı, Duygu Değeri 4	Metin	120-120	4487,27	16,12
Yok	Stanza	Sadece Cashtag	SVM	EOM, Yazar Sayısı, Duygu Değeri 5	Metin ve Emoji	120-120	4393,46	22,44
Var	Vader	Tüm Gönderiler	RF	KCP, CCI, Yazar Sayısı, Duygu Değeri 3	Metin ve Emoji	60-60	4352,19	12,32

Tablo 5.26. (Devam) Üç sınıflı sınıflandırmada öznitelik seçim dönemi için en yüksek toplam kar değerleri

Yok	Vader	Sadece Cashtag	SVM	EOM, Yazar Sayısı, Duygu Değeri 6	Metin	120-120	4339,29	23,37
Yok	Vader	Tüm Gönderiler	SVM	KCP, EOM, Yazar Sayısı, Duygu Değeri 6	Metin ve Emoji	60-60	4309,62	8,57
Var	Textblob	Sadece Cashtag	RF	CCI, Stoch RSI, Yazar Sayısı, Duygu Değeri 2	Metin ve Emoji	120-120	4252,12	24,53

En yüksek karlılık değerine, RF algoritması kullanılarak, 120-120 zaman periyodunda ulaşılmıştır. KCP teknik indikatörü, önceki hareket yönü, yazar sayısı ve Duygu değeri 5'in kullanıldığı bu durum için dört aylık sürede 5 şirket için toplam 6162\$ kazanç sağlanmıştır. Duygu durum değeri, Vader kütüphanesi ile tüm gönderilerden oluşan veri seti için hem sosyal medya paylaşımlarının metin duygu değerleri hem de Emojilerin duygu değerlerini ortalaması alınarak kullanılmasıyla elde edilmiştir. Her bir Paylaşımın duygu durum değeri, Post etki değeri ile çarpılıp periyod boyunca toplanarak Duygu değeri 5 elde edilmiştir. Sosyal medya paylaşımlarının metin verisine ait kelime vektörünün dahil edildiği durumda en yüksek karlılık değerine ulaşılmıştır.

SPSS programı kullanılarak en yüksek toplam kar olan 6161,77\$ değerine Tek Örneklem T-test'i uygulanarak ortalama anlamlı farklılığı incelenmiştir. Bu analiz için H_0 ve H_1 hipotezleri aşağıdaki gibidir:

H_0 = En iyi öznitelik kombinasyonunun toplam kar miktarı, tüm öznitelik kombinasyonlarının toplam karlılık değerlerinin ortalamasından farklı değildir

H_1 = En iyi öznitelik kombinasyonunun toplam kar miktarı, tüm öznitelik kombinasyonlarının toplam karlılık değerlerinin ortalamasından farklıdır.

Tablo 5.27'de gösterilen sonuçlar incelendiğinde P(sig) değerinin 0.05 değerinden küçük olması nedeniyle H_0 hipotezi reddedilir ve belirlenen öznitelik kombinasyonuna ait toplam kar değerinin ortalama anlamlı derecede farklı olduğu söylenebilir.

Tablo 5.27. Üç sınıflı sınıflandırmada toplam kar-zarar değerinin T-testi değerleri

Tek-Örneklem Test						
	Test Değeri = 6161.77					
	t	df	Sig. (2-kuyruklu)	Ortalama Fark	Farkın %95'lik Güven Aralığı	
					Alt	Üst
Toplam Kar-Zarar	-481,467	89999	,000	-19366.13	-19444.96	-19287.29

Öznitelik belirleme dönemi için elde edilen en iyi değerler karşılaştırıldığında en iyi sonuçların duygu değerinin dahil olduğu parametre kombinasyonlarında elde edildiği görülmektedir. Bu durumdan yola çıkarak duygu durum değerinin analizlere dahil edilmesinin karlılığı arttığı söylenebilir.

Belirlenen en iyi öznitelikler kullanılarak üç farklı test dönemi için analizler gerçekleştirilmiştir. Seçilen şirketler için market simülasyonu ile test dönemi sonunda ulaşılan kar veya zarar, bu kar ve zarar sonucunda dönem sonunda ulaşılan toplam yatırım değeri, market simülasyonu ile alım-satım işlemlerinin sayısı ve nakitte kalma oranı, algoritma tahminlerinin doğruluk yüzdesi, algoritma tahminlerinin F1-skor değeri, Şirket hisse senetleri ile Al-Tut stratejisi uygulanması halinde hisse senetlerinin Kar-Zarar miktarları ve Al-Tut stratejisi sonucunda hisse senetlerinin dönem sonu toplam yatırım değeri hesaplanmıştır.

Her bir test dönemi için 2 farklı durum oluşturulmuş ve karşılaştırılmıştır. İlk durumda en iyi sonucu veren değişken kombinasyonu için sonuçlar elde edilmiştir. İkinci durumda en iyi değişken kombinasyonunun sosyal medya verilerinden elde edilen özniteliklerin çıkarılmış hali için sonuçlar elde edilmiştir. İlk durumda belirlenen KCP teknik indikatörü, önceki hareket yönü, yazar sayısı ve Duygu değeri 5 değeri ile kelime vektörü de dahil edilip RF algoritması kullanılarak 120 dk zaman periyodu için analizler gerçekleştirilmiştir. İkinci durumda ise Duygu değeri 5, yazar sayısı ve kelime vektörü analizden çıkarılmış ve sadece KCP teknik indikatörü ve önceki hareket yönü kullanılarak RF algoritması ile 120 dk zaman periyodu için analiz gerçekleştirilmiştir. Borsanın yükseliş durumunu temsil eden 1. Test dönemi için karşılaştırmalı sonuçlar Tablo 5.28’de görülebilmektedir. En iyi değere RF algoritması ile ulaşıldığı bilinmektedir. Tablo 5.29’da ise birinci test döneminde RF algoritması ile elde edilen sonuçlar ile ARIMA yöntemi kullanılarak elde edilen sonuçların karşılaştırması gösterilmiştir. ARIMA yöntemi ile öncelikle hisse senetlerinin belirlenen öznitelikteki zaman periyodu için getiri

tahmini gerçekleştirilmiş ve daha sonrasında üç sınıflı sınıflandırma probleminde kullanmak üzere getirinin <0.001 olduğu durumlar -1, >0.001 olduğu durumlar +1 ve getirinin ± 0.001 aralığında olduğu durumlar ise 0 olarak etkilenmiştir. Bu 3 sınıflı olarak etiketlenen ARIMA tahminleri, market simülasyonunda kullanılmış böylece diğer tablo değerlerinin hesaplaması gerçekleştirilmiştir. Tablolardaki kar-zarar miktarları ve toplam yatırım değerleri 5 şirketin değerlerinin toplamı alınarak gösterilmiştir. Tablolardaki dönem boyunca nakitte kalma oranı, algoritma ile alım-satım işlem sayısı, doğruluk yüzdesi ve F1-skor değeri, 5 şirketin ortalaması alınarak gösterilmiştir.

Tablo 5.28. Üç sınıflı sınıflandırmada birinci test dönemi analiz sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Algoritma ile Elde Edilen Kazanç (\$)	16627,93	2800,07
Dönem Sonu Toplam Yatırım Değeri (\$)	66627,93	52800,07
Dönem Boyunca Nakitte Kalma Oranı (%)	11	61
Algoritma ile Alım-Satım İşlem Sayısı	27	143
Doğruluk Oranı (%)	22,26	38,16
F1-Skor Değeri	0,22	0,38
Al-Tut Stratejisi ile Kazanç (\$)	19333,29	19333,29
AL-Tut Dönem Sonu Toplam Yatırım Değer (\$)	69333,29	69333,29

Tablo 5.29. Üç sınıflı sınıflandırmada birinci test dönemi yöntem karşılaştırması

	Belirlenen Öznitelikler ile Analiz Sonucu	ARIMA Yöntemi ile Ulaşılan Sonuç
Algoritma ile Elde Edilen Kazanç (\$)	16627,93	3748,61
Dönem Sonu Toplam Yatırım Değeri (\$)	66627,93	53748,61
Dönem Boyunca Nakitte Kalma Oranı (%)	11	44
Algoritma ile Alım-Satım İşlem Sayısı	27	106
Doğruluk Oranı (%)	22,26	38,33
F1-Skor Değeri	0,22	0,38
Al-Tut Stratejisi ile Kazanç (\$)	19333,29	19197,35
AL-Tut Dönem Sonu Toplam Yatırım Değer (\$)	69333,29	69197,35

Belirlenen Öznitelikler ve RF algoritması kullanılarak 120dk zaman periyodu için elde edilen tahminler sonucunda %22,26 doğruluk yüzdesi elde edilmiştir. Analizlerin toplam işlem süresi 48 saniyedir. Algoritma ile elde edilen tahminlerin market

simülasyonunda kullanılması sonucunda ise 5 Şirket için 16628\$'lık kar elde edilmiş ve dönem sonunda 66628\$ toplam yatırım değerine ulaşılmıştır. 5 şirket için Al-Tut stratejisi ile elde edilen kazanç miktarı 19133\$ ve dönem sonundaki toplam yatırım değeri 69333\$ olarak hesaplanmıştır. Borsanın yükseliş dönemini temsil eden 1. Test döneminde 16628\$ kar elde edilmesine rağmen hisse senetlerinin Al-Tut strateji sonucunda kazandığı değer in altında kazançlar elde edilebilmiştir.

Tablo 5.28'deki 2 farklı durum incelendiğinde Belirlenen özniteliklerden sosyal medya verilerinin çıkarılması durumunda elde edilen kar miktarının 16628\$'dan 2800\$'a, toplam yatırım miktarı ise 66628\$'dan 52800\$'a gerilediği görülmektedir. Sonuçlar değerlendirildiğinde alım-satım işlem sayısı sosyal medya verilerinin çıkarıldığı durumda ortalama 143 adete artış göstermiştir. Alım-satım işlem sayısının yüksek olması her işlem başına %0.1 oranındaki işlem ücreti nedeniyle karlılık miktarını düşürmektedir. Sonuçlar değerlendirildiğinde 1. Test dönemi için sosyal medya verilerinin kullanılmasının karlılık miktarını arttırdığı söylenebilir.

Tablo 5.29'daki sonuçlar değerlendirildiğinde belirlenen öznitelik değerleri ile RF algoritması kullanılarak 16628\$ kar elde edilirken ARIMA yöntemi tahminlerinin market simülasyonunda kullanılması sonucunda 3749\$ kar elde edildiği görülmektedir. RF algoritması ile 66628\$'lık toplam yatırım değerine ulaşılırken, ARIMA yöntemiyle 53749\$ toplam yatırım değerine ulaşılabilmiştir. Alım-satım işlem sayılarına bakıldığında RF algoritmasıyla ortalama 27 adet işlem gerçekleştirilirken, ARIMA yöntemiyle ortalama 106 adet işlem gerçekleştirilmiştir. İşlem adetinin bu kadar yüksek olması %0.1'lik işlem ücreti nedeniyle ciddi bir kar kaybına yol açmaktadır.

Borsanın düşüş dönemini temsil eden İkinci test dönemi için aynı öznitelikler ve parametreler kullanılarak analizler tekrarlanmıştır. Sosyal medya verilerinin olduğu ve sosyal medya verilerinin çıkarıldığı iki durum için karşılaştırmalı sonuçlar Tablo 5.30'da gösterilmiştir. İkinci test dönemi için RF ile elde edilen sonuçların ARIMA yöntemi sonuçlarıyla karşılaştırılması ise Tablo 5.31'de görülebilmektedir.

Tablo 5.30. Üç sınıflı sınıflandırmada ikinci test dönemi analiz sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Algoritma ile Elde Edilen Kazanç (\$)	-8987,16	-9545,22
Dönem Sonu Toplam Yatırım Değeri (\$)	41012,84	40454,78
Dönem Boyunca Nakitte Kalma Oranı (%)	23	61
Algoritma ile Alım-Satım İşlem Sayısı	37	133
Doğruluk Oranı (%)	21,07	45,92
F1-Skor Değeri	0,21	0,46
AI-Tut Stratejisi ile Kazanç (\$)	-11383,98	-11384
AI-Tut Dönem Sonu Toplam Yatırım Değer (\$)	38616,02	38616,02

Tablo 5.31. Üç sınıflı sınıflandırmada ikinci test dönemi yöntem karşılaştırması

	Belirlenen Öznitelikler ile Analiz Sonucu	ARIMA Yöntemi ile Ulaşılan Sonuç
Algoritma ile Elde Edilen Kazanç (\$)	-8987,16	-12708,8
Dönem Sonu Toplam Yatırım Değeri (\$)	41012,84	37291,19
Dönem Boyunca Nakitte Kalma Oranı (%)	23	52
Algoritma ile Alım-Satım İşlem Sayısı	37	103
Doğruluk Oranı (%)	21,07	44,41
F1-Skor Değeri	0,21	0,45
AI-Tut Stratejisi ile Kazanç (\$)	-11384	-11708,5
AI-Tut Dönem Sonu Toplam Yatırım Değeri (\$)	38616,02	38291,48

Belirlenen Öznitelikler ve RF algoritması kullanılarak 120dk zaman periyodu için elde edilen tahminler sonucunda ikinci test döneminde %21,07 doğruluk yüzdesi elde edilebilmiştir. Analizlerin toplam işlem süresi 51 saniyedir. Algoritma ile elde edilen tahminlerin market simülasyonunda kullanılması sonucunda ise 5 Şirket için 8987\$ zarar edilmiş ve dönem sonunda 41013\$ toplam yatırım değeri elde edilebilmiştir. 5 şirket için AI-Tut stratejisi ile elde edilen zarar miktarı ise 11384\$ ve dönem sonundaki toplam yatırım değeri 38616\$ olarak hesaplanmıştır. Borsanın düşüş dönemini temsil eden 2. Test döneminde toplam 8987\$ zarar edilmesine rağmen gerçekleşen zarar hisse senetlerinin AI-Tut strateji sonucunda kaybettiği değerin çok altında kalmıştır.

Tablo 5.30'daki 2 farklı durum incelendiğinde Belirlenen özniteliklerden sosyal medya verilerinin çıkarılması durumunda ulaşılan zarar miktarının 8987\$'dan 9545\$'a çıktığı ve toplam yatırım miktarının 41013\$'dan 40454\$'a gerilediği görülmektedir.

Sonuçlar incelendiğinde alım-satım işlem sayısı ve nakitte kalma oranının, sosyal medya verilerinin çıkarıldığı durumda daha yüksek olduğu görülmektedir. Alım-satım işlem sayısının yüksek olması her işlem başına %0.1 oranındaki işlem ücreti nedeniyle karlılık miktarını düşürmektedir. Sonuçlar değerlendirildiğinde 2. Test dönemi için sosyal medya verilerinin kullanılmasının karlılık miktarını arttırdığı söylenebilir.

Tablo 5.31'deki sonuçlar değerlendirildiğinde belirlenen öznitelik değerleri ile RF algoritması kullanılarak 8987\$ zarar edilirken ARIMA yöntemi tahminlerinin market simülasyonunda kullanılması sonucunda 12709\$ zarar edildiği görülmektedir. RF algoritması ile 41013\$'lık toplam yatırım değeri elde edilirken, ARIMA yöntemiyle 37291\$ toplam yatırım değerine ulaşılabilmiştir. Alım-satım işlem sayılarına bakıldığında SVM algoritmasıyla ortalama 37 adet işlem gerçekleştirilirken, ARIMA yöntemiyle ortalama 103 adet işlem gerçekleştirilmiştir. İşlem adetinin bu kadar yüksek olması %0.1'lik işlem ücreti nedeniyle ciddi bir kar kaybına yol açmaktadır.

Üçüncü test dönemi içinde aynı öznitelikler ve parametreler kullanılarak analizler tekrarlanmıştır. Sosyal medya verilerinin olduğu ve sosyal medya verilerinin çıkarıldığı iki durum için karşılaştırmalı sonuçlar Tablo 5.32'de gösterilmiştir. İkinci test dönemi için RF ile elde edilen sonuçların ARIMA yöntemi sonuçlarıyla karşılaştırılması ise Tablo 5.33'te görülebilmektedir.

Tablo 5.32. Üç sınıflı sınıflandırmada üçüncü test dönemi analiz sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Algoritma ile Elde Edilen Kazanç (\$)	760,72	313,26
Dönem Sonu Toplam Yatırım Değeri (\$)	50760,72	50313,26
Dönem Boyunca Nakitte Kalma Oranı (%)	19	58
Algoritma ile Alım-Satım İşlem Sayısı	29	126
Doğruluk Oranı (%)	20,27	43,02
F1-Skor Değeri	0,20	0,43
Al-Tut Stratejisi ile Kazanç (\$)	-609,83	-609,83
AL-Tut Dönem Sonu Toplam Yatırım Değer (\$)	49390,17	49390,17

Tablo 5.33. Üç sınıflı sınıflandırmada üçüncü test dönemi yöntem karşılaştırması

	Belirlenen Öznitelikler ile Analiz Sonucu	ARIMA Yöntemi ile Ulaşılan Sonuç
Algoritma ile Elde Edilen Kazanç (\$)	760,72	-5375,8
Dönem Sonu Toplam Yatırım Değeri (\$)	50760,72	44624,2
Dönem Boyunca Nakitte Kalma Oranı (%)	19	49
Algoritma ile Alım-Satım İşlem Sayısı	29	97
Doğruluk Oranı (%)	20,27	43,83
F1-Skor Değeri	0,20	0,44
Al-Tut Stratejisi ile Kazanç (\$)	-609,83	-379,05
AL-Tut Dönem Sonu Toplam Yatırım Değer (\$)	49390,17	49620,95

Belirlenen Öznitelikler ile gerçekleştirilen tahminler sonucunda %20,27 doğruluk yüzdesi elde edilmiştir. Analizlerin toplam işlem süresi 45 saniyedir. Algoritma ile elde edilen tahminlerin market simülasyonunda kullanılması sonucunda ise 5 Şirket için 761\$'lık kar elde edilmiş ve dönem sonunda 50761\$ toplam yatırım değerine ulaşılmıştır. 5 şirket için Al-Tut stratejisi ile 610\$ zarar edilmiş ve dönem sonundaki toplam yatırım değeri 49390\$ olarak hesaplanmıştır. 3. Test döneminde Al-Tut stratejisi ile hisse senedi 610 dolar zarar etmesine rağmen Algoritmanın tahminleri kullanılarak gerçekleştirilen market simülasyonu sonucunda 761\$ kar elde edilebilmiştir ve toplam yatırım değerinde 50761\$ değerine ulaşmıştır.

Tablo 5.32'deki 2 farklı durum incelendiğinde Belirlenen özniteliklerden sosyal medya verilerinin çıkarılması durumunda elde edilen kar miktarının 761\$'dan 313\$'a, toplam yatırım miktarı ise 50761\$'dan 50313\$'a gerilediği görülmektedir. Sonuçlar değerlendirildiğinde alım-satım işlem sayısının sosyal medya verilerinin çıkarıldığı durumda ortalama 126 değerine yükseldiği görülmektedir. Alım-satım işlem sayısının yüksek olması her işlem başına %0.1 oranındaki işlem ücreti nedeniyle karlılık miktarını düşürmektedir. Sonuçlar değerlendirildiğinde 3. Test dönemi için sosyal medya verilerinin kullanılmasının karlılık miktarını arttırdığı söylenebilir.

Tablo 5.33'teki sonuçlar değerlendirildiğinde belirlenen öznitelik değerleri ile RF algoritması kullanılarak 761\$ kar elde edilirken ARIMA yöntemi tahminlerinin market simülasyonunda kullanılması sonucunda 5375\$ zarar edildiği görülmektedir. RF algoritması ile 50761\$'lık toplam yatırım değerine ulaşılırken, ARIMA yöntemiyle 44624\$ toplam yatırım değerine ulaşılabilmiştir. Alım-satım işlem sayılarına bakıldığında RF algoritmasıyla ortalama 29 adet işlem gerçekleştirilirken, ARIMA

yöntemiyle ortalama 97 adet işlem gerçekleştirilmiştir. İşlem adetinin bu kadar yüksek olması %0.1'lik işlem ücreti nedeniyle ciddi bir kar kaybına yol açmaktadır.

Öznitelik belirleme dönemi için yapılan tahminlerin market simülasyonunda kullanılması sonrasında en yüksek doğruluk yüzdesine ulaşan parametre kombinasyonları Tablo 5.34.'de görülebilmektedir.

Tablo 5.34. Üç sınıflı sınıflandırmada öznitelik seçim dönemi için en yüksek doğruluk yüzdeleri

TF-IDF	Duygu Analiz	Veri Seti	Algo ritma	Değişkenler	Duygu Değeri	Zaman Periyod (dk)	Doğruluk (%)	Toplam Kar-Zarar (\$)
Yok	Stanza	Sadece Cashtag	LR	KCP, Önceki Hareket Yönü, Duygu Değeri 4	Metin	120	52,33	-7413,31
Yok	Textblob	Sadece Cashtag	LR	EOM, CCI, Duygu Değeri 4	Metin ve Emoji	120	52,33	-5789,48
Yok	Stanza	Sadece Cashtag	LR	EOM, Yazar Sayısı, Duygu Değeri 2	Metin	120	52,32	-5597,96
Yok	Vader	Tüm Gönderiler	LR	CCI, Önceki Hareket Yönü, Duygu Değeri 4	Metin ve Emoji	120	52,28	-3222,96
Yok	Textblob	Tüm Gönderiler	LR	EOM, Önceki Hareket Yönü, Duygu Değeri 4	Metin	120	52,28	-3010,36
Yok	Textblob	Tüm Gönderiler	LR	Önceki Hareket Yönü, Duygu Değeri 2	Metin	120	52,28	-2919,72
Yok	Textblob	Sadece Cashtag	LR	KCP, EOM, Yazar Sayısı, Duygu Değeri 3	Metin ve Emoji	120	52,28	-2214,1
Yok	Stanza	Sadece Cashtag	LR	KCP, EOM, Yazar Sayısı, Duygu Değeri 4	Metin	120	52,28	-2761,4
Yok	Stanza	Sadece Cashtag	LR	KCP, Önceki Hareket Yönü, Duygu Değeri 4	Metin ve Emoji	120	52,28	-7266,42
Yok	Textblob	Sadece Cashtag	LR	EOM, CCI, Duygu Değeri 4	Metin	120	52,27	-5798,29

En yüksek doğruluk değerine, LR algoritması kullanılarak, 120-120 zaman periyodunda ulaşılmıştır. KCP teknik indikatörü ve önceki hareket yönü ile Duygu değeri

4'ün kullanıldığı bu durum için dört aylık sürede 5 şirket için ortalama %52,33 doğruluk yüzdesine ulaşılabilmektedir. Duygu durum değeri, Stanza kütüphanesi ile sadece cashtag içeren paylaşımlardan oluşan veri seti için sosyal medya paylaşımlarının metin verileri kullanılarak elde edilmiştir. Her bir Paylaşımın duygu değeri, yazar etki değeri ile çarpılıp periyod boyunca toplanıp paylaşım sayısına bölünerek Duygu değeri 3 elde edilmiştir. Sosyal medya paylaşımlarının metin verisine ait kelime vektörünün dahil edilmediği durumda en yüksek doğruluk yüzdesine ulaşılmıştır.

SPSS programı kullanılarak en yüksek ortalama doğruluk yüzdesi olan 52,33 değerine Tek Örneklem T-test'i uygulanarak ortalamadan anlamlı farklılığı incelenmiştir. Bu analiz için H_0 ve H_1 hipotezleri aşağıdaki gibidir:

H_0 = En iyi öznitelik kombinasyonunun doğruluk yüzdesi, tüm öznitelik kombinasyonlarının doğruluk yüzdesi ortalamasından farklı değildir

H_1 = En iyi öznitelik kombinasyonunun doğruluk yüzdesi, tüm öznitelik kombinasyonlarının doğruluk yüzdesi ortalamasından farklıdır.

Tablo 5.35'te gösterilen sonuçlar incelendiğinde P(sig) değerinin 0.05 değerinden küçük olması nedeniyle H_0 hipotezi reddedilir ve belirlenen öznitelik kombinasyonuna ait doğruluk yüzdesi değerinin ortalamadan anlamlı derecede farklı olduğu söylenebilir.

Tablo 5.35. Üç sınıflı sınıflandırmada ortalama doğruluk yüzdesi değerlerinin T-testi değerleri

Tek-Örneklem Test						
	Test Değeri = 52.33					
	t	df	Sig. (2-kuyruklu)	Ortalama Fark	Farkın %95'lik Güven Aralığı	
					Alt	Üst
Doğruluk Yüzdesi	-489,374	89999	,000	-21,887	-21,974	-21,799

Öznitelik belirleme dönemi için elde edilen en iyi değerler karşılaştırıldığında en iyi sonuçların duygu değerinin dahil olduğu öznitelik kombinasyonlarında elde edildiği görülmektedir. Bu durumdan yola çıkarak duygu durum değerinin analizlere dahil edilmesinin doğruluk yüzdesini arttığı söylenebilir. Tablo 5.26'daki değerler ile karşılaştırıldığında yüksek doğruluk yüzdelерinin elde edildiği sonuçlarda toplam kar miktarlarının oldukça düşük olduğu görülmektedir. %52,33 doğruluk yüzdesine ulaşıldığı durumda algoritma tahminleri kullanılarak gerçekleştirilen market simülasyonu

sonucunda 7413\$ zarar edilmiştir. Bu sonuçlardan yola çıkarak doğruluk yüzdesinin artmasının karlılığın artmasını sağlamadığı söylenebilir.

Belirlenen en iyi parametreler kullanılarak üç farklı test dönemi için analizler gerçekleştirilmeden önce algoritmaların hiper-parametrelerinin belirlenmesi ve optimize edilmesi için iki sınıflı sınıflandırma analizindeki benzer çalışma yürütülmüştür. Optimizasyon için Tablo 5.15.'de hiper-parametre alternatifleri aynı şekilde kullanılmıştır. Daha sonrasında algoritmalar için belirlenen parametreler ait bütün kombinasyonlar Scikit-Learn kütüphanesindeki GridSearchCV fonksiyonu kullanılarak denenmiş ve en iyi parametre kombinasyonları tespit edilmiştir.

Gerçekleştirilen hiper-parametre optimizasyon analizi sonucunda LR algoritması için hiper parametreler $C=1$ ve Solver= newton-cg olarak belirlenmiştir. Belirlenen en iyi hiper-parametreler ve öznitelikler kullanılarak üç farklı test dönemi için yeniden analizler gerçekleştirilmiştir. Seçilen şirketler için algoritma tahminlerinin eğitim ve test dönemleri için doğruluk yüzdeleri, aynı şekilde eğitim ve test dönemleri için F1-skor değerleri, market simülasyonu sonucunda dönem sonunda ulaşılan kar veya zarar miktarı ve market simülasyonu sırasında gerçekleşen alım-satım işlem sayısı hesaplanmıştır.

Her bir test dönemi için 2 farklı durum oluşturulmuş ve karşılaştırılmıştır. İlk durumda en iyi sonucu veren değişken kombinasyonu için sonuçlar elde edilmiştir. İkinci durumda en iyi değişken kombinasyonunun sosyal medya verilerinden elde edilen özniteliklerin çıkarılmış hali için sonuçlar elde edilmiştir. İlk durumda belirlenen KCP teknik indikatörü, önceki hareket yönü ve Duygu değeri 4 ile LR algoritması kullanılarak 120 dk zaman periyodu için analizler gerçekleştirilmiştir. İkinci durumda ise Duygu değeri 4 analizden çıkarılmış ve sadece KCP teknik indikatörü ve önceki hareket yönü kullanılarak, LR algoritması ile 120 dk zaman periyodu için analiz gerçekleştirilmiştir. Borsanın yükseliş durumunu temsil eden 1. Test dönemi için karşılaştırmalı sonuçlar Tablo 5.36'da görülebilmektedir. Tablo 5.37'de ise birinci test döneminde LR algoritması ile elde edilen sonuçlar ile ARIMA yöntemi kullanılarak elde edilen sonuçların karşılaştırması gösterilmiştir. Tablo 5.36'deki iki farklı durum için LR algoritmasına alternatif olarak derin öğrenme yöntemlerinden LSTM ağırları kullanılarak da analizler gerçekleştirilmiş ve sonuçlar Tablo 5.38'de gösterilmiştir. Derin öğrenme ile gerçekleştirilen analizlerde ikili sınıflandırma analizindeki ağ yapısı ve hiper parametreler üç sınıflı sınıflandırma analizinde de aynen kullanılmıştır.

ARIMA yöntemi ve LSTM algoritması ile öncelikle hisse senetlerinin belirlenen öznitelikteki zaman periyodu için getiri tahmini gerçekleştirilmiş ve daha sonrasında üç sınıflı sınıflandırma probleminde kullanmak üzere getirinin <0.001 olduğu durumlar -1, >0.001 olduğu durumlar +1 ve getirinin ± 0.001 aralığında olduğu durumlar ise 0 olarak etkilendirilmiştir. Bu 3 sınıflı olarak etiketlenen tahminler, market simülasyonunda kullanılmış böylece diğer tablo değerlerinin hesaplaması gerçekleştirilmiştir. Tablolardaki doğruluk yüzdesi, F1-skor değeri, dönem boyunca nakitte kalma oranı ve algoritma ile alım-satım işlem sayısı 5 şirketin ortalaması alınarak gösterilmiştir. Tablolardaki kar-zarar miktarları 5 şirketin değerlerinin toplamı alınarak gösterilmiştir.

Tablo 5.36. Üç sınıflı sınıflandırmada birinci test dönemi LR algoritması sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Eğitim Doğruluk Yüzdesi (%)	48,41	48,74
Eğitim F1-skor Değeri	0,32	0,37
Test Doğruluk Yüzdesi (%)	47,12	45,93
Test F1-skor Değeri	0,30	0,35
Algoritma ile Elde Edilen Kazanç (\$)	11304,45	12661,49
Algoritma ile Alım-Satım İşlem Sayısı	81	77

Tablo 5.37. Üç sınıflı sınıflandırmada birinci test dönemi LR- ARIMA yöntem karşılaştırması

	Belirlenen Öznitelikler ile Analiz Sonucu	ARIMA Yöntemi ile Ulaşılan Sonuç
Test Doğruluk Yüzdesi (%)	47,12	38,33
Test F1-skor Değeri	0,30	0,39
Algoritma ile Elde Edilen Kazanç (\$)	11304,45	3748,61
Algoritma ile Alım-Satım İşlem Sayısı	81	106

Tablo 5.38. Üç sınıflı sınıflandırmada birinci test dönemi LSTM algoritması sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Eğitim Doğruluk Yüzdesi (%)	88,18	82,81
Eğitim F1-skor Değeri	0,87	0,81
Test Doğruluk Yüzdesi (%)	78,53	76,49
Test F1-skor Değeri	0,73	0,72
Algoritma ile Elde Edilen Kazanç (\$)	6726,93	5868,03
Algoritma ile Alım-Satım İşlem Sayısı	130	135

Belirlenen öznitelikler ve hiper-parametre değerleri için LR algoritmasıyla 1. Test dönemi için gerçekleştirilen analizler sonucunda test verisinde %47,12 doğruluk yüzdesine ulaşıldığı Tablo 5.36’da görülmektedir. Test verisinde 0,30 F1-skor değeri ve algoritma tahminleri kullanılarak 11304\$ kar elde edilmiştir. Analizlerin toplam işlem süresi 6 saniyedir. Belirlenen özniteliklerden sosyal medya verilerinin çıkartıldığı ikinci durumda doğruluk yüzdesinin %45,93’e gerilediği görülmektedir. Tablo 5.37’deki iki durum incelendiğinde belirlenen öznitelikler ve LR algoritması kullanılarak %47,12 test doğruluk yüzdesine ulaşılırken, ARIMA yöntemiyle %38,33 doğruluk yüzdesine ulaşılabilmiştir. LR algoritması tahminlerinin market simülasyonunda kullanılması sonucunda 81 adet alım-satım işlemi sonucunda 11304\$ kar elde edilebilirken, ARIMA yöntemi sonucunda 106 adet alım-satım işlemi sonucunda 3748\$ kar elde edilebilmiştir.

Tablo 5.38’deki sonuçlar değerlendirildiğinde 1. Test dönemi için belirlenen öznitelikler ve LSTM algoritması kullanılarak %78,53 test doğruluk yüzdesine ulaşılmıştır. LSTM algoritmasının kullanılması durumunda, LR algoritmasına kıyasla doğruluk yüzdesinde %31,41 artış sağlanmıştır. LSTM algoritmasıyla toplam işlem süresi 2 dakika 27 saniyedir. LSTM algoritması ile elde edilen tahminlerin market simülasyonunda kullanılması halinde 130 alım-satım işlemi sonrasında 6727\$ kar elde edilmiştir. Bu sonuçlarda doğruluk oranının artmasının kar miktarında net bir artış sağlamadığını göstermektedir. Belirlenen öznitelikler kullanılarak %78,53 doğruluk yüzdesine ulaşılabilirken, sosyal medya verilerinin kullanılmadığı durumda doğruluk yüzdesi %76,49’a gerilemektedir. Bu sonuçtan yola çıkarak 1. Test dönemi için sosyal medya verilerinin kullanılması halinde daha yüksek doğruluk yüzdesi değerlerine ulaşılabilirdiği söylenebilir. LSTM algoritması ile elde edilen tahminleri market

simülasyonunda kullanırken alım-satım işlem ücreti olmaması halinde elde edilen kar miktarı 149540\$ olarak gerçekleşmektedir.

Borsanın düşüş dönemini temsil eden İkinci test dönemi için aynı öznitelikler ve parametreler kullanılarak analizler tekrarlanmıştır. Sosyal medya verilerinin olduğu ve sosyal medya verilerinin çıkarıldığı iki durum için LR algoritması ile elde edilen karşılaştırmalı sonuçlar Tablo 5.39’da gösterilmiştir. İkinci test dönemi için LR ile elde edilen sonuçların ARIMA yöntemi sonuçlarıyla karşılaştırılması ise Tablo 5.40’ta görülebilmektedir. Sosyal medya verilerinin olduğu ve sosyal medya verilerinin çıkarıldığı iki durum için LSTM algoritmasıyla elde edilen karşılaştırmalı sonuçlar Tablo 5.41’de gösterilmiştir.

Tablo 5.39. Üç sınıflı sınıflandırmada ikinci test dönemi LR algoritması sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Eğitim Doğruluk Yüzdesi (%)	48,68	48,66
Eğitim F1-skör Değeri	0,37	0,37
Test Doğruluk Yüzdesi (%)	47,45	44,84
Test F1-skör Değeri	0,36	0,34
Algoritma ile Elde Edilen Kazanç (\$)	-2791,05	-12617,1
Algoritma ile Alım-Satım İşlem Sayısı	39	63

Tablo 5.40. Üç sınıflı sınıflandırmada ikinci test dönemi LR- ARIMA yöntem karşılaştırması

	Belirlenen Öznitelikler ile Analiz Sonucu	ARIMA Yöntemi ile Ulaşılan Sonuç
Test Doğruluk Yüzdesi (%)	47,45	44,41
Test F1-skör Değeri	0,36	0,45
Algoritma ile Elde Edilen Kazanç (\$)	-2791,05	-12708,8
Algoritma ile Alım-Satım İşlem Sayısı	39	103

Tablo 5.41. Üç sınıflı sınıflandırmada ikinci test dönemi LSTM algoritması sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Eğitim Doğruluk Yüzdesi (%)	84,5	86,17
Eğitim F1-skor Değeri	0,84	0,85
Test Doğruluk Yüzdesi (%)	69,06	63,7
Test F1-skor Değeri	0,67	0,63
Algoritma ile Elde Edilen Kazanç (\$)	-15521,9	-12408,9
Algoritma ile Alım-Satım İşlem Sayısı	128	122

Belirlenen öznitelikler ve hiper-parametre değerleri için LR algoritmasıyla 2. Test dönemi için gerçekleştirilen analizler sonucunda test verisinde %47,45 doğruluk yüzdesine ulaşıldığı Tablo 5.9’da görülmektedir. Test verisinde 0,36 F1-skor değerine ulaşılmış ve algoritma tahminleri kullanılarak 39 alım-satım işlemi sonucunda 2794\$ zarar edilmiştir. Analizlerin toplam işlem süresi 6 saniyedir. Belirlenen özniteliklerden sosyal medya verilerinin çıkartıldığı ikinci durumda doğruluk yüzdesinin %44,84’e gerilediği görülmektedir. Tablo 5.40’taki iki durum incelendiğinde Belirlenen öznitelikler ve LR algoritması kullanılarak %45,47 test doğruluk yüzdesine ulaşılırken, ARIMA yöntemiyle %44,41 doğruluk yüzdesine ulaşılabilmiştir. LR algoritması tahminlerinin market simülasyonunda kullanılması sonucunda 39 adet alım-satım işlemi sonucunda 2791\$ zarar edilirken, ARIMA yöntemi sonucunda 103 adet alım-satım işlemi sonucunda 12709\$ zarar edilmiştir.

Tablo 5.41’deki sonuçlar değerlendirildiğinde 2. Test dönemi için belirlen öznitelikler ve LSTM algoritması kullanılarak %69,06 test doğruluk yüzdesine ulaşılmıştır. LSTM algoritmasının kullanılması durumunda, LR algoritmasına kıyasla doğruluk yüzdesinde %23,6 artış sağlanmıştır. LSTM algoritmasıyla toplam işlem süresi 2 dakika 25 saniyedir. LSTM algoritması ile elde edilen tahminlerin market simülasyonunda kullanılması halinde 146 alım-satım işlemi sonrasında 15521\$ zarar edilmiştir. Belirlenen öznitelikler kullanılarak %69,06 doğruluk yüzdesine ulaşılabilirken, sosyal medya verilerinin kullanılmadığı durumda doğruluk yüzdesi %63,7’ye gerilemektedir. Bu sonuçtan yola çıkarak 2. Test dönemi için sosyal medya verilerinin kullanılması halinde daha yüksek doğruluk yüzdesi değerlerine ulaşılabilirdiği söylenebilir. LSTM algoritması ile elde edilen tahminleri market simülasyonunda

kullanırken alım-satım işlem ücreti olmaması halinde ulaşılan zarar 10594\$'a gerilemektedir.

Üçüncü test dönemi için aynı öznitelikler ve parametreler kullanılarak analizler tekrarlanmıştır. Sosyal medya verilerinin olduğu ve sosyal medya verilerinin çıkarıldığı iki durum için LR algoritması ile elde edilen karşılaştırmalı sonuçlar Tablo 5.42'de gösterilmiştir. Üçüncü test dönemi için LR ile elde edilen sonuçların ARIMA yöntemi sonuçlarıyla karşılaştırılması ise Tablo 5.43'te görülebilmektedir. Sosyal medya verilerinin olduğu ve sosyal medya verilerinin çıkarıldığı iki durum için LSTM algoritmasıyla elde edilen karşılaştırmalı sonuçlar Tablo 5.44'te gösterilmiştir.

Tablo 5.42. Üç sınıflı sınıflandırmada üçüncü test dönemi LR algoritması sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Eğitim Doğruluk Yüzdesi (%)	48,83	48,41
Eğitim F1-skor Değeri	0,37	0,32
Test Doğruluk Yüzdesi (%)	45,63	44,99
Test F1-skor Değeri	0,35	0,28
Algoritma ile Elde Edilen Kazanç (\$)	-3814,24	-3019,64
Algoritma ile Alım-Satım İşlem Sayısı	59	56

Tablo 5.43. Üç sınıflı sınıflandırmada üçüncü test dönemi LR- ARIMA yöntem karşılaştırması

	Belirlenen Öznitelikler ile Analiz Sonucu	ARIMA Yöntemi ile Ulaşılan Sonuç
Test Doğruluk Yüzdesi (%)	45,63	43,83
Test F1-skor Değeri	0,35	0,44
Algoritma ile Elde Edilen Kazanç (\$)	-3814,24	-5375,8
Algoritma ile Alım-Satım İşlem Sayısı	59	97

Tablo 5.44. Üç sınıflı sınıflandırmada üçüncü test dönemi LSTM algoritması sonuçları

	Belirlenen Öznitelikler ile Analiz Sonucu	Sosyal Medya Verilerinin Çıkarıldığı Durum
Eğitim Doğruluk Yüzdesi (%)	88,59	86,23
Eğitim F1-skor Değeri	0,88	0,85
Test Doğruluk Yüzdesi (%)	80,66	75,1
Test F1-skor Değeri	0,80	0,74
Algoritma ile Elde Edilen Kazanç (\$)	-5696,66	-4479,09
Algoritma ile Alım-Satım İşlem Sayısı	151	109

Belirlenen öznitelikler ve hiper-parametre değerleri için LR algoritmasıyla 3. Test dönemi için gerçekleştirilen analizler sonucunda test verisinde %45,63 doğruluk yüzdesine ulaşıldığı Tablo 5.42’de görülmektedir. Test verisinde 0,35 F1-skor değerine ulaşılmış ve algoritma tahminleri kullanılarak 59 alım-satım işlemi sonucunda 3814\$ zarar edilmiştir. Analizlerin toplam işlem süresi 6 saniyedir. Belirlenen özniteliklerden sosyal medya verilerinin çıkartıldığı ikinci durumda doğruluk yüzdesinin %44,99’a gerilediği görülmektedir. Tablo 5.43’teki iki durum incelendiğinde belirlenen öznitelikler ve LR algoritması kullanılarak %45,63 test doğruluk yüzdesine ulaşılırken, ARIMA yöntemiyle %43,83 doğruluk yüzdesine ulaşılabilmiştir. LR algoritması tahminlerinin market simülasyonunda kullanılması sonucunda 59 adet alım-satım işlemi sonucunda 3814\$ zarar edilirken, ARIMA yöntemi sonucunda 97 adet alım-satım işlemi sonucunda 5375\$ zarar edilmiştir.

Tablo 5.44’teki sonuçlar değerlendirildiğinde 3. Test dönemi için belirlenen öznitelikler ve LSTM algoritması kullanılarak %80,66 test doğruluk yüzdesine ulaşılmıştır. LSTM algoritmasının kullanılması durumunda, LR algoritmasına kıyasla doğruluk yüzdesinde %35,03 artış sağlanmıştır. LSTM algoritmasıyla toplam işlem süresi 2 dakika 26 saniyedir. LSTM algoritması ile elde edilen tahminlerin market simülasyonunda kullanılması halinde 151 alım-satım işlemi sonrasında 5697\$ zarar edilmiştir. Belirlenen öznitelikler kullanılarak %80,66 doğruluk yüzdesine ulaşılabilirken, sosyal medya verilerinin kullanılmadığı durumda doğruluk yüzdesi %75,1’e gerilemektedir. Bu sonuçtan yola çıkarak 3. Test dönemi için sosyal medya verilerinin kullanılması halinde daha yüksek doğruluk yüzdesi değerlerine ulaşılabilirdiği söylenebilir. LSTM algoritması ile elde edilen tahminleri market simülasyonunda kullanırken alım-satım işlem ücreti olmaması halinde 1491\$ kar elde edilmektedir.

5.4. Analiz Sonuçlarının Karşılaştırılması

İkili ve üç sınıflı sınıflandırma analizi kullanılarak elde edilen tahminler kullanılarak gerçekleştirilen market simülasyonu sonucunda elde edilen toplam kar-zarar değerleri ve diğer çıktıların, üç farklı test dönemi için karşılaştırması Tablo 5.45’te gösterilmiştir.

Tablo 5.45. Toplam kar-zarar değeri için test dönemlerinin karşılaştırması

	Algoritma Kazanç (\$)	Toplam Yatırım Değeri (\$)	Nakit Kalma Oranı (%)	Alım-Satım İşlem Sayısı	Doğruluk Oranı (%)	F1-Skor Değeri	Al-Tut Kazanç (\$)	AL-Tut Toplam Yatırım Değeri (\$)
İki sınıflı 1. Test Dönemi	12304,97	62304,97	10	41	51,48	0,65	19320,93	69320,93
İki sınıflı 2. Test Dönemi	-4575,88	45424,12	34	27	49,55	0,47	-11200,3	38799,74
İki sınıflı 3. Test Dönemi	4292,92	54292,92	34	25	50,57	0,49	-476,23	49523,77
Üç sınıflı 1. Test dönemi	16627,93	66627,93	11	27	22,26	0,22	19333,29	69333,29
Üç sınıflı 2. Test dönemi	-8987,16	41012,84	23	37	21,07	0,21	-11384	38616,02
Üç sınıflı 3. Test dönemi	760,72	50760,72	19	29	20,27	0,20	-609,83	49390,17

Elde edilen sonuçları gösteren Tablo 5.45. incelendiğinde, ikinci ve üçüncü test dönemlerinde algoritma tahminleri kullanılmasıyla, hisse senetlerinin alınıp tutulmasına kıyasla daha yüksek toplam yatırım değerine ulaşılmış fakat birinci test döneminde daha altında kalınmıştır. Tablodaki ortalama nakit kalma oranlarının %10-%34 aralığında değiştiği görülmektedir. Market simülasyonu sırasında gerçekleştirilen alım-satım sayılarının ortalamalarının ise 25-41 aralığında gerçekleştiği görülmektedir.

Öznitelik belirleme dönemi için en yüksek toplam kar değerine ulaşılan öznitelik kombinasyonları incelendiğinde, ikili analizde KCP teknik indikatörleri ile Duygu değeri 5 değişkeni en iyi sonucu verirken, üç sınıflı analizde KCP teknik indikatörü, önceki hareket yönü, yazar sayısı ve Duygu değeri 5 değişkenleri en iyi sonucu vermektedir. Her iki durumda da yazar etki ve post etki değerlerini içeren duygu durum değişkenlerinde en iyi sonuca ulaşıldığı görülmektedir. Her iki durumda da KCP teknik indikatörünün kullanıldığı durumda en yüksek toplam kar değerine ulaşılmıştır. Her iki durumda da tüm gönderilerden oluşan veri seti ile en iyi sonuca ulaşılmıştır. İkili analizde kelime vektörünün dahil edilmediği durumda en iyi sonuca ulaşılrken, üç sınıflı analizde kelime vektörünün dahil edildiği durumda en iyi sonuca ulaşılmıştır. İki sınıflı analizde 30dk zaman periyodunda en yüksek kar değerine ulaşılrken, üç sınıflı analizde 120 dk zaman periyodunda en yüksek kar değerine ulaşılmıştır.

Toplam kar değeri açısından değerlendirildiğinde, birinci test dönemi için üç sınıflı analizde daha yüksek toplam yatırım değerine ulaşılırken, ikinci ve üçüncü test dönemlerinde iki sınıflı analiz sonucunda daha yüksek toplam yatırım değerlerine ulaşıldığı görülmüştür.

Tüm sonuçlar göz önüne alındığında borsanın yükseliş dönemlerinde algoritma tahminlerinin yatırım için elverişli olmadığı fakat borsanın düşüş ve daha dengeli olduğu dönemlerde algoritma tahminlerinin yatırımda kullanılmasını daha yüksek toplam yatırım değeri sağladığı görülmektedir.

İkili ve üç sınıflı sınıflandırma analizi için makine öğrenmesi algoritmaları ile elde edilen tahminlerinin doğruluk yüzdelerinin karşılaştırması Tablo 5.46’da, LSTM ağırları ile elde edilen tahminlerin doğruluk yüzdelerinin karşılaştırması ise Tablo 5.47’de gösterilmiştir.

Tablo 5.46. *ML algoritması ile ortalama doğruluk değerleri için test dönemlerinin karşılaştırması*

	Eğitim Doğruluk Yüzdesi (%)	Eğitim F1-skor Değeri	Test Doğruluk Yüzdesi (%)	Test F1-skor Değeri	Algoritma ile Elde Edilen Kazanç (\$)	Alım-Satım İşlem Sayısı
İki sınıflı 1. Test Dönemi	55,44	0,69	53,37	0,53	11901,42	44
İki sınıflı 2. Test Dönemi	55,44	0,69	49,73	0,58	-12923,6	75
İki sınıflı 3. Test Dönemi	55,44	0,69	50,05	0,56	-4389,9	62
Üç sınıflı 1. Test dönemi	48,41	0,32	47,12	0,30	11304,45	81
Üç sınıflı 2. Test dönemi	48,68	0,37	47,45	0,36	-2791,05	39
Üç sınıflı 3. Test dönemi	48,83	0,37	45,63	0,35	-3814,24	59

Tablo 5.47. *LSTM algoritması ile ortalama doğruluk değerleri için test dönemlerinin karşılaştırması*

	Eğitim Doğruluk Yüzdesi (%)	Eğitim F1-skor Değeri	Test Doğruluk Yüzdesi (%)	Test F1-skor Değeri	Algoritma ile Elde Edilen Kazanç (\$)	Alım-Satım İşlem Sayısı
İki sınıflı 1. Test Dönemi	95,81	0,96	86,89	0,90	6686,49	139
İki sınıflı 2. Test Dönemi	91,33	0,92	80,53	0,75	-15393,3	146
İki sınıflı 3. Test Dönemi	92,15	0,93	86,74	0,87	-8851,77	152
Üç sınıflı 1. Test dönemi	88,18	0,87	78,53	0,73	6726,93	130
Üç sınıflı 2. Test dönemi	84,5	0,84	69,06	0,67	-15521,9	128
Üç sınıflı 3. Test dönemi	88,59	0,88	80,66	0,80	-5696,66	151

Elde edilen sonuçları gösteren Tablo 5.46. ve Tablo 5.47. incelendiğinde finansal zaman serilerinin analizinde LSTM ağlarının kullanılması durumunda makine öğrenmesi algoritmalarına kıyasla daha yüksek doğruluk oranlarına ulaşılabilirdiği görülmektedir.

Makine öğrenmesi algoritmalarıyla %53,37 test doğruluk oranına ulaşılabilirken, LSTM ağları için bu değer %86,89 olarak gerçekleşmiştir. Ulaşılan eğitim ve test doğruluk yüzdelerinin birbirine yakın değerlerde olması aşırı uyum problemi yaşanmadığını göstermektedir.

Öznitelik belirleme dönemi için en yüksek doğruluk yüzdesi değerine ulaşılan öznitelik kombinasyonları incelendiğinde, ikili analizde KCP teknik indikatörü ve önceki hareket yönü ile Duygu değeri 3 değişkeni en iyi sonucu verirken, üç sınıflı analizde KCP teknik indikatörü ve önceki hareket yönü ile Duygu değeri 4 değişkenleri en iyi sonucu vermektedir. Her iki durumda da yazar etki değerini içeren duygu durum değerleri ile en iyi sonuca ulaşıldığı görülmektedir. Her iki durum içinde KCP teknik indikatörünü ve önceki hareket yönü kullanılarak daha yüksek doğruluk yüzdeleri elde edilmiştir. Her iki durumda da kelime vektörünün dahil edilmediği durumda en iyi sonuca ulaşılmıştır. İkili ve üç sınıflandırma analizlerinin ikisinde de 120dk zaman periyodunda en yüksek doğruluk yüzdesi değerine ulaşılmıştır.

Elde edilen sonuçlar değerlendirildiğinde ikili sınıflandırma analizi sonucunda üç sınıflı sınıflandırma analizi sonuçlarına kıyasla daha yüksek doğruluk yüzdelerine ulaşıldığı görülmektedir. İki durum içinde 120 dk zaman periyodunda en yüksek doğruluk yüzdesine ulaşılmasından yola çıkarak düşük zaman periyodlarında yüksek doğruluk yüzdesi elde etmenin daha zor olduğu söylenebilir.

Yüksek ortalama doğruluk yüzdesine ulaşılan durumlarda toplam kar miktarının daha düşük olurken, yüksek toplam kar miktarına ulaşılan durumlarda ortalama doğruluk miktarının daha düşük olduğu görülmüştür. Aynı zaman da yüksek ortalama doğruluk değerlerinin elde edildiği durumlarda alım-satım sinyalleri daha yüksek sayıda oluşmakta ve %0.1'lik işlem ücretinden kaynaklı olarak toplam kar miktarları azalmaktadır.

6. SONUÇ VE ÖNERİLER

Bu çalışma kapsamında NASDAQ borsasında yer alan beş adet teknoloji firması seçilmiştir. Sosyal medya verileri X aracılığıyla toplanmış ve hisse senedi verileriyle birlikte birçok öznitelik hesaplanarak analizler gerçekleştirilmiştir. X paylaşımını paylaşan kişinin etkisinin ve paylaşımın ulaştığı etki büyüklüğünün algoritmaya dahil edilmesi amacıyla gönderi istatistikleri kullanılarak yazar etki değeri ve post etki değeri öznitelikleri hesaplanmış ve bu öznitelikler gönderilerin duygu değerleriyle birlikte analize dahil edilmiştir. Toplanan veriler Eğitim, Öznitelik seçimi ve Test olmak üzere üç farklı döneme ayrılmıştır ve Test dönemi içinde üç farklı alternatif oluşturulmuştur. Tahmin edilen hisse senedi hareket yönü kullanılarak bir market simülasyonu oluşturulmuş ve toplam kar-zarar, toplam yatırım değeri ve ortalama doğruluk yüzdeleri hesaplanmıştır. Çalışma kapsamında öznitelik belirleme dönemi için;

- Beş farklı şirket (Tesla, Microsoft, Apple, Netflix ve Nvidia)
- Beş farklı zaman periyodunda (5 dk – 15 dk – 30 dk – 60 dk – 120 dk),
- İki farklı veri seti için (tüm gönderiler ve sadece cashtag içeren gönderiler)
- Üç farklı duygu analiz algoritması (Stanza, Textblob, Vader)
- Sadece Metin ve hem metin hem de emoji duygu durumlarının olduğu 2 farklı durum
- Kelime vektörünün dahil edildiği ve edilmediği 2 farklı durum
- Altı farklı duygu değeri değişkeni
- 45 farklı bağımsız değişken kombinasyonu
- Dört farklı makine öğrenmesi algoritması (Rastgele orman, Karar ağacı, Destek vektör makinesi ve Lojistik regresyon)
- İki farklı sınıflandırma analizi (iki sınıflı ve üç sınıflı)

İçin toplamda 900,000 farklı analiz gerçekleştirilmiştir.

Öznitelik belirleme dönemi için elde edilen en iyi değerler karşılaştırıldığında en iyi sonuçların duygu değerinin dahil olduğu parametre kombinasyonlarında elde edildiği görülmektedir. Bu durumdan yola çıkarak duygu durum değerinin analizlere dahil edilmesinin karlılığı arttığı söylenebilir. Sonuçlar incelendiğinde en iyi sonuçlara yazar veya post etki değerlerinin katıldığı duygu durum değişkenleri ile ulaşılmıştır. Bu sonuçlardan yola çıkarak yazar ve post etki değerlerinin analizlere dahil edilmesinin toplam karı ve ortalama doğruluk yüzdesini artırdığı sonucuna ulaşılmıştır.

En iyi sonucu veren öznitelik kombinasyonları incelendiğinde KCP teknik indikatörünün diğer teknik indikatörlere kıyasla daha iyi sonuçlar verdiği görülmüştür. LSTM ağı kullanılarak yapılan tahminler sonucunda makine öğrenmesi algoritmalarına kıyasla çok daha yüksek doğruluk yüzdelere ulaşılmıştır. İki sınıflı analizde RF algoritmasıyla %53,37 doğruluk yüzdesine ulaşılabilmektedirken LSTM algoritmasıyla %86,89 doğruluk yüzdesi değerine ulaşılabilmektedir. Üç sınıflı analizlerde LR algoritmasıyla %47,45 doğruluk yüzdesine ulaşılabilmektedirken, LSTM ağı kullanılarak %80,66 doğruluk yüzdesi değerine ulaşılmıştır. LSTM ağı ile elde edilen sonuçlarda Makine öğrenmesi yöntemlerine kıyas daha fazla alım-satım sinyali oluşmuştur.

Elde edilen sonuçlar değerlendirildiğinde ikili sınıflandırma analizi sonucunda üç sınıflı sınıflandırma analizi sonuçlarına kıyasla daha yüksek doğruluk yüzdelere ulaşıldığı görülmektedir. İki durum içinde 120 dk zaman periyodunda en yüksek doğruluk yüzdesine ulaşılmış olan yola çıkarak düşük zaman periyodlarında yüksek doğruluk yüzdesi elde etmenin daha zor olduğu söylenebilir.

Çalışma kapsamında gerçekleştirilen analizlerin süreleri değerlendirildiğinde makine öğrenmesi algoritmalarıyla, derin öğrenme algoritmalarına kıyasla çok daha kısa sürede sonuçlar elde edilebildiği görülmüştür. Kelime vektörünün dahil edildiği öznitelik kombinasyonları için analiz sürelerinde ciddi bir artış olmaktadır.

Yapılan çalışma kapsamında sonuçlar değerlendirildiğinde en yüksek doğruluğu sağlayan öngörü modelinin yüksek karı sağlamadığı görülmüştür. Yüksek frekanslı finansal verilerdeki yüksek oynaklık nedeniyle, bu verilerin yüksek doğrulukla tahmin edilmesi alım-satım sinyallerinin sayısını arttırmaktadır. Borsadaki alım-satım işlemleri için ödenen işlem ücretine bağlı olarak yüksek alım-satım sayıları yatırımın kar miktarı ciddi miktarda azaltmaktadır.

Ayı piyasasının yaşandığı ikinci test dönemi sonunda algoritma tahminlerinin kullanılması durumunda hisse senetlerinin alınıp tutulmasına kıyasla toplam yatırım değerinde daha az kayıp yaşandığı görülmektedir. Daha dengeli olan üçüncü test döneminde ise algoritma tahminlerinin kullanılması halinde hisse senetlerinin alınıp tutulmasına kıyasla daha yüksek getiri sağlanmıştır. Boğa piyasasının yaşandığı birinci test döneminde ise algoritma tahminlerinin kullanılmasıyla ulaşılan toplam yatırım değerinin, hisse senetlerinin alınıp tutulması halindeki toplam yatırım değerinin altında kaldığı görülmektedir.

Çalışma kapsamında elde edilen bulgulardan yola çıkarak boğa piyasanın yaşandığı dönemlerde al-tut stratejisinin daha karlı olduğu fakat ayı piyasasının yaşandığı ya da büyük değişimlerin yaşanmadığı dönemlerde belirlenen öznitelikler ve sosyal medya verilerini kullanan algoritmalar ile elde edilen tahminlerinin yatırımların yönetilmesinde kullanılması durumunda al-tut stratejisine kıyasla daha yüksek karlar elde edilebileceği bulunmuştur.

Bu analizler kapsamında kullanılan sosyal medya paylaşımlarına ait istatistiksel etkileşim verilerin paylaşım yapıldıktan sonra ileri bir tarihte toplanması nedeniyle aldığı etkileşim miktarında zamana bağlı olarak biriken bir artış olacağı bilinmektedir. Bu nedenle gelecekteki çalışmalarda sosyal medya paylaşımlarının ve bu paylaşımlara ait istatistiksel etkileşim verilerinin anlık olarak hisse senedi verileriyle birlikte toplanması halinde paylaşımların hisse senetleri üzerindeki etkisi daha etkili bir şekilde ortaya konabilir. Gelecek çalışmalarda daha farklı değişken kombinasyonları ve farklı makine öğrenmesi algoritmaları kullanılarak daha kapsamlı deneyler gerçekleştirilebilir. Gelecek çalışmalarda farklı zaman periyotları için analizler gerçekleştirilerek daha kapsamlı çalışmalar gerçekleştirilebilir.

KAYNAKÇA

- [1] Hoven, T.C. (2020). *Do President Donald Trump' s tweets affect the stock market?: An analysis of company-specific tweets.*
- [2] Kinyua, J.D., Mutigwe, C., Cushing, D.J. ve Poggi, M. (2021). An analysis of the impact of President Trump's tweets on the DJIA and S&P 500 using machine learning and sentiment analysis. *Journal of behavioral and experimental finance.* 29, 100447.
- [3] Piñeiro-Chousa, J., Vizcaíno-González, M. ve Pérez-Pico, A.M. (2017). Influence of Social Media over the Stock Market. *Psychology & Marketing.* 34(1), 101-108.
- [4] Zhou, Z., Gao, M., Liu, Q. ve Xiao, H. (2020). Forecasting stock price movements with multiple data sources: Evidence from stock market in China. *Physica A: Statistical Mechanics and its Applications.* 542, 123389.
- [5] Ren, R., Wu, D.D. ve Liu, T. (2019). Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal.* 13(1), 760-770.
- [6] Liu, J.-X., Leu, J.-S. ve Holst, S. (2023). Stock price movement prediction based on Stocktwits investor sentiment using FinBERT and ensemble SVM. *PeerJ Computer Science.* 9, e1403.
- [7] Ghanavati, M., Wong, R.K., Chen, F., Wang, Y. ve Fong, S. (2016). A generic service framework for stock market prediction. *2016 IEEE international conference on services computing (SCC)*, IEEE, 283-290.
- [8] Bouktif, S., Fiaz, A. ve Awad, M. (2020). Augmented Textual Features-Based Stock Market Prediction. *Ieee Access.* 8, 40269-40282.
- [9] Saurabh, S. ve Dey, K. (2020). Unraveling the relationship between social moods and the stock market: Evidence from the United Kingdom. *Journal of behavioral and experimental finance.* 26, 100300.
- [10] Khan, W., Ghazanfar, M.A., Azam, M.A., Karami, A., Alyoubi, K.H. ve Alfakeeh, A.S. (2020). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing.* 1-24.
- [11] Qiu, Y., Song, Z. ve Chen, Z. (2022). Short-term stock trends prediction based on sentiment analysis and machine learning. *Soft Computing.* 26(5), 2209-2224.
- [12] Xiao, Q. ve Ihnaini, B. (2023). Stock trend prediction using sentiment analysis. *PeerJ Computer Science.* 9, e1293.
- [13] Coyne, S., Madiraju, P. ve Coelho, J. (2017). Forecasting stock prices using social media analysis. *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl*

- [14] Nam, K. ve Seong, N. (2019). Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market. *Decision Support Systems*. 117, 100-112.
- [15] Chakraborty, P., Priya, U.S., Rony, M.R.A.H. ve Majumdar, M.A. (2017). Predicting stock movement using sentiment analysis of Twitter feed. *2017 6th International Conference on Informatics, Electronics and Vision & 2017 7th International Symposium in Computational Medical and Health Technology (ICIEV-ISCMHT)*, IEEE, 1-6.
- [16] Maqsood, H., Mehmood, I., Maqsood, M., Yasir, M., Afzal, S., Aadil, F., Selim, M.M. ve Muhammad, K. (2020). A local and global event sentiment based efficient stock exchange forecasting using deep learning. *International Journal of Information Management*. 50, 432-451.
- [17] Mndawe, S.T., Paul, B.S. ve Doorsamy, W. (2022). Development of a stock price prediction framework for intelligent media and technical analysis. *Applied Sciences*. 12(2), 719.
- [18] Li, J., Bu, H. ve Wu, J. (2017). Sentiment-aware stock market prediction: A deep learning method. *2017 international conference on service systems and service management*, IEEE, 1-6.
- [19] Verma, I., Dey, L. ve Meisheri, H. (2017). Detecting, quantifying and accessing impact of news events on Indian stock indices. *Proceedings of the International Conference on Web Intelligence*, 550-557.
- [20] Chen, M.-Y., Liao, C.-H. ve Hsieh, R.-P. (2019). Modeling public mood and emotion: Stock market trend prediction with anticipatory computing approach. *Computers in human behavior*. 101, 402-408.
- [21] Wang, Z. ve Shirai, K. (2018). Measuring Financial Crisis Index for Risk Warning through Analysis of Social Network. *24th Annual Conference of the Association for Natural Language Processing*,
- [22] Li, Y., Bu, H., Li, J. ve Wu, J. (2020). The role of text-extracted investor sentiment in Chinese stock price prediction with the enhancement of deep learning. *International Journal of Forecasting*. 36(4), 1541-1562.
- [23] Albahli, S., Irtaza, A., Nazir, T., Mehmood, A., Alkhalifah, A. ve Albattah, W. (2022). A machine learning method for prediction of stock market using real-time twitter data. *Electronics*. 11(20), 3414.
- [24] Vargas, M.R., De Lima, B.S. ve Evsukoff, A.G. (2017). Deep learning for stock market prediction from financial news articles. *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, IEEE, 60-65.

- [25] Chen, W., Yeo, C.K., Lau, C.T. ve Lee, B.S. (2018). Leveraging social media news to predict stock index movement using RNN-boost. *Data & Knowledge Engineering*. 118, 14-24.
- [26] Wang, Q., Xu, W. ve Zheng, H. (2018). Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. *Neurocomputing*. 299, 51-61.
- [27] Huynh, H.D., Dang, L.M. ve Duong, D. (2017). A new model for stock price movements prediction using deep neural network. *Proceedings of the Eighth International Symposium on Information and Communication Technology*, 57-62.
- [28] Kraus, M. ve Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems*. 104, 38-48.
- [29] Zhuge, Q., Xu, L. ve Zhang, G. (2017). LSTM Neural Network with Emotional Analysis for prediction of stock price. *Engineering letters*. 25(2).
- [30] Wu, S., Liu, Y., Zou, Z. ve Weng, T.-H. (2022). S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis. *Connection Science*. 34(1), 44-62.
- [31] Minh, D.L., Sadeghi-Niaraki, A., Huy, H.D., Min, K. ve Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *Ieee Access*. 6, 55392-55404.
- [32] Ding, X., Zhang, Y., Liu, T. ve Duan, J. (2015). Deep learning for event-driven stock prediction. *Twenty-fourth international joint conference on artificial intelligence*,
- [33] Akita, R., Yoshihara, A., Matsubara, T. ve Uehara, K. (2016). Deep learning for stock prediction using numerical and textual information. *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, IEEE, 1-6.
- [34] Hagenau, M., Liebmann, M., Hedwig, M. ve Neumann, D. (2012). Automated news reading: Stock price prediction based on financial news using context-specific features. *2012 45th Hawaii International Conference on System Sciences*, IEEE, 1040-1049.
- [35] Dang, M. ve Duong, D. (2016). Improvement methods for stock market prediction using financial news articles. *2016 3rd National foundation for science and technology development conference on information and computer science (NICS)*, IEEE, 125-129.
- [36] Das, S., Behera, R.K. ve Rath, S.K. (2018). Real-time sentiment analysis of twitter streaming data for stock prediction. *Procedia Computer Science*. 132, 956-964.

- [37] Xu, F. ve Keelj, V. (2014). Collective sentiment mining of microblogs in 24-hour stock price movement prediction. *2014 IEEE 16th conference on business informatics*, IEEE, 60-67.
- [38] Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M. ve Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. *PloS one*. 10(9), e0138441.
- [39] Chang, C.Y., Zhang, Y., Teng, Z., Bozanic, Z. ve Ke, B. (2016). Measuring the information content of financial news. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3216-3225.
- [40] Jangid, H., Singhal, S., Shah, R.R. ve Zimmermann, R. (2018). Aspect-based financial sentiment analysis using deep learning. *Companion Proceedings of the The Web Conference 2018*, 1961-1966.
- [41] Sohangir, S., Wang, D., Pomeranets, A. ve Khoshgoftaar, T.M. (2018). Big Data: Deep Learning for financial sentiment analysis. *Journal of Big Data*. 5(1), 3.
- [42] Mahmoudi, N., Docherty, P. ve Moscato, P. (2018). Deep neural networks understand investors better. *Decision Support Systems*. 112, 23-34.
- [43] Piao, G. ve Breslin, J.G. (2018). Financial aspect and sentiment predictions with deep neural networks: an ensemble approach. *Companion Proceedings of the The Web Conference 2018*, 1973-1977.
- [44] Atsalakis, G.S. ve Valavanis, K.P. (2009). Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert systems with applications*. 36(7), 10696-10707.
- [45] Dağlı, H. (2004). Sermaye piyasası ve portföy analizi. *Trabzon: Derya Kitabevi*.
- [46] McMillan, M., Pinto, J.E., Pirie, W.L. ve Van de Venter, G. (2011). *Investments: Principles of portfolio and equity analysis*. John Wiley & Sons.
- [47] Edwards, R.D., Magee, J. ve Bassetti, W.C. (2018). *Technical analysis of stock trends*. CRC press.
- [48] Box, G.E. ve Jenkins, G.M. (1970). *Time series analysis. Forecasting and control*. San Francisco: Holden-Day.
- [49] Whittle, P. (1951). *Hypothesis testing in time series analysis*.
- [50] Box, G.E., Jenkins, G.M., Reinsel, G.C. ve Ljung, G.M. (2016). *Time series analysis: forecasting and control (fifth edition)*. John Wiley & Sons.
- [51] Bush, V. (1945). As we may think. *The atlantic monthly*. 176(1), 101-108.
- [52] Turing, A.M. (1950). Computing Machinery and Intelligence.

- [53] Smith, C., McGuire, B., Huang, T. ve Yang, G. The History of Artificial Intelligence 2006: University of Washington.
- [54] Arf, C. (1959). Makine düşünebilir mi ve nasıl düşünebilir. *Atatürk Üniversitesi-Üniversite Çalışmalarını Muhite Yayma ve Halk Eğitimi Yayınları Konferanslar Serisi*.(1), 91-103.
- [55] Samuel, A.L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*. 3(3), 210-229.
- [56] Nilsson, N.J. (1965). Learning machines.
- [57] Duda, R.O. ve Hart, P.E. (1973). *Pattern recognition and scene analysis*. Wiley, New York.
- [58] Deng, L. ve Yu, D. (2014). Deep learning: methods and applications. *Foundations and trends® in signal processing*. 7(3–4), 197-387.
- [59] Rokach, L. ve Maimon, O. (2010). *Data mining and knowledge discovery handbook*. Springer New York.
- [60] Kothari, R. ve Dong, M. (2001). Decision trees for classification: A review and some new results. *Pattern recognition: from classical to modern approaches*. 169-184.
- [61] Morgan, J.N. ve Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*. 58(302), 415-434.
- [62] Breiman, L., Friedman, J., Stone, C.J. ve Olshen, R. (1984). *Classification and Regression Trees*. CRC Press.
- [63] Quinlan, J.R. (1986). Induction of decision trees. *Machine learning*. 1, 81-106.
- [64] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
- [65] Kuhn, M. ve Johnson, K. (2013). *Applied predictive modeling*. Springer.
- [66] Breiman, L. (2001). Random forests. *Machine learning*. 45, 5-32.
- [67] Cortes, C. ve Vladimir, V. (1995). Support-Vector Networks. *Machine learning*. 20, 273–297.
- [68] Crammer, K. ve Singer, Y. (2001). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*. 2(Dec), 265-292.
- [69] Soman, K., Loganathan, R. ve Ajay, V. (2009). *Machine learning with SVM and other kernel methods*. PHI Learning Pvt. Ltd.

- [70] Smola, A.J. ve Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*. 14, 199-222.
- [71] Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American statistical association*. 39(227), 357-365.
- [72] Cox, D.R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 20(2), 215-232.
- [73] Cox, D.R. (1966). Some procedures associated with the logistic qualitative response curve. *Research papers in statistics: Festschrift for J. Neyman*. 55-71.
- [74] Alpar, C. (2017). Uygulamalı çok değişkenli istatistiksel yöntemler.
- [75] Berenson, M., Levine, D., Szabat, K.A. ve Krehbiel, T.C. (2012). Basic business statistics: Concepts and applications.
- [76] Hochreiter, S. ve Schmidhuber, J. (1997). Long Short-Term Memory. *Neural computation*. 9(8), 1735-1780.
- [77] Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R. ve Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*. 28(10), 2222-2232.
- [78] Maron, M.E. (1961). Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*. 8(3), 404-417.
- [79] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*. 34(1), 1-47.
- [80] Ortigosa, A., Martín, J.M. ve Carro, R.M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in human behavior*. 31, 527-541.
- [81] Nasukawa, T. ve Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd international conference on Knowledge capture*, 70-77.
- [82] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. ve Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*. 61(12), 2544-2558.
- [83] Medhat, W., Hassan, A. ve Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*. 5(4), 1093-1113.
- [84] Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*. 2(2010), 627-666.
- [85] Ikonomakis, M., Kotsiantis, S. ve Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS transactions on computers*. 4(8), 966-974.

- [86] Luhn, H.P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*. 1(4), 309-317.
- [87] Manning, C.D. (2008). *Introduction to information retrieval*. Syngress Publishing.
- [88] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*. 28(1), 11-21.
- [89] Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation*. 60(5), 503-520.
- [90] Fragopanagos, N. ve Taylor, J.G. (2005). Emotion recognition in human–computer interaction. *Neural Networks*. 18(4), 389-405.
- [91] Davis, M. ve Edberg, P. Unicode emoji2015: Technical Report 51.
- [92] Kralj Novak, P., Smailović, J., Sluban, B. ve Mozetič, I. (2015). Sentiment of emojis. *PloS one*. 10(12), e0144296.
- [93] Kralj Novak, P., Smailović, J., Sluban, B. ve Mozetič, I. (2015). Emoji sentiment ranking 1.0.
- [94] Hutto, C. ve Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the international AAAI conference on web and social media*, 216-225.
- [95] Qi, P., Zhang, Y., Zhang, Y., Bolton, J. ve Manning, C.D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.
- [96] Colby, R.W. ve Meyers, T.A. (1988). The encyclopedia of technical market indicators. (*No Title*).
- [97] Keltner, C.W. (1960). How to make money in commodities. (*No Title*).
- [98] Arms Jr, R.W. (1990). Ease Of Movement. *Technical Analysis of Stocks*. 8(Issue).
- [99] Schlossberg, B. (2006). *Technical analysis of the currency market: classic techniques for profiting from market swings and trader sentiment*. John Wiley & Sons.
- [100] Chande, T.S. ve Kroll, S. (1994). *The new technical trader: boost your profit by plugging into the latest indicators*.
- [101] Wilder, J.W. (1978). *New concepts in technical trading systems*. Greensboro, NC.
- [102] Pedregosa, F. (2011). Scikit-learn: Machine learning in python Fabian. *Journal of machine learning research*. 12, 2825.

- [103] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G. ve Isard, M. (2016). TensorFlow: a system for Large-Scale machine learning. *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 265-283.
- [104] Seabold, S. ve Perktold, J. (2010). Statsmodels: econometric and statistical modeling with python. *SciPy*. 7(1).
- [105] Smith, T.G. (2017). pmdarima: ARIMA estimators for Python.
- http-1:** <https://www.bbc.com/turkce/haberler-dunya-52511235> (Eriřim tarihi: 04.03.2022)
- http-2:** <https://www.ntv.com.tr/ekonomi/elon-musk-onerdi-yanlis-signal-hisseleri-artti-yuzde-bin-300-artis,qEYhpf2AkUWLHkSlwj6NFg> (Eriřim tarihi: 04.03.2022)
- http-3:** <https://www.independent.co.uk/news/business/news/celebrity-stock-market-influence-twitter-kylie-jenner-snapchat-tweet-donald-trump-oprah-winfrey-weight-watchers-mcafee-cryptocurrencies-a8225266.html> (Eriřim tarihi: 04.03.2022)
- http-4:** <https://www.marieclaire.com/celebrity/a18740061/celebrities-stock-market-influence/> (Eriřim tarihi: 04.03.2022)
- http-5:** https://cdn.emogi.com/docs/reports/2015_emoji_report.pdf (Eriřim tarihi: 22.01.2022)

ÖZGEÇMİŞ

ORCID NO: 0000-0003-3399-8658

Ad Soyad : Hakan GÖKDAŞ

Yabancı Dil : İngilizce

Eğitim ve Mesleki Geçmişi:

- 2013, Eskişehir Osmangazi Üniversitesi, Mühendislik Mimarlık Fakültesi, İnşaat Mühendisliği Bölümü
- 2014, Eskişehir Osmangazi Üniversitesi, Mühendislik Mimarlık Fakültesi, Endüstri Mühendisliği Bölümü (Çift Anadal)
- 2016, Anadolu Üniversitesi, İşletme Fakültesi, İşletme Bölümü
- 2018, Eskişehir Osmangazi Üniversitesi, Fen Bilimleri Enstitüsü, Endüstri Mühendisliği Anabilim Dalı, Yöneylem Araştırması Bilim Dalı
- 2018, Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, İnşaat Mühendisliği Anabilim Dalı, Yapı Bilim Dalı