MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

ANALYSIS OF KAP FINANCIAL DISCLOSURES AND CREATION OF KAP INDEX

MUHLİS SARIYER

MASTER THESIS

Department of Data Engineering

ADVISORS

Assoc. Prof. Dr. Murat Can GANİZ

Prof. Dr. Lokman GÜNDÜZ

ISTANBUL, 2022

MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

ANALYSIS OF KAP FINANCIAL DISCLOSURES AND CREATION OF KAP INDEX

MUHLİS SARIYER

MASTER THESIS

Department of Data Engineering

ADVISORS
Assoc. Prof. Dr. Murat Can GANİZ
Prof. Dr. Lokman GÜNDÜZ

ISTANBUL, 2022

MARMARA UNIVERSITY

INSTITUTE FOR GRADUATE STUDIES IN PURE AND APPLIED SCIENCES

Muhlis SARIYER, a Master of Science student of Marmara University Institute for Graduate Studies in Pure and Applied Sciences, defended his thesis entitled "**Analysis of KAP Financial Disclosures and Creation of KAP Index**", on June 15, 2022 and has been found to be satisfactory by the jury members.

Jury Members

Assoc. Prof. Dr. Murat Can GANİZ (Advisor)
Marmara University…………………………………………………. (SIGN)……………

Assoc. Prof. Dr. Mustafa Ağaoğlu (Jury Member)
Marmara University…………………………………………………. (SIGN)……………

Prof. Dr. Selim Akyokuş (Jury Member)
Istanbul Medipol University………………………………………. (SIGN)…………

APPROVAL

Marmara University Institute for Graduate Studies in Pure and Applied Sciences Executive Committee approves that Muhlis SARIYER be granted the degree of Master of Science in Data Engineering Program, on , . (Resolution no:).

Director of the Institute
Prof. Dr.

DECLARATION OF AUTHORSHIP

I, Muhlis SARIYER, hereby declare that this thesis titled “Analysis of KAP Financial Disclosures and Creation of KAP Index” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other
qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ____

Date: ____

ACKNOWLEDGMENT

I want to acknowledge the Granter of all Knowledge's allowance to the completion of this work, for without His gift there is no knowing.

I find it my responsibility to thank the Great Scholars of History, without them there would not be any basis of our knowledge. Among them I specifically thank my advisors Doç. Dr. Murat Can Ganiz and Prof. Dr. Lokman Gündüz.

There is no doubt that this work would not come to fruition if not for the unwavering support of my Mother and my Beloved Wife, my indebtedness to them is incalculable.

TABLE OF CONTENTS

ÖZET

KAP FİNANSAL BİLDİRİMLERİNİN ANALİZİ VE KAP ENDEKSİNİN OLUŞTURULMASI

Bu çalışmada tekil hisselerin fiyatlarının ve günlük işlem hacimlerinin tahmin edilmesi için makine öğrenmesi modelleri kullanılmıştır. Kamuyu Aydınlatma Platformu (KAP) Borsa İstanbul'da kayıtlı şirketlerin bildirimlerini üzerinden yapmakla yükümlü oldukları düzenleyici platformdur.

Borsa İstanbul yatırımcıları Twitter'ı hisselere dair duygularını ifade etmek için kullanmaktadırlar. Tahminleme modelimiz yatırımcıların Twitter'da ifade ettikleri duygularını BIST30 endeksinde bulunan şirketlerin KAP bildirimlerinden elde edilen puanlarla birleştirerek hisselerin günlük işlem hacmi ve fiyat değişimlerini tahmin etmektedir. BIST30, DJI endeksi ve USD ile Ons Altın fiyatlarını içeren piyasanın durumuna dair finansal veriler modelin doğruluk oranını arttırmak için eklenmiştir. Bildirimlerden özellikle etkilenen firmalarda tekil hisse fiyatı tahmininde 80%'i aşan doğruluk oranına ulaşılmıştır. Bütün BIST30 şirketlerinde ise hacim tahmininde ortalama 74.7% doğruluk oranına ulaşılmıştır.

Her bir şirket için tekil makine öğrenmesi modelleri güvenilir şekilde yüksek doğruluk oranları yakalaması sebebiyle bütün şirketlere dair genel bir KAP endeksi oluşturulmasının mümkün olmadığı görülmüştür.

ABSTRACT

ANALYSIS OF KAP FINANCIAL DISCLOSURES AND CREATION OF KAP INDEX

In this study machine learning models are used for predicting individual stock price and daily volume changes using sentiments from public disclosures and tweets. Public Disclosure Platform (KAP) is the mandated regulatory platform for disclosing news about companies listed in Borsa Istanbul Stock Exchange.

Investors in Borsa Istanbul use Twitter to express their sentiments for stocks. By combining people's sentiment in Twitter and companies' disclosures, our prediction model predicts the volume and price changes of individual company stocks listed in BIST30. Financial data regarding market conditions consisting of daily price changes of BIST30, DJI, USD and Gold per Ounce are also added to enhance prediction f1 score of the model. Above 80% individual stock price prediction f1 score is achieved for companies with high susceptibility to news. 74.7% mean volume prediction f1 score is achieved across all BIST30 companies.

It has been found that building individual machine learning models for each company produces reliably higher f1 score therefore creation of a general KAP index about all companies is not feasible.

SYMBOLS

l	: Tweet Like Count
RE	: Tweet Reply Count
RT	: Tweet Retweet Count
f	: Tweet Follower Count
IS	: Interaction Score for tweets
z	: Standard score
log	: logarithm
σ	: standard deviation of the sample
μ	: mean of the sample

ABBREVIATIONS

ML **:** Machine Learning

KAP **:** Public Disclosure Platform

BIST **:** Borsa Istanbul Stock Exchange

BIST30 **:** BIST30 Index

DJI **:** Dow Jones Index

NLP **:** Natural Language Processing

BERT **:** Bidirectional Encoder Representations from Transformers

ANN **:** Artificial Neural Network

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Stock Market Studies

The study of stock markets is a study of human behavior [1]. This assumes people that participate in the market behave emotionally and their collective movements are impacted by news about companies [2]. Efficient markets hypothesis assumes that the participants in an efficient market make their decisions based on information available to the public [2]. Borsa Istanbul provides investors with KAP (Public Disclosure Platform) for companies to disclose information publicly to provide equal opportunity of information for all investors [4]. Stock price prediction using financial information is a rigorously studied area due to its potential applications for building profitable trading strategies. Most of the studies in the field are done to predict market indices such as BIST30 or BIST100 [6].

BIST30 and BIST100 are the indices that represent stock performance of the 30 and 100 most valuable companies in Borsa Istanbul Stock Exchange, respectively. Public Disclosure Platform (KAP) is the mandated regulatory platform for disclosing news about companies.

In conjunction with our research focus, we train different machine learning models for each company to observe the impact of public disclosures on individual stocks. This allow us to achieve higher confidence towards a trading strategy as well as provide us the ability to distinguish the impact of news sentiments on companies. Some companies are more effected by disclosures than others.

1.2 Efficient Markets Hypothesis

Market efficiency is a financial assessment of markets with regards to their relationship with information. If information access is more transparent for all possible investors these markets are assumed to be more efficient. If information regarding a company is directly represented in its stock price and such information is accessible by everyone than information is efficiently reflected in the stock. No parties can have an asymmetric information advantage [4].

1.3 Natural Language Processing

We build predictive models on individual stocks to predict their direction if they will increase or decrease that are listed in BIST30. We focus on creating models for predicting the individual stock price changes.

The literature of sentiment analysis driven stock price prediction consists mostly of financial news sentiment analysis [16], [17]. In this study, we focus on public disclosures because they are the first source of information about a company [9]. We also focus on Twitter so we can determine how information regarding a company spreads in public and sentiment of investors regarding the stock.

We create several individual stock price and volume change prediction models to foresee volatility and price changes using public disclosures and Twitter sentiments. We improved our model by including data about market conditions. We tested and implemented imputation methods to assess news relevance across time.

Sentiment Analysis is done by using Natural Language Processing algorithms such as BERT. These machine learning models provide a sentiment score for a given text.

1.4 Purpose of the Study

Study of stock market prediction using NLP is generally done for stock market indices and not on individual stocks. For Borsa Istanbul case single stock prediction is especially applicable due to the existence of KAP and high information traffic on Twitter. KAP provides information regarding every company in BIST and therefore can be used as the initial source of information about companies and provide a level investment ground for all investors. The use of KAP with NLP for stock price prediction is an important area of research for understanding the behavior of stock market investors and their relationship with public information. This study aims to analyze the efficiency of Borsa Istanbul in terms of information residing in KAP.

1.5 Thesis Structure

The thesis is structured as follows:

- In Related Work, literature on stock price prediction is analyzed with regards to Sentiment Analysis using financial news and tweets. Studies on Borsa Istanbul stock prices and KAP are analyzed in conjunction with the purpose of this work.
- In the Approach, the gathering process of Tweets, KAP disclosures and market indices are explained. Processing of the gathered data is explained in depth for consolidation into a data structure that is usable in machine learning modeling.
- In the Experiment Results and Discussion, findings of this study are presented, and their implications are discussed.
- In Thesis Conclusion, the steps and processes of this study are summarized.
- In Future Work, areas of improvement about this work for higher prediction f1 score and more comprehensive results are discussed.

2. RELATED WORK

2.1 The Use of Financial News for Stock Price Prediction

Individual stock prediction can be seen in these [4], [19] studies with lower f1 score with respect to index predictions. Compared to our work these studies are not usable for building a trading strategy.

Although the use of financial news [14] in sentiment analysis about stock market companies is more prevalent than the use of Twitter, there are many studies with Twitter data cases, these approaches produce reliably strong results that can be seen in their reported f1 score levels [5], [6]. KAP disclosure use for prediction is rarely seen in the literature and only with approaches focusing on indices and not individual companies' stock [7]. KAP data is unstructured and is seen to be difficult to structure for ML models, therefore it is avoided in many studies in the field.

KAP disclosures such as quarterly financial reports contain detailed financial information therefore consist mostly of numeric values. Generally, NLP models are not compatible with numeric values if they are not specifically built to do so [33]. Numeric values are only meaningful in context. Understanding the context of a numeric value from a small set of data is not reliable.

As a multidisciplinary study building predictive machine learning models of financial information, sources its efforts from both fields. Building statistical models from financial information regarding companies has been studied since Fama and continues to be prevalent on the financial aspects of stock price prediction [32].

As computing power got cheaper and building complex artificial neural network models became more feasible, research around ANNs have entered the field of stock price predictions [33]. Financial information about any company is a data that requires deep analysis and has high orders of correlation therefore the increase in computability and the spread of new pre-trained models and easily implementable systems have increased the speed of progress in the field.

2.2 The Use of Twitter for Stock Price Prediction

Twitter data use for stock prediction falls under two categories, first is the hashtag use [8], second is a broader keyword detection implementation [9], we found hashtags to be more reliable because they provide a more precise impression for the companies. Keyword detection implementations are comprehensive in planning and provide bigger data that can be used for Deep Learning applications. Our application required precise impressions therefore hashtag method is used.

2.3 Borsa Istanbul

BIST100 companies represent the majority of Borsa Istanbul’s trading volume and BIST30 companies are the biggest and most prestigious companies in BIST100. BIST30 Index and BIST30 companies are widely used for news sentiment analysis by studies conducted on Borsa Istanbul [10], [11]. Companies in BIST30 index use KAP to disclose information much more frequently. Also these companies attract much more attention on social media, especially Twitter. Using BIST100 companies would produce a much sparse dataset that would yield inconclusive results, therefore we decided to use BIST30 as our target.

Gunduz et al. [15][29] have pioneered news sentiment analysis using ML models for Borsa Istanbul in 2013 and achieved greater prediction f1 score in their study in 2018 [30].

Literature on finance and machine learning overlaps in papers regarding stock market analysis. Therefore, there are many papers with data included from financial analyses.

Compared to other studies in the field from the perspective of price prediction f1 score financial models can improve models [31] but require a more manual approach.

General sentiment analysis for assessing predictability of stock returns is done by You et al [13]. Predictability assessment allows the work in the field to be done systematically.

Combination of sentiments and market information in our method can be compared to aggregate/ensemble learning done by Pasupulety et al [12]. These aggregate models are used to contain wider market information and specific data regarding a company.

2.4 BERT Natural Language Processing Pre-Trained Model

BERT [24] is an NLP machine learning pre-trained model that has proven its versatility in many academic studies. Natural language processing requires a deep understanding of human communication. This requires an understanding of context, generalization of such context across many fields can be done only with sizable datasets, using pre-trained models transfers information that exists inside the wider context of all data it has been trained with.

Research that implements NLP solutions requires pre-trained ML models. It is impractical for each study to build a new NLP model. Pre-trained models can be further enhanced by implementing them and by improving their interval weights and parameters to support a certain area of expertise.

Versatility of BERT can also be seen in its wide use in many different languages such as Spanish, French, and Turkish. Financial jargon can change the meaning of words such that their intention can be reversed, therefore area specific NLP models are used to understand textual information for specific cases.

2.5 Market Efficiency

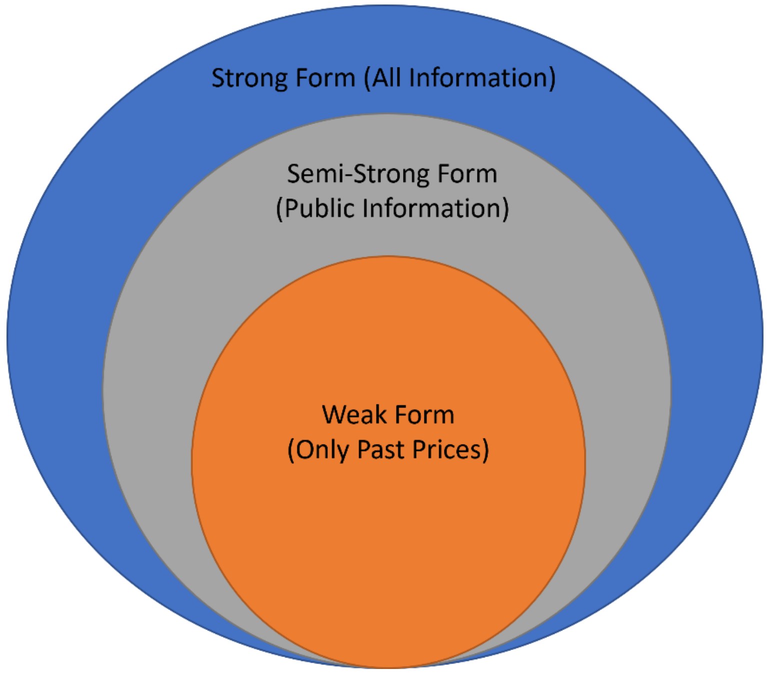


Figure 2.1: Market Efficiency Forms

There are three forms of market efficiency that are comprehensive of one another according to Fama [2]. These forms predict how efficiently stock markets represent information regarding companies inside a stock market.

Weak form assumes all past trading information is reflected inside the current stock price. Including past trading prices and daily trading volumes. Trading with weak form assumptions and using only past prices is called technical analysis [34]. Technical analysis differs from fundamental analysis where news information regarding companies and performance metrics such as dividend and profit announcements are analyzed to predict future prices of a stock and investment decisions are made based on these data. Fundamental analysis is a method with a semi-strong form assumption about the market.

Semi-strong form assumes all public information about a company is reflected in a company's stock including all past trading information that is considered in the weak form. Semi-strong form assumption can be tested by analyzing the impact of news about a company.

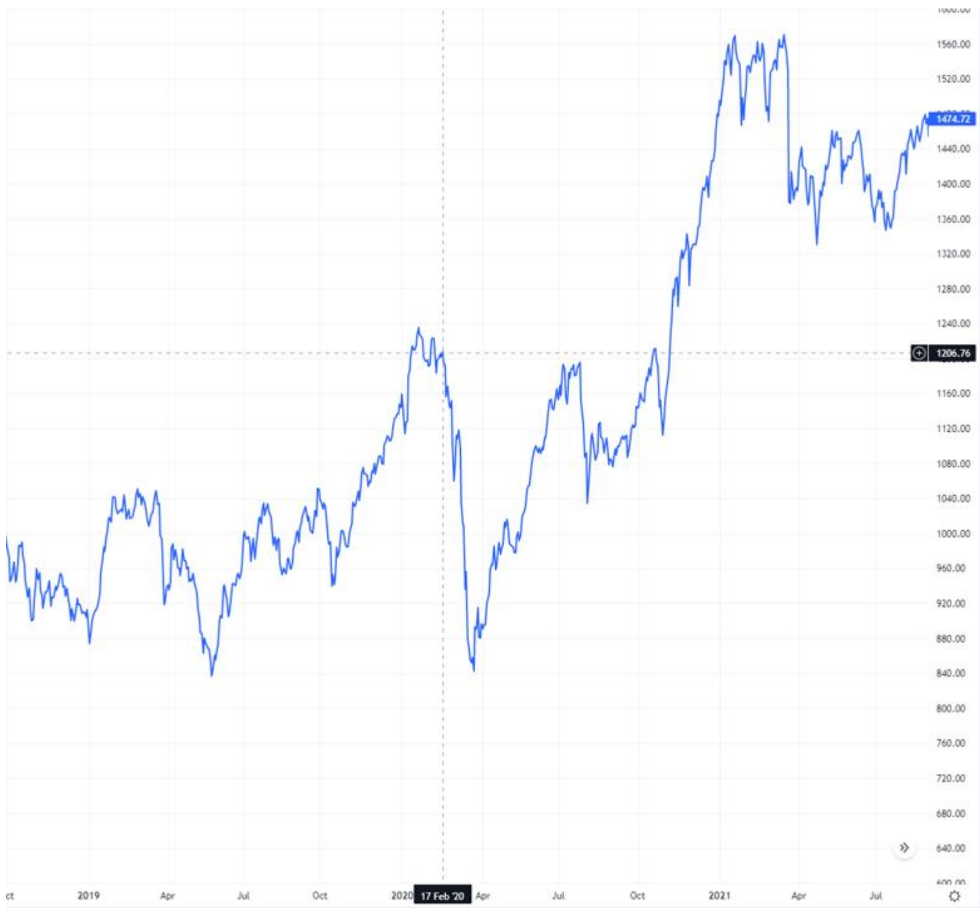


Figure 2.2: COVID-19 Impact on BIST100 Index

An example of this phenomena can be seen in the price fluctuations caused by COVID-19 pandemic which increased volatility and decreased valuation of companies due to a concern for economic downturn caused by lockdowns, decreased economic activity and possible financial crisis [36]. News about COVID-19 on BIST100 can be seen on February 17th, 2020, which is a downwards turning point for BIST100 due to the increased news coverage of the pandemic worldwide that went on to decrease total value by 30%.

Strong form is predicated on the idea that both private and public information is reflected in the markets' pricing of the stocks. The impact of private information regarding a company’s stock price can be seen when a company’s stock prices increase unusually before a big announcement is made by the company. This kind of price movement is assumed to be caused by insider traders [37] who know the announcement before it is made, where they buy the company’s stock thus increasing its price. Such strong form can be calculated by interpolating the sudden changes of a stock price with its announcement dates. If there are strong price changes before announcements are being made by a company and not after the announcements are made there is increased likelihood of insider trading.

This assumption is included in our study by means of including the stock price changes of a company’s stock inside the ML prediction model built for the company by varying intervals, these intervals are as follows: 3, 7, 10, 13 days. If information about a company is reflected in its price it can be seen within different periods.

2.6 Volume Prediction

Daily trading volume refers to the number of shares of an individual stock that is bought and sold each day [35]. This number can be greater than the total number of shares of the company because same shares can be bought and sold in the same day. Volume prediction is an under researched area compared to stock price prediction. Accurate prediction of future stock prices can be used in building long term trading strategies, but volume prediction can be used for building day trading strategies which is a fast form of trading that relies on high price volatility.

3. APPROACH

For financial analysis of big data, deep learning methods are used as machine learning models [4], [16], [20]. In our case, a total of 38,864 data points exist for training the final ML model. Due to the scarcity of labeled data, deep learning methods performed poorly in our experiments. Therefore, we conducted our experiments with ML models that are being used in literature for similar cases.

BERT [23], [24] is a pre-trained machine learning model for general purpose natural language processing developed by Google. BERTurk [25] is a community driven machine learning model developed to calculate sentiments in Turkish sentences. BERTurk is the state-of-the-art sentiment analysis model for Turkish texts and is used in financial news sentiment analysis in many studies [26], [27].

Research that implements NLP solutions requires pre-trained ML models. It is impractical for each study to build a new NLP model. Pre-trained models can be further enhanced by improving their internal weights and parameters to support a certain area of expertise.

The literature on news sentiment analysis for stock market prediction is overwhelmingly done for predicting Market or Sector Indices such as BIST100 or banking sector. BIST100 is the index to represent stock performance of the 100 most valuable companies in Borsa Istanbul (BIST). Banking sector index in Borsa Istanbul is the index for all banks that are listed in BIST.

As a novel approach we choose the prediction of individual stocks rather than a market or sector index, these predictions can form trading strategies to maximize profit in a systemic and reliable way.

Our literature analysis shows that indices and sectors are preferred in this area of research because of their predictable nature. And information about individual companies is too fragmented and sparse to form into a machine learning model. Our study overcomes this fragmentation and sparsity by enriching public disclosure data and combining it with Twitter and wider market data.

3.1 Twitter Data Collection

Stock symbols/tickers are unique abbreviations for each company's stock in a stock exchange such as THYAO for Turkish Airlines or MSFT for Microsoft. For Twitter data "snscrape" python module is used and all the tweets that mentioned a stock symbol in BIST30 is collected with additional data that will be used to give weights such as like count of the tweet, quote count of the tweet, reply count of the tweet, retweet count of the tweet, follower count of the account at the time of tweeting.

In total we gathered 1,494,108 tweets regarding BIST30 companies and their interactions from Twitter between 1st of January 2016 and 1st of January 2021.

Figure 3.1, Figure 3.2 and Figure 3.3 shows examples of good and bad tweets in terms of interpretability in terms of consisting mostly of textual data, legitimacy by means of being real person and not a bot.

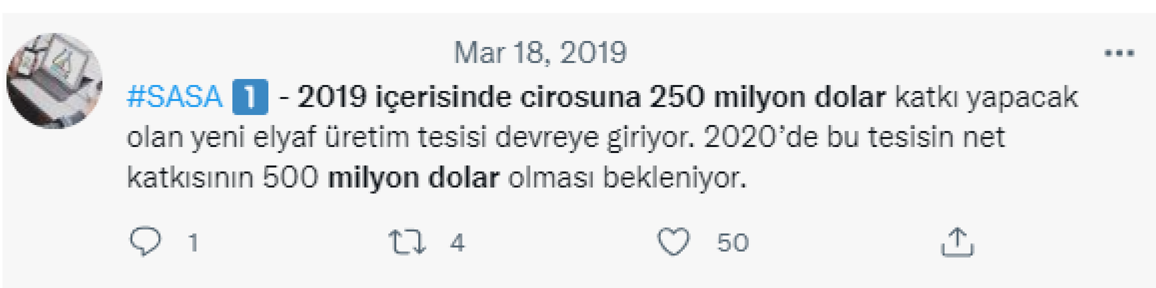


Figure 3.1: Example of an interpretable and authentic tweet

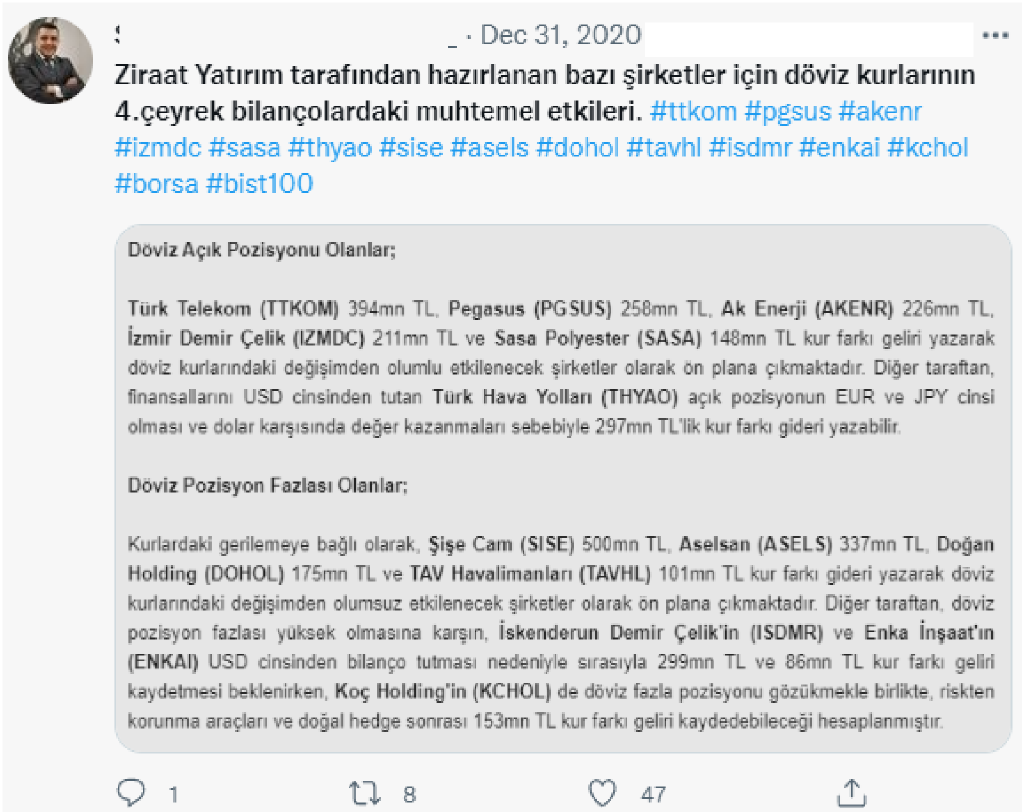


Figure 3.2: Example of legitimate but not interpretable tweet

Figure 3.3: Example of a bot's tweet

A Twitter bots tweets are generally repetitive and frequently use hashtags to collect attention. Therefore, these types of tweets can be discarded when building an ML model. In this study, Twitter bots could not be identified therefore they are not discarded.

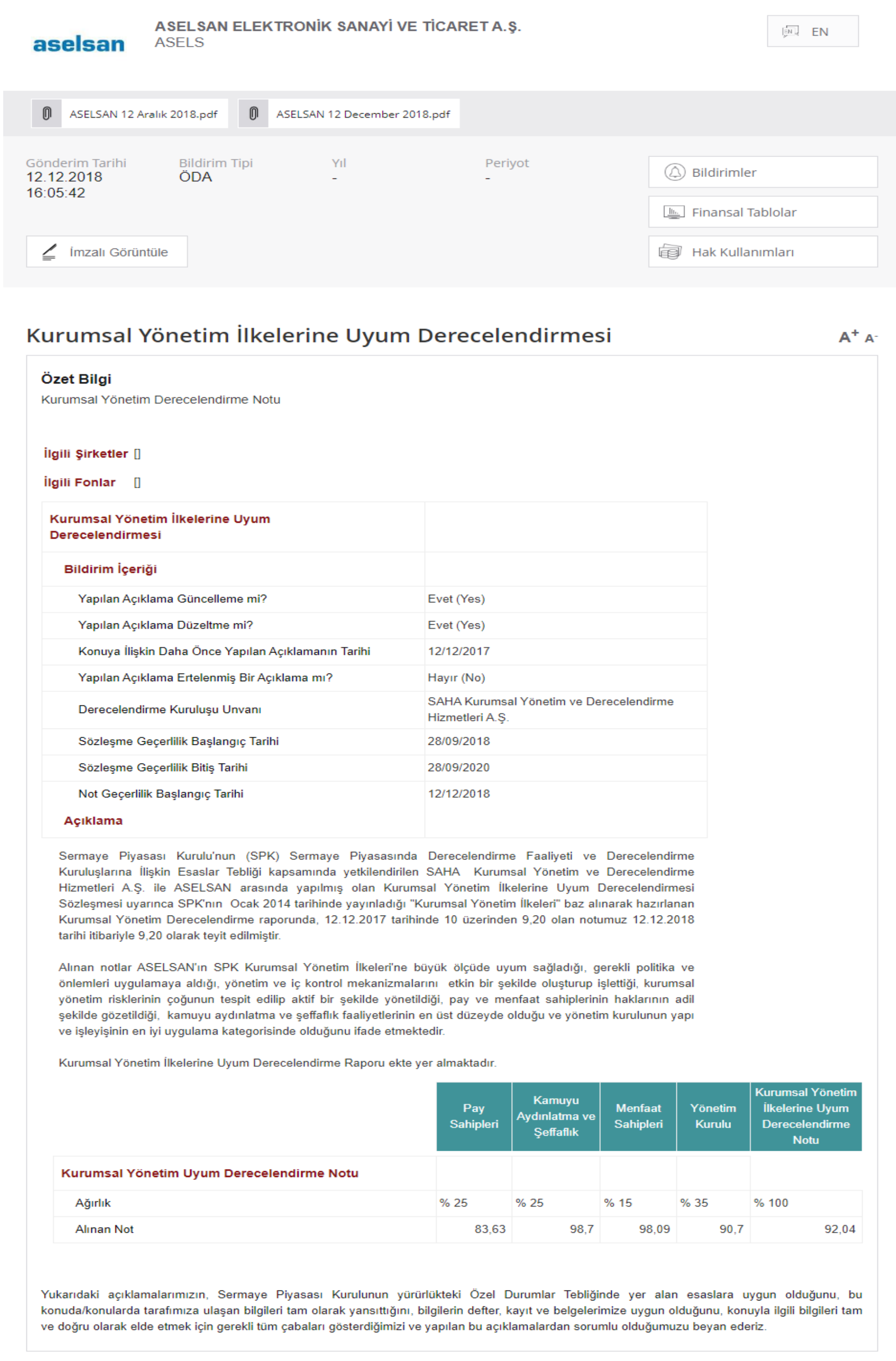
3.2 Public Disclosure Data Collection

For public disclosures in KAP data web scraping tools are used to gather all disclosures about companies listed under BIST30 index.

In total we gathered **20.667** public disclosures from KAP's website between 1st of January 2016 and 1st of January 2021. **1294** days of trading occurred inside the timeframe.

Public Disclosures are in html format, and they are not accessible in structured data formats. Therefore, parsing of html files to extract relevant information is done by removing html tags and repeated content in all disclosures.

As the main source of information regarding companies that is used in this study gathering of Public Disclosures without data loss is not possible because some of the data inside the disclosures are not relevant such as html tags, images, or static items.



aselsan

ASELSAN ELEKTRONİK SANAYİ VE TİCARET A.Ş.
ASELS

EN

ASELSAN 12 Aralık 2018.pdf | ASELSAN 12 December 2018.pdf

Gönderim Tarihi	Bildirim Tipi	Yıl	Periyot
12.12.2018 16:05:42	ÖDA	-	-

Bildirimler

Finansal Tablolar

Hak Kullanımları

İmzalı Görüntüle

Kurumsal Yönetim İlkelerine Uyum Derecelendirmesi

A+ A-

Özet Bilgi

Kurumsal Yönetim Derecelendirme Notu

İlgili Şirketler []

İlgili Fonlar []

Kurumsal Yönetim İlkelerine Uyum Derecelendirmesi	
Bildirim İçeriği	
Yapılan Açıklama Güncelleme mi?	Evet (Yes)
Yapılan Açıklama Düzeltme mi?	Evet (Yes)
Konuya İlişkin Daha Önce Yapılan Açıklamanın Tarihi	12/12/2017
Yapılan Açıklama Ertelenmiş Bir Açıklama mı?	Hayır (No)
Derecelendirme Kuruluşu Unvanı	SAHA Kurumsal Yönetim ve Derecelendirme Hizmetleri A.Ş.
Sözleşme Geçerlilik Başlangıç Tarihi	28/09/2018
Sözleşme Geçerlilik Bitiş Tarihi	28/09/2020
Not Geçerlilik Başlangıç Tarihi	12/12/2018
Açıklama	

Sermaye Piyasası Kurulu'nun (SPK) Sermaye Piyasasında Derecelendirme Faaliyeti ve Derecelendirme Kuruluşlarına İlişkin Esaslar Tebliği kapsamında yetkilendirilen SAHA Kurumsal Yönetim ve Derecelendirme Hizmetleri A.Ş. ile ASELSAN arasında yapılmış olan Kurumsal Yönetim İlkelerine Uyum Derecelendirmesi Sözleşmesi uyarınca SPK'nın Ocak 2014 tarihinde yayınladığı "Kurumsal Yönetim İlkeleri" baz alınarak hazırlanan Kurumsal Yönetim Derecelendirme raporunda, 12.12.2017 tarihinde 10 üzerinden 9,20 olan notumuz 12.12.2018 tarihi itibariyle 9,20 olarak teyit edilmiştir.

Alınan notlar ASELSAN'ın SPK Kurumsal Yönetim İlkeleri'ne büyük ölçüde uyum sağladığı, gerekli politika ve önlemleri uygulamaya aldığı, yönetim ve iç kontrol mekanizmalarını etkin bir şekilde oluşturup işlettiği, kurumsal yönetim risklerinin çoğunun tespit edilip aktif bir şekilde yönetildiği, pay ve menfaat sahiplerinin haklarının adil şekilde gözetildiği, kamuyu aydınlatma ve şeffaflık faaliyetlerinin en üst düzeyde olduğu ve yönetim kurulunun yapı ve işleyişinin en iyi uygulama kategorisinde olduğunu ifade etmektedir.

Kurumsal Yönetim İlkelerine Uyum Derecelendirme Raporu ekte yer almaktadır.

	Pay Sahipleri	Kamuyu Aydınlatma ve Şeffaflık	Menfaat Sahipleri	Yönetim Kurulu	Kurumsal Yönetim İlkelerine Uyum Derecelendirme Notu
Kurumsal Yönetim Uyum Derecelendirme Notu					
Ağırlık	% 25	% 25	% 15	% 35	% 100
Alınan Not	83,63	98,7	98,09	90,7	92,04

Yukarıdaki açıklamalarımızın, Sermaye Piyasası Kurulunun yürürlükteki Özel Durumlar Tebliğinde yer alan esaslara uygun olduğunu, bu konuda/konularda tarafımıza ulaşan bilgileri tam olarak yansıttığını, bilgilerin defter, kayıt ve belgelerimize uygun olduğunu, konuyla ilgili bilgileri tam ve doğru olarak elde etmek için gerekli tüm çabaları gösterdiğimizi ve yapılan bu açıklamalardan sorumlu olduğumuzu beyan ederiz.

Figure 3.4: KAP disclosure example

Figure 3.4 shows an example of a KAP disclosure. This example illustrates information that is not relevant to a machine learning model such as headers and titles as well as information that is relevant residing inside of “Açıklama” part. Sentiments can best be derived from proper sentences due to a context that is intended to be in the sentence.

3.3 Twitter Data Preparation

While collecting the twitter data we also gathered extra metadata belonging to the tweets including like count (l), reply count (RE), retweet count (RT), follower count(f). Using these values, we generated an interaction score (IS) (2) by normalizing each of the values and generating Z-score values with the normalized values (1).

$$z(f) = (f - \mu) / \sigma \quad (1)$$

$$IS(l, RE, RT, f) = log(l)z(f)\sqrt{}(RE + RT) \quad (2)$$

After the calculation of interaction score, we calculate a final score value with multiplication of interaction scores and sentiment scores that is used in the final model.

Twitter is being used for various purposes. We use tweets to capture investor sentiment but according to [33] 5% of all Twitter users are bots and they make up over 44% of interaction on the platform. We found that a significant amount of our data is made up by bots. To remove bot posts from our data, we filtered repeated tweets that have the exact same content. There are bots that post tweets that are very similar but not exactly same with another tweet. Filtering these types of tweets are not feasible therefore is left as is. User based filtering is a field of study for these applications. The impact of such filtering to our work is proven to be negligible in terms of accuracy improvement therefore, user-based filtering is not applied for twitter data.

3.4 Public Disclosure Text Preparation

Our time interval for data collection is between January of 2016 and 2021. This period is chosen because it presents a time of high volatility and increased informational awareness in public both present on Twitter and in public disclosures.

Disclosures are in text format without any length limit therefore their length varies greatly, information regarding a disclosure's length can provide an indication of its relevance towards the change in price.

To enrich public disclosures data, we formed the length of disclosures into a feature for our ML model to capture readability, we also added the ratio of numeric to alphabetic characters in each disclosure to capture the interpretability. Higher ratios are expected to have higher interpretable results.

To obtain sentiment scores using information relevant to the stock we preprocessed public disclosures by removing all the html tags from the texts as well as English versions that exist in 5% of the disclosures.

3.5 Sentiment Score Calculation

For sentiment score calculations for both KAP and twitter datasets the sentiment BERT model introduced in [3], [18] is used, the model has a character limit of 512. For longer KAP contents we divided up the contents into chunks of size 512 characters and used the average of the calculated sentiments. These sentiments are combined for the purpose of finding their overall impact in a manner that can represent all the contents of the disclosure therefore can hold up to the sentiment analysis of NLP models. This approach is necessary for using pre-trained models such as BERTurk. While combinatorial relations inside the disclosures are not properly represented with this method these relations proved to be of little impact on the accuracy.

Table 3.1 shows the number of positive and negative sentiments derived from Twitter and KAP disclosures.

Table 3.1: Sentiments of Data Points of Different Platforms

Sentiment	**Twitter**	**KAP**
Positive	6364	3914
Negative	3013	5463

3.6 Market Data Preparation

We chose prices of Gold, USD/TRY, DJI and BIST30 indices as market indicators. Price of Gold per Ounce is generally seen as an indicator for perceived risk in financial markets. Ratio of USD/TRY is used for benchmarking against inflation in Turkish Lira. We included DJI index as it represents the global trend in stock market investments.

For the BIST30 index and the price change of individual stocks we calculated 1, 3, 5, 8, 10, 20, 30, 60, 90 days change ahead of the trading day to understand the expected descent of news' relevance and our ability to predict the price changes going into the future.

Trading volume is the number of shares bought and sold in a stock exchange for a given stock in each day, it represents the activity and attention on a company's stock by investors [22]. We calculated volume changes of stocks and included them to predict volatility. Volume changes vary between -100% to 500%, which is grouped into equal parts of 6, 12, 24 bins respectively. -100% means no trading is done for that stock in that day. 500% means 4 times more trading compared to its all-time average trading volume is done on that stock for that day. Table 3.2 shows the 6 bins with respect to their intervals:

Table 3.2: Bin Ids with Range of Bins for Volume Change Prediction

Bin Id	Volume Change Range
0	(-1.0, 0.0)
1	(0.0, 1.0)
2	(1.0, 2.0)
3	(2.0, 3.0)
4	(3.0, 4.0)
5	(4.0, 5.0)

3.7 Data Consolidation

Trading days refer to days that a company's stock can be traded within a stock exchange. We consolidated all the gathered data with respect to trading days. Our analysis is conducted for both a combined model where all stocks are joined together, and predictions are made on a single model, and we also trained separate ML models for each stock.

To combine different stocks, we used one hot encoding method where each stock is turned into a column with a one or a zero value. We combined stocks into a single machine learning model to provide a unified system and be able to build more complex models.

In conjunction with our research focus we train different machine learning models for each company to observe the impact of disclosures on single stocks. This allow us to achieve higher confidence towards a trading strategy as well as provide us the ability to distinguish information efficiency on news sentiments of companies. Some companies are more affected by disclosures than others.

Distinguishing companies and building multiple ML models proved to be more successful in predicting stock price changes therefore further study is conducted on building ML models for each company and combining results by taking maximum, minimum and mean of predictions made on single stocks.

Companies do not publish public disclosures every trading day therefore to make use of data in those days, imputation is made. Imputation [28] is the practice of determining what values should be used when there are gaps in a dataset.

Zero fill method involves filling the gaps with 0 values. Forward filling method refers to using previous data points. Decay filling means filling values by using several data points before and assigning decaying weight with respect to their distance. Decay fill is intuitively more appropriate for news because news don’t lose their informational value in a single day but that value decays over time. Results showed no significant difference between zero fill and decay fill in imputation and because of decay fills computational complexity zero fill method is used.

Table 3.3: Disclosure and Tweet Counts of all Companies

Ticker	Disclosure Count	Tweet Count		Ticker	Disclosure Count	Tweet Count
AKBNK	2180	49341		KRDMD	346	64914
ARCLK	316	32019		PETKM	323	88333
ASELS	412	115649		PGSUS	349	45682
BIMAS	326	30313		SAHOL	291	26328
DOHOL	454	24594		SASA	305	71718
EKGYO	1761	56089		SISE	435	62696
EREGL	363	49431		TAVHL	274	30464
FROTO	284	17424		TCELL	545	47081
GARAN	2122	115628		THYAO	367	125831
GUBRF	401	17541		TKFEN	190	30559
HALKB	1197	69017		TTKOM	457	39402
ISCTR	1919	40266		TUPRS	302	60041
KCHOL	264	27306		VAKBN	1929	31811
KOZAA	155	42708		VESTL	359	28756
KOZAL	289	45345		YKBNK	1751	33549

Following columns are used in this study, these are:

- Day10Change (Stock price change in the last 10 trading day)
- Day1Change (Stock price change in the last trading day)
- DJIDay1Change (Index value change in the last trading day)
- GOLDDay1Change (Index value change in the last trading day)
- USDDay1Change (Index value change in the last trading day)
- XU30Day1Change (BIST30 Index value change in the last trading day)
- KAP-score (sentiment scores obtained from KAP disclosure)
- Tweet-score (sentiment scores obtained from Twitter)
- Predicted value: Stock price will go up or down

4. EXPERIMENT RESULTS AND DISCUSSION

Accuracy of the models are calculated using F1 score and 80% of the data is used for training and 20% is used for testing. Stocks price changes are time dependent therefore random sampling is not feasible for our case. Division for training and test is done by splitting data timewise into before and after a certain trading day.

To detect the most accurate model for the problem various ML models are tested by means of calculating their F1 score compared on the test set.

Table 4.1 shows the comparison of F1 scores of 9 different Machine Learning models:

- Logistic Regression (LR)
- Bagging Classifier (BagC)
- Random Forest (RandF)
- AdaBoost (AdaB)
- K Nearest Neighbors (KNN)
- Decision Trees (DecisionT)
- ExtraTrees Classifier (ExtraTC)
- Support Vector Classifier (SVC)
- XGBoost (XGB)

These algorithms are applied to get maximum, minimum, and mean predictions of BIST30 individual stocks' price changes. Due to their faster processing times, novelty, and interpretability of results, four of the models (LR, RandF, AdaB, DecisionT and ExtraTC) are chosen to be used in our studies.

As the target of this study to find the highest accuracy and most reliable prediction results using ML models many models are tested and some are found to be more suitable to use in this work.

Table 4.1: F1 Scores for Stock Price Prediction with different ML Models

	Max	**Min**	**Mean**
LR	58.6	47.5	52.3
BagC	62.7	52.8	56.9
RandF	60.1	51.9	56.2
AdaB	63.6	51	56.6
KNN	**62.5**	**52**	**57.3**
DecisionT	59.9	52.2	56.6
ExtraTC	59.6	52.7	56.4
SVC	62.6	48.3	54
XGB	62	51.3	57.3

Different ML models yields different accuracy levels; therefore, some are more fit to be used in this area. But most of the ML models yield similar test accuracies, this can be interpreted as a certainty towards the relationship between input values to the model and predicted values. Table 4.1 shows F1 score that is around 61 is a reliable estimate for prediction accuracy of a ML learning with the corresponding data.

Table 4.2 shows the results of predictions using 2 bin predictions with 3 imputation methods and best results are achieved by using decay and zero fill and to achieve more precise results further research is done by only using zero fill method.

Decay imputation fills empty data by looking at previous data.

Zero imputation fills only empty data with 0 as default value.

Forward imputations fill gaps in data by looking at future data.

Decay and forward imputations are done by multiplying scores of the previous days by a value that decreases as it goes back. Zero fill and decay fill has produces very similar results therefore zero fill method is preferred.

Table 4.2: F1 Scores for Stock Price prediction using different imputation methods

Imputation	**LR**	**RandF**	**AdaB**	**DecisionT**	**ExtraTC**
Forward Fill	38.3	50.1	49.2	49.9	50.1
Decay Fill	52.3	56.2	56.6	56.5	56.3
Zero Fill	52.3	56.2	56.6	56.6	56.1

4.1 Stock Trading Volume Change Prediction

In literature, daily stock volume prediction is an understudied topic in terms of sentiment analysis. Our results portrayed in Table 4.3 shows strong correlation between sentiment scores and volume changes.

Volume changes are calculated by taking average daily volume of a stock and dividing current days volume to the average value. As bin counts increases radical drop in F1 score is observed.

Bins refer to the classes that are being predicted with the ML model. For stock volume prediction the range is between -100% and 500% of the average trading volume. In 6 bins example this interval is split into 6 equal parts and has proven to be most successful.

Table 4.3: F1 Scores for Volume prediction on Different Bin sizes

F1 Score	LR	RandF	AdaB	DecisionT	ExtraTC
6 bins	**74.7**	62.4	70.2	61.8	62.8
12 bins	34.6	34.4	35.7	34.2	34.1
24 bins	17.8	20.2	18.9	20.1	20.1

4.2 Stock Price Change Prediction

In Table 4.4 results of predictions using Twitter and KAP sentiment scores are shown using different Machine Learning models. Maximum, minimum and mean of different stocks' predictions show that different companies have different susceptibility to sentiment across KAP and Twitter.

As the focus of this study stock price prediction for individual companies yields accuracies above 80% but for some companies the F1 Score falls below 50% which indicates a failure of prediction. Mean value of all company's prediction F1 Score is highest in logistic regression and random forest models.

Table 4.4: F1 Scores for Stock Price predictions of different stocks for 2 Bins

F1 Score	**LR**	**RandF**	**AdaB**	**DecisionT**	**ExtraTC**
Max	81.76	81.31	81.58	77.00	76.68
Min	45.9	49.15	5.82	45.81	48.22
Mean	67.82	67.72	59.82	62.85	62.82

Table 4.5 shows the F1 Score decreasing as the resolution of prediction increases. This is the expected behavior of ML models and provides confidence towards connection of the findings to actual real world results of data.

Table 4.5: F1 Scores for mean Stock Price change prediction for different Bins

F1 Score	**LR**	**RandF**	**AdaB**	**DecisionT**	**ExtraTC**
2 Bin	52.7	60.4	62.2	62.7	62.4
4 Bin	51.9	58.9	55.6	61.1	61.1
6 Bin	49.1	54.8	51.9	56.3	56.7
8 Bin	44	48.9	45.8	50.7	50.7
10 Bin	44	48.7	45.6	50.1	50.4

Inclusion of disclosure length and alphabetic/numeric character ratios improves prediction F1 Score of models between 1-5%.

Figure 4.1 shows that first day of prediction yields higher F1 Scores as KAP and Twitter sentiment directly impact the price and after the first day a sharp decrease in F1 Score occurs. Our analysis is conducted for predicting next 1, 3, 5, 8, 10, 20, 30, 60 and 90 days.

In 30 trading days' time we can see the information in the sentiment integrating into the price of the stock and stabilizing afterwards. This means that the F1 Scores of predictions gets higher for longer periods because noise of daily price fluctuations can make predictions harder in the short term. But after 30 days the impact of disclosures and tweets are stabilized.

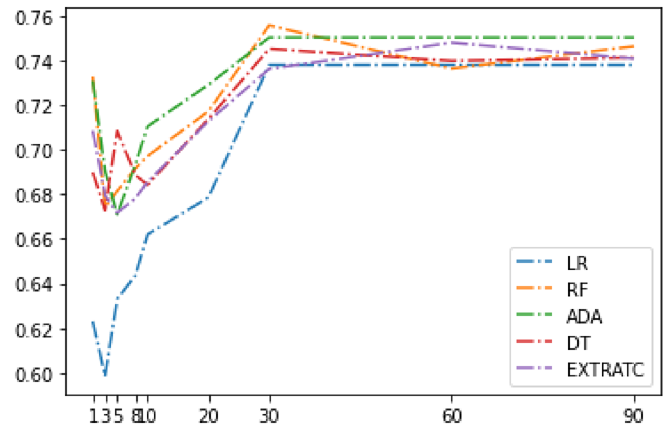


Figure 4.1: F1 scores for price prediction for different periods of time (day)

The increase in price predictability makes long term investment decisions more viable for investors.

5. CONCLUSION AND FUTURE WORK

5.1 Thesis Conclusion

In this study, we focus on building ML models for predicting individual stocks that is compatible to forming a trading strategy based on public disclosures published on KAP and people’s sentiments on Twitter. We use state-of-the-art BERTurk sentiment model for sentiment analysis of tweets and disclosures. To build our dataset and make predictions for days without tweets or disclosures we test several imputation methods to assess their value, days after their publishment.

We append financial data regarding market conditions consisting of daily price changes of BIST30, DJI, USD and Gold per Ounce to improve prediction accuracy and integrate markets’ overall sentiments and risk appetite.

Our study also consists of predicting daily volume changes of stocks as a novel study field. We can predict the volume changes with high accuracy using sentiments from public disclosures and tweets. We analyze the volatility of stocks by predicting daily volume changes of stocks and this provides reliable correlation between the sentiments and volume changes for the upcoming trading day with 74.7% prediction accuracy. But because volume changes are generally considered to be reactionary the prediction falls quickly for further days.

Building a trading strategy using a ML model with accuracy levels higher than 60% can yield higher than market average profits and our results can support such levels of accuracy of our predictions. Our study shows prediction accuracies for BIST30 companies' individual stocks of 67%.

Our study suggests that predicting the stock price changes for longer periods of time provides higher accuracy levels therefore by using larger timeframes and more advanced sentiment models better trading strategies can be formed for long term investment.

It has been found that building individual machine learning models for each company produces reliably higher accuracy therefore creation of a general KAP index is not feasible.

Because higher overall accuracy is achieved using individual ML models for each company our study focuses on individual models for each company.

5.2 Future Work

Using only sentiment scores for disclosures misses out on a lot of information; therefore, more information regarding KAP disclosures will be extracted using word vectors, and these word vectors will be combined for higher F1 scores.

Research on dividend and stock split data to predict and explain the behavior of changes in stock prices goes back to Eugene Fama's 1970 paper on Efficient markets hypothesis [2] and is proven to be reliable information. Compared to the disclosure and tweet data dividend and stock splits occur less frequently but can be integrated into the model for a study that covers a longer period.

Imputation methods in financial news [21] is an unexplored area of research that can provide fruitful results towards assessing the relevance of financial news over a period. In our research we made progress in this area and in the future studies imputation methods can append greater value towards NLP in financial news. NLP models specific to finance in Turkish are not publicly available therefore a clear improvement in this area will be the development and implementation of financial NLP models for Turkish as well as using established financial NLP models in English for KAP disclosures that are published in English.

REFERENCES

[1] Shiller, R. J. (1999). Human behavior and the efficiency of the financial system. Handbook of macroeconomics, 1, 1305-1340.

[2] Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), 383–417. https://doi.org/10.2307/2325486

[3] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, Third Edition, 2012

[4] Aslan, B., & Erdur, R. C. (2020, October). Stock Market Prediction with Deep Learning Using Public Disclosure Platform Data. In 2020 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-5). IEEE.

[5] Acikalin, U. U., Bardak, B., & Kutlu, M. (2020, October). Turkish sentiment analysis using bert. In 2020 28th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[6] Ateş, E., & Güran, A. (2021). Pearson correlation and Granger causality analysis of Twitter sentiments and the daily changes in Bist30 index returns. J. Fac. Eng. Archit. Gazi Univ, 36, 1687-1701.

[7] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.

[8] Bozma, G., & Kul, S. (2020). Can Twitter Forecast Uncertainty of Stocks?. Sosyoekonomi, 28(45).

[9] Çörtük, O., & Erten, M. (2016). Türkiye'de Kamuyu Aydınlatmanın Sermaye Piyasasına Etkisi. İstanbul Üniversitesi İşletme Fakültesi Dergisi, 45(1), 65-77.

[10] Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications, 42(24), 9603-9611.

[11] Sabuncu, İ. (2021). Forecasting Stock Value Based on Data from Social Media and Investment Instruments. Acta Infologica, 5(2), 267-285.

[12] Pasupulety, U., Anees, A. A., Anmol, S., & Mohan, B. R. (2019, June). Predicting stock prices using ensemble learning and sentiment analysis. In 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) (pp. 215-222). IEEE.

[13] You, W., Guo, Y., & Peng, C. (2017). Twitter's daily happiness sentiment and the predictability of stock returns. Finance Research Letters, 23, 58-64.

[14] Gunduz, H., & Cataltepe, Z. (2015). Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection. Expert Systems with Applications, 42(22), 9001-9011

Gunduz, H., & Cataltepe, Z. (2015). Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection. Expert Systems with Applications, 42(22), 9001-9011.

[15] Gunduz, H., Yaslan, Y., & Cataltepe, Z. (2017). Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations. Knowledge-Based Systems, 137, 138-148.

[16] Kilimci, Z. H., & Duvar, R. (2020). An efficient word embedding and deep learning based model to forecast the direction of stock exchange market using Twitter and financial news sites: a case of Istanbul stock exchange (BIST 100). IEEE Access, 8, 188186-188198.

[17] Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Information Processing & Management, 57(5), 102212.

[18] Li, Y. H., & Jain, A. K. (1998). Classification of text documents. The Computer Journal, 41(8), 537-546.

[19] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. Expert systems with applications, 42(1), 259-268.

[20] Yildirim, S. (2020). Comparing Deep Neural Networks to Traditional Models for Sentiment Analysis in Turkish Language. In Deep Learning-Based Approaches for Sentiment Analysis (pp. 311-319). Springer, Singapore.

[21] Kaur, S., Sharma, N., & Singh, K. (2018). Missing Value Treatment using Effective Optimization on Data from Multiple Social Media. Research Cell: An International Journal of Engineering Sciences, 30(SP), 135-147.

[22] Lo, A. W., & Wang, J. (2010). Stock market trading volume. In Handbook of financial econometrics: Applications (pp. 241-342). Elsevier.

[23] Guven, Z. A. (2021, September). Comparison of BERT Models and Machine Learning Methods for Sentiment Analysis on Turkish Tweets. In 2021 6th International Conference on Computer Science and Engineering (UBMK) (pp. 98-101). IEEE.

[24] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[25] Mutlu, M. M., & Özgür, A. (2022). A Dataset and BERT-based Models for Targeted Sentiment Analysis on Turkish Texts. arXiv preprint arXiv:2205.04185.

[26] Othan, D., Kilimci, Z. H., & Uysal, M. (2019, December). Financial sentiment analysis for predicting direction of stocks using bidirectional encoder representations from transformers (BERT) and deep learning models. In International Conference on Innovative and Intelligent Technologies (ICIIT-19), Istanbul, Turkey.

[27] Dogra, V., Singh, A., Verma, S., Jhanjhi, N. Z., & Talib, M. N. (2021). Analyzing DistilBERT for Sentiment Classification of Banking Financial News. In Intelligent Computing and Innovation on Data Science (pp. 501-510). Springer, Singapore.

[28] Hunt, L. A. (2017). Missing data imputation and its effect on the accuracy of classification. In Data Science (pp. 3-14). Springer, Cham.

[29] Gunduz, H., & Cataltepe, Z. (2013). Prediction of Istanbul stock exchange (ISE) direction based on news articles. In Conference: The Third International Conference on Digital Information Processing and Communications (ICDIPC2013) (pp. 320-330).

[30] Gündüz, H., Yaslan, Y., & Çataltepe, Z. (2018). Stock market prediction with deep learning using financial news. In 2018 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[31] Guo, X., & Li, J. (2019). A novel twitter sentiment analysis model with baseline correlation for financial market prediction with improved efficiency. In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 472-477). IEEE.

[32] Al-Radaideh, Q. A., Assaf, A. A., & Alnagi, E. (2013, December). Predicting stock prices using data mining techniques. In The International Arab Conference on Information Technology (ACIT'2013) (pp. 1-8).

[33] Iglesias Caride, M., Bariviera, A. F., & Lanzarini, L. (2018). Stock returns forecast: An examination by means of artificial neural networks. In Complex Systems: Solutions and Challenges in Economics, Management and Engineering (pp. 399-410). Springer, Cham.

[34] Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A systematic review of fundamental and technical analysis of stock market predictions. Artificial Intelligence Review, 53(4), 3007-3057.

[35] Ajinkya, B. B., & Jain, P. C. (1989). The behavior of daily stock market trading volume. Journal of accounting and economics, 11(4), 331-359.

[36] Göker, İ. E. K., Eren, B. S., & Karaca, S. S. (2020). The Impact of the COVID-19 (Coronavirus) on The Borsa Istanbul Sector Index Returns: An Event Study. Gaziantep University Journal of Social Sciences, 19(COVID-19 Special Issue), 14-41.

[38] Chau, M., & Vayanos, D. (2008). Strong-form efficiency with monopolistic insiders. The Review of Financial Studies, 21(5), 2275-2306.