University of Tartu

Faculty of Science and Technology

Institute of Computer Science

Ardi Aasmaa

Predicting stock price based on media monitoring

Master's Thesis (30 ECTS)
Software Engineering Curriculum

Supervisors:

Rajesh Sharma PhD

Tartu 2020

Abstract:
Predicting stock price based on media monitoring

Using automated systems for finding investment ideas becomes more popular every year and the models are getting more complex. In recent years a lot of studies have been conducted that have researched the possibilities of using social media sentiment as input for stock prediction models. However, the results have been contradicting as the problem is complex. In this research, data was collected from Twitter about Standard Poor's 100 companies over a period of six months. Also, financial data with one minute interval was collected from Alpha Vantage. Five different machine learning algorithms were used to predict maximum profit and maximum loss for the prediction horizon of five trading days. It was investigated whether adding social media based features to financial data based features would improve the results and if so, then tweets from what type of users would give the highest information gain. It was found out that adding social media data as input is beneficial for both, predicting maximum loss and maximum profit. For the explainability part, Shap library was used. As found out, features extracted from financial data were most important. For social media based features, most information was gained from tweets posted by news agencies and by users having relatively few followers.
CERCS: P170 Computer science, numerical analysis, systems, control
Keywords: machine learning, sentiment analysis, social media monitoring

Aktsia hindade ennustamine sotsiaalmeedia monitooringu põhjal

Aasta aastalt on investeerimisideede leidmiseks üha enam kasutatud automaatseid süsteeme ja need on muutunud järjest keerulisemaks. Viimastel aastatel on läbi viidud mitmeid uuringuid selle kohta, kuidas kasutada sotsiaalmeedia andmeid aktsiahindade ennustamiseks. Senised tulemused on olnud vasturääkivad, kuna ülesanne on keeruline ja sõltub väga paljudest faktoritest. Antud uurimustöös koguti Twitterist kuue kuu jooksul andmeid 102 Standard Poor'i indeksisse kuuluva firma kohta. Lisaks sotsiaalmeedia andmestikule kasutati finantsandmeid, mis koguti Alpha Vantage'i kaudu. Kasutati viite masinõppe algoritmi, et ennustada maksimaalset kasumit ja kahjumit viie järgneva kauplemispäeva jooksul. Uuriti, kas sotsiaalmeedia andmete lisamine mudelile parandab tulemusi ja kui parandab, siis millist tüüpi kasutajate postitused annavad ennustamisel kõige suurema kasu. Leiti, et sotsiaalmeedia andmete lisamine on kasulik nii maksimaalse kasumi, kui maksimaalse kahjumi ennustamiseks. Mudeli väljundi selgitamiseks kasutati teeki Shap. Leiti, et kõige olulisemateks sisendparameetriteks olid finantsandmed. Sotsiaalmeedia puhul oli kõige rohkem kasu uudisteagentuuride poolt postitatud sõnumitest ja sõnumitest, mille postitasid suhteliselt väikese jälgijate arvuga kasutajad.
CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)
Keywords: masinõpe, sentimentaalne analüüs, sotsiaalmeedia monitooring

Contents

1 Introduction

It has been estimated that on an average trading day in the US stock market about 50 to 60% of trading is done fully automatically by trading programs [7]. The strategies those programs are using vary significantly, but the goal remains the same. Investors have an idea of how to make a profit and they are automating the process from collecting data to making transactions. The most basic models are usually making predictions based purely on price and trading volume statistics. More advanced models are also trying to take advantage of finding the correlation between some stocks, instruments or markets. For example, a trading strategy could be to buy a stock when its price is ten percent lower than the previous 50 days rolling average or when the price of another stock from the same sector increases significantly. [12]

Investors are constantly looking for new opportunities for making their predictions more accurate. Because the usage of social media has tripled in the last decade, the correlation between social media sentiment and stock prices has been widely researched in recent years [21]. The results of previous studies are hard to compare, mainly because the data sets used are not public and study periods are relatively short. Researchers are still arguing whether price movements and social media sentiment are significantly correlated and if they are then for which prediction horizons (intraday, some days or months) and for which markets or sectors? [8]

Most of the related studies have simplified the problem, by predicting only the direction (rise, fall) of the price change, not the magnitude. From the investors' point of view, knowing only the price change direction is not enough. A small increase in price could still mean loss because of trading costs and other taxes. The most popular target has been the close price of the next trading day. However, for an investor, it would be significantly more beneficial to accurately predict the maximum price of some period rather than the closing price. For example, if the stock price is currently 90 dollars and it is predicted to close at 88 dollars the next trading day it would not seem like an investing opportunity. On the other hand, if it would be predicted that at some point before the closing time of the next trading day the price would maximally rise to 95 dollars, a trigger could be set to sell the stock as soon as it exceeds some predefined level.

In this study, data was collected from Twitter over a period of 6 months for the S&P100 (Standard and Poor's 100) companies. S&P100 is a stock market index that contains 102 most valuable stocks from the U.S. stock market. [30] Standard and Poor's index was picked because it is one of the most important and known indexes in the world and the companies belonging to there are mostly very popular on social media. [5] In addition to Twitter data, price and volume data were collected from Alpha Vantage with one minute interval. After discarding irrelevant tweets, sentiment scores were found and the results were aggregated by companies they mention. Sul et al. divided tweets into two groups depending on if the tweet author had more than 172 followers or less. They were suggesting that information spreads significantly faster and has higher impact if it is posted by a user who has a lot of followers. [29] In this

thesis, this idea was developed further. Tweets were grouped into eight groups based on who was the author of the tweet. The first group of authors contained official company-related accounts. The second group contained major news agencies' accounts (BBC, NY Times, Wall Street Journal, etc). Another group had users brought out by Forbes Magazine to be the most influential finance-related Twitter accounts [26]. Other users were divided into five groups based on how many followers they have. In addition to aggregating sentiment scores by user types, another grouping was done, based on if the tweet contains a cashtag or not. Cashtag is a dollar symbol that is followed by a stock symbol. For example $AAPL. On Twitter, they are used to mark that tweet is about a certain stock. [14] With that grouping, it is found out, if the assumption that tweets containing cashtags have a higher correlation to stock prices holds.

The contribution of this thesis is exploring the possibilities of using social media data for predicting the maximum profit and maximum loss of the next five trading days. Using maximum and minimum price instead of closing price as in previous studies could be used to find additional investment opportunities. An additional objective is to find out whether combining social media based features with financial data helps to improve prediction results. Also to compare which kind of Twitter users affect the model performance most significantly. During this research, a program was made to collect and process data from a live feed.

This thesis is structured into six sections.

1. In the related studies chapter, an overview is given about the most important studies that have researched the usage of social media data for stock price prediction.

2. Chapter 3 describes how social media and stock price data were collected for this thesis. In addition, multiple charts are shown to illustrate the main characteristics of the data.

3. The next chapter describes how the data was processed and the features that were extracted for building the machine learning models.

4. In the results chapter, the performance of five different models are compared. The importance of features extracted from social media data compared to price statistics based features are also evaluated.

5. In the final chapter, it is discussed whether social media data could be profitably used for trading and proposals for future studies are made.

2 Related work

This section gives an overview of the most relevant related studies. In recent years a lot of effort has been put into researching the possibilities of using social media data as input for stock price prediction. Numerous researches have been conducted about different social media platforms (Twitter, Facebook, StockTwits, Xueqiu, etc.), instruments (stocks, gold, currencies, cryptocurrencies, etc.) and markets (US, China, India, etc.). [21] However, the problem of predicting stock prices is complicated and depends on a high amount of variables [27]. To better compare the results with this study, the following inclusion criteria were used for the related studies:

- Twitter as a social media for collecting the data;
- the stocks are listed on one of the United States markets;
- at least 20 companies are included in the study;
- the study period is longer than three months;
- the dataset contains more than 2 million tweets;
- the study period starts 2008 or later.

The reason why studies covering other social media platforms were excluded is that they have so different characteristics that the results about one wouldn't say much about others. In this research, Twitter was taken as a social media representative because it is one of the largest social media platforms. When compared to StockTwits, which is a platform used especially for posting about stocks, the posts are less formal and more subjective. However, the amount of Twitter users is 1500 times larger than StockTwits. While StockTwits represents more what professional investors think about stocks, Twitter represents the sentiment of a larger crowd, but also includes the professionals. Other platforms have smaller data feed or are mainly used for other purposes. As for excluding the studies containing only a couple of companies, the reason is that the social media popularity of different companies varies a lot. The effect the sentiment has on prices depends among other factors on the capitalization of the company. The higher the amount of small investors owning the stock the higher is the sensitivity of the price. [23] Thus when trying different ideas, the number of companies must be large enough. Otherwise, it would be hard to prove that the idea applies to a wider range of companies and for periods outside the study period.

There are some differences about what was collected from Twitter, as can be seen from Table 1. For example, in the study [6], the authors collected all tweets that directly expressed the emotions of the tweet authors. They were interested in how the overall mood of the Twitter users affects the financial markets, rather than the sentiment about some specific stock. In

the study [15], they were trying to replicate the results of work [6] and thus used the same dataset. The rest of the studies selected a group of companies and collected only tweets that were mentioning those companies. Since Twitter introduced cashtags, it is the most convenient way to mark that the tweet is about a certain stock. Thus cashtags were the most popular filters in the related studies. Some of the researchers added additional keywords to the filters they were searching for. For example, the authors of research [19] went further by adding hashtags ('#AAPL') and pure stock symbols ('AAPL'), because they noticed that these versions are often used as alternatives for cashtags. The largest amount of data was collected in the study [16]. They included messages mentioning the company names, products or services as well, which significantly increased the daily collection volume. They collected on average 1.1 million tweets per company per month (200 million in total) compared to research [29], where 310 tweets per company per month were collected. Another reason why the data feed had so large difference is that, in research [29], the authors were collecting tweets about companies that were much smaller and less known.

Table 1: Dataset description.

Reference	Study period	Filters	Dataset size	Market
Sul et al. (2014) [29]	03.2011 - 01.2013	cashtags	3.4 million	Stocks (S&P500)
Bollen et al. (2011) [6], Lachanski, Pav (2017) [15]	28.02.2008 - 19.12.2008	all tweets (not only finance-related) expressing emotions ("I feel", "I am feeling", "I'm feeling", "I don't feel", "I'm", "Im", "I am" and "makes me")	9.8 million	Index (DJIA)
Makrehchi et al. (2013) [19]	27.03.2012 - 13.07.2012	stock symbols (pure, with prefix $ and with prefix #), company name	30 million	Stocks (DJI30)
Li et al. (2017) [16]	10.2011 - 03.2012	company names, products and services related to the company	200 million	30 random companies from (NYSE, NASDAQ)
Oliveira et al. (2017) [23]	22.12.2012 - 29.10.2015	cashtags	31 million	all stocks traded in the US markets (nearly 3800 stocks)

The next important difference between the related studies is how they cleaned the textual data and which methods were used for finding sentiment scores of individual tweets. Tweets are short messages, with a maximum length of 140 characters. However, what makes the sentiment extraction complicated is that tweets include a significant amount of slang, emoticons, URLs, cashtags ($), hashtags (#) and user mentions (@) as part of the sentences. The most common approach for extracting sentiment values was using lexicon-based methods, as can be seen from Table 2. In most of the studies, general lexicons (for example Harvard-IV dictionary) were used for the task, but the authors of the study [23] went further by using a dictionary that had been specially adapted for financial tweets. In their study, they also used Kalman Filter to aggregate sentiment values of different monthly surveys to daily sentiment scores of tweets. In research [19], despite using public dictionaries, they created their own. They were looking for significant price changes (more than + or - 3%) and marking the tweets that had been posted previously on the same day accordingly as positive if the price rose or negative if it dropped. In the study [16], a wide range of tweets about the companies and their products and services were collected. They created a conceptual map for the companies and their products. The total sentiment score was found by aggregating scores of each concept. The scores of the concept were adjusted by different margins. The authors of the [29], discarded all tweets that had been retweeted or posted by a user that had more than 172 followers. Other studies didn't take followers count and retweet count into consideration. Several researchers [6, 15, 19] tried to find out which moods and how strongly were expressed by the tweets. They aggregated different mood scores and tried to predict stock prices from that.

After cleaning the data and finding sentiment values, machine learning models were used to predict prices for different timeframes. Table 2 shows that researchers experimented with different prediction horizons from the end of the next trading day (T+1), which was the most common target, up to 20 days forward (T+20). In some studies [16, 19], only the price movement direction was predicted. The latter [19] used a quite simple approach. They were calculating the net sentiment of the current day. If it was higher than zero they predicted the price would go up and vice versa. For investors, it is significantly more useful to predict the magnitude of the price change, as minor price increases would still not be profitable when also considering the trading fees. Thus other studies predicted the price itself, as a regression problem. In research [29], they used the Gradual Information model for prediction. They suggested that the amount of time it takes for sentiment value changes to be reflected on prices depends on how many times the tweets had been retweeted and how much followers the author had. They argued that only sentiment of tweets from users with few followers are correlated to price changes a couple of days later, as highly followed users affect the prices faster, in a couple of minutes. Some researchers [6, 15] were also looking for the correlation between different moods (six moods) detected from tweets and stock prices by using Granger causality analysis. Couple of studies also compared the performance of different algorithms [16, 23].

The studies that were collecting social media data about companies were able to prove

Table 2: Sentiment analysis.

Reference	Sentiment extraction	Sentiment values	Features
Sul et al. (2014) [29]	word analysis (Harvard-IV dictionary)	(NEG, POS)	followers count (FEW, MANY), retweets count
Bollen et al. (2011) [6], Lachanski, Pav (2017) [15]	GPOMS, OpinionFinder	(NEG, POS) 6 moods	6 moods, positive/negative score
Makrehchi et al. (2013) [19]	trained tweets classifier using data from significant changes (over 3%)	(NEG, POS) 4 moods	4 moods (dictionary), positive/negative classification (model trained based on significant changes (+-3%))
Li et al. (2017) [16]	SMeDA-SA, NLP methods to classify	(Positive+, Positive, Neutral, Negative, Negative-)	concept maps (company, service, product, product attributes), gives weights to all concepts to calculate aggregated sentiment about the company
Oliveira et al. (2017) [23]	lexicon adapted to financial tweets (Oliveira et al. 2016)	sentiment score	uses Kalman Filter to aggregate survey sentiment indicators (AAII, II, UMSX, Sentix) to daily Twitter sentiment

that there exists a significant correlation between Twitter data and stock prices. The authors of study [6], who were collecting all tweets that were expressing emotions claimed that there exists significant causality relation for mood "Calm". They used Granger causality analysis. However, they collected data for 10 months but used only 19 day period for evaluation. The goal of research [15], was to replicate the results of the study [6] by acquiring the same dataset and using the same techniques. Instead of using only 19 day period, they used the whole 10 month data. They were not able to find a correlation between any of the moods and stock prices and brought out the weaknesses of the study [6]. As can be seen from Table 4 different studies are hard to compare, as different evaluation methods were used. Only a couple of them [16, 23] compared different algorithms. In both of the works, good results were obtained by support

Table 3: Stock price prediction.

Reference	Algorithm	Prediction horizon	Prediction type
Sul et al. (2014) [29]	Gradual Information Diffusion model (Hong & Stein, 1999)	T + 1 (closing price of the next trading day), T + 10, T + 20	Regression
Bollen et al. (2011) [6], Lachanski, Pav (2017) [15]	Self-organizing fuzzy neural network (SOFNN)	from T + 2 to T + 6	Regression
Makrehchi et al. (2013) [19]	Calculate the net sentiment of current day, if higher than zero then UP else DOWN	T + 1	Classification (up, down)
Li et al. (2017) [16]	Compare SmeDa-Sa, SVM, C4.5, Naive Bayes	from T + 1 to T + 6	Classification (up+, up, flat, down, down+)
Oliveira et al. (2017) [23]	Compare MR, NN, SVM, RF and EA	T + 1	Regression

vector machine (SVM) models. The only model that was able to outperform SVM was an algorithm created by Li et al. called SMeDA-SA. [16] Li et al. compared the results of different sectors and got the best result for companies belonging to the IT sector. [16] The directional accuracy for the IT sector was 76% which was 10% higher than the mean accuracy of all sectors. As was brought out in the study [23], the results are affected by the degree of capitalization of the stock. Their study indicated that stock prices have a higher correlation with social media sentiment when the capitalization is low. They also compared how different markets perform and obtained the best results for the Standard & Poor's 500 index. Diebold-Mariano test was used to evaluate the statistical significance of the predictions. In some of the studies, trading strategies were created and tested to find out whether the models could be used profitably. The idea of the authors of the study [29], was to buy the stocks whose sentiment score was in the top 10% of the most positive sentiment scores. Additionally, they sold the stocks with the lowest sentiment scores. In the simulation, they used a transfer fee of 0.1%. The transfer fees provided by Estonian banks are more than two times higher [20]. On the other hand, there exist multiple international platforms that don't take any money for making transactions but high fees are collected for keeping the stock overnight. In the simulation, when they set the maximum holding period to be one day and considered also the taxes, their trading algorithm had a 28.57% annual loss. For longer holding periods (10, 20 days) their model was profitable

(15.65%, 11.41%). A trading simulation was also implemented in the study [19]. They used one day holding period, but didn't use trading fees. That way they were able to outperform the S&P 500 market by 20% over a four-month test period.

Table 4: Results.

Reference	Evaluation methods	Results	Profitability simulation
Sul et al. (2014) [29]	SMAPE	T+1 (0.9678), T+10 (0 .9399), T+20 (0 .9216)	T+1 (11.44%, –28.57%), T+10 (17.91%, 15.65%), T+20 (12.59%, 11.41%) transfer fee 0.1
Bollen et al. (2011) [6]	MAPE, directional accuracy, Granger causality analysis	significant Granger causality relation only for mood "Calm", lowest MAPE 1.79, highest accuracy 86.7%	-
Lachanski, Pav (2017) [15]	MAPE, directional accuracy, Granger causality analysis	didn't find significant correlation for any of the prediction horizons	-
Makrehchi et al. (2013) [19]	-	-	beat the S&P 500 index by 20% over the period of 4 month, transfer fees not included
Li et al. (2017) [16]	directional accuracy	SMeDA-SA: (76.12% - best accuracy, 66.48% - mean accuracy), SVM: (70.75%, 63.34%), C4.5: (67.63%, 60.15%), Naïve Bayesian: (65.02%, 59.79%)	-
Oliveira et al. (2017) [23]	Diebold-Mariano test, NMAE	S&P500 NMAE (MR: 7.90, RF: 8.32, SVM 7.87, NN: 8.59, EA: 7.92)	-

In conclusion, the related studies showed promising results and presented some very good

ideas. On the other hand, the field is still relatively new and there exist many opportunities to study every sub-process more deeply. In addition, the way investors act changes constantly and is different for different markets and companies.

3 Data description

In this research data processing pipeline was set up for collecting real-time data from Twitter about 102 companies (S&P 100) traded in the US stock market. Additionally, price statistics (open, close, min, max, volume) were collected about those companies through Alpha Vantage API with one minute interval. The data collection started on 18th November 2019 and lasted six months, that is, until 17th May 2020. Within this time period, 66 million tweets posted by 13,7 million different users were collected. The total dataset size was 45,8GB. A large portion of those tweets (19 million tweets, 13,1GB of data) was discarded after initial processing because it was found that they were not actually mentioning any of the S&P 100 companies.

3.1 Companies selection

S&P 100 was selected because it is one of the most popular, most valuable and most actively traded stock indexes. Companies belonging to that index are also popular on social media. [5] The companies selected for this study belong to eleven different sectors as shown in Figure 1.

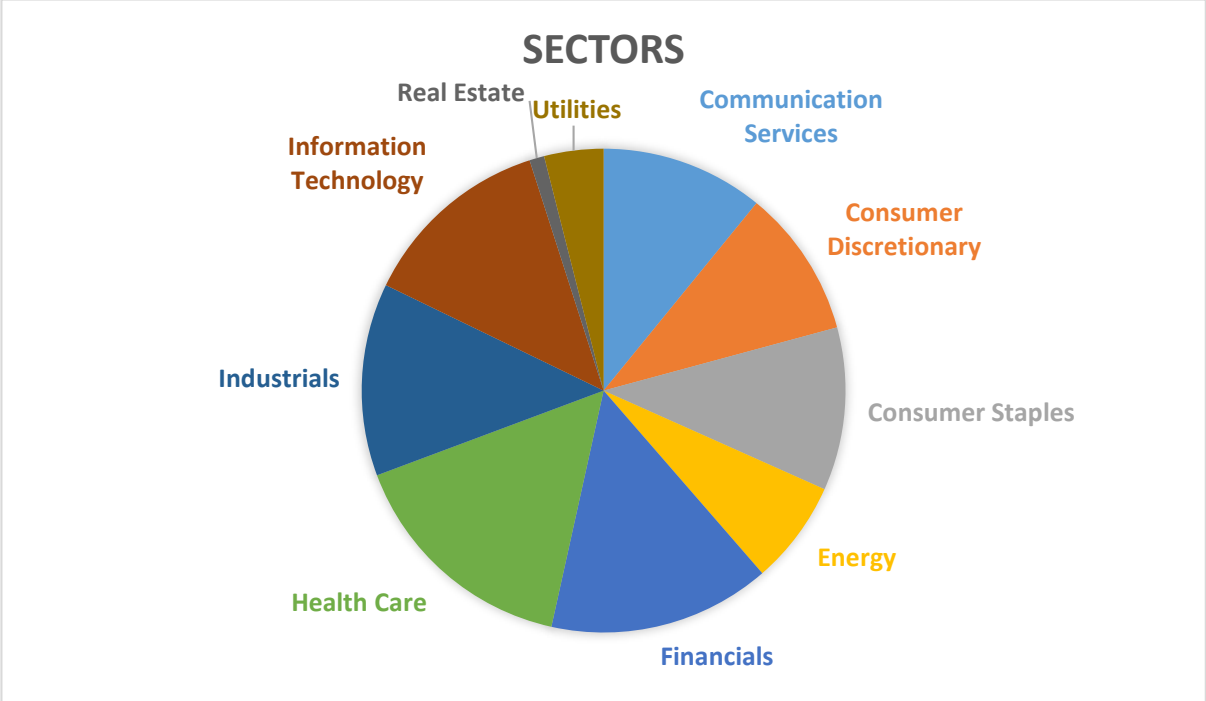


Figure 1: Sectors of companies

As can be seen from Figure 2 companies from Communication Services, Information Technology and Consumer Discretionary sectors are significantly more popular on social media than the rest. One of the reasons might be that products (Apple, Samsung, etc) and services (Facebook, Google, Amazon, etc) of those companies are used more often and more widely than for example services of some real estate company. During the study, it was realized that the number of tweets mentioning companies belonging to the Real Estate, Utilities and Energy sectors are

too small. For those sectors on an average four tweets per hour per company were collected compared to over 1000 tweets collected per company per hour for more popular sectors. However, those companies were still kept to study how much impact it has on information gain for social media related features.

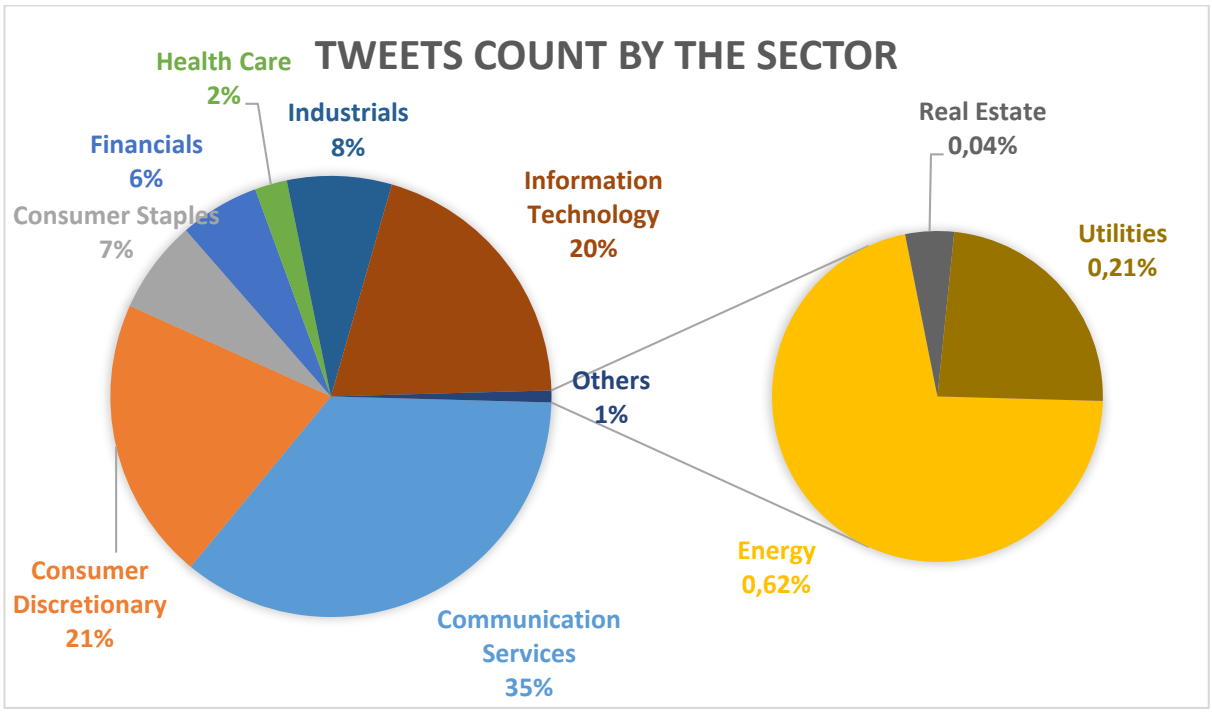


Figure 2: Tweets count by the sector

3.2 Tweets collection

Twitter is one of the largest social media platforms with 500 million posts per day on average. [28] For collecting the data through Twitter Streaming API, a developer account was created. Only 1% of the full feed (about 5 million tweets daily) is allowed to be collected with free developer accounts. For that, the stream can be pre-filtered, by defining which keywords should be included in the tweet or by which users they are posted. The maximum limit for keywords is 400. This means, on average four keywords per company for this research. In Twitter and other social media platforms, hashtags ('#' followed by keywords) are used to mark that the post is about a specific topic. In addition, user mentions are marked by the '@' symbol followed by the user name. In 2012, the social media platform introduced cashtags ('$' symbol followed by stock symbol) that are used to mark that the tweet is about a certain stock. [14] In this research, the filters contain cashtags (for example "$AAPL"), hashtags ("AAPL") and user mentions. Additionally, user mentions of the company official accounts ("@APPLE") and the company CEO's official ("@TIM_COOK") accounts are looked for. Also, a list of user names was created that contained company official accounts, CEO's official accounts, major news companies (The Wall Street Journal, NY Times, Bloomberg, Reuters, BBC, etc) and accounts

Forbes marked as the best financial Twitter accounts [8]. Tweets mentioning those users (over 300 user accounts) or posted by those users were collected in addition to collecting all the tweets including at least one of the 400 keywords mentioned earlier. After collection, additional filtering was needed to discard the tweets that were not actually connected to companies used in this study. One of the problems is that the same stock symbols are used for different companies in different markets. [10] Furthermore, some of the stock symbols have an alternative meaning (CAT - Cattepillar Inc, VISA - Visa Inc, etc). Also, abbreviations can be used differently. The stock symbol of Visa is V, but most of the tweets containing '#V', '@V' or '$V' are actually mentioning popular Korean singer.

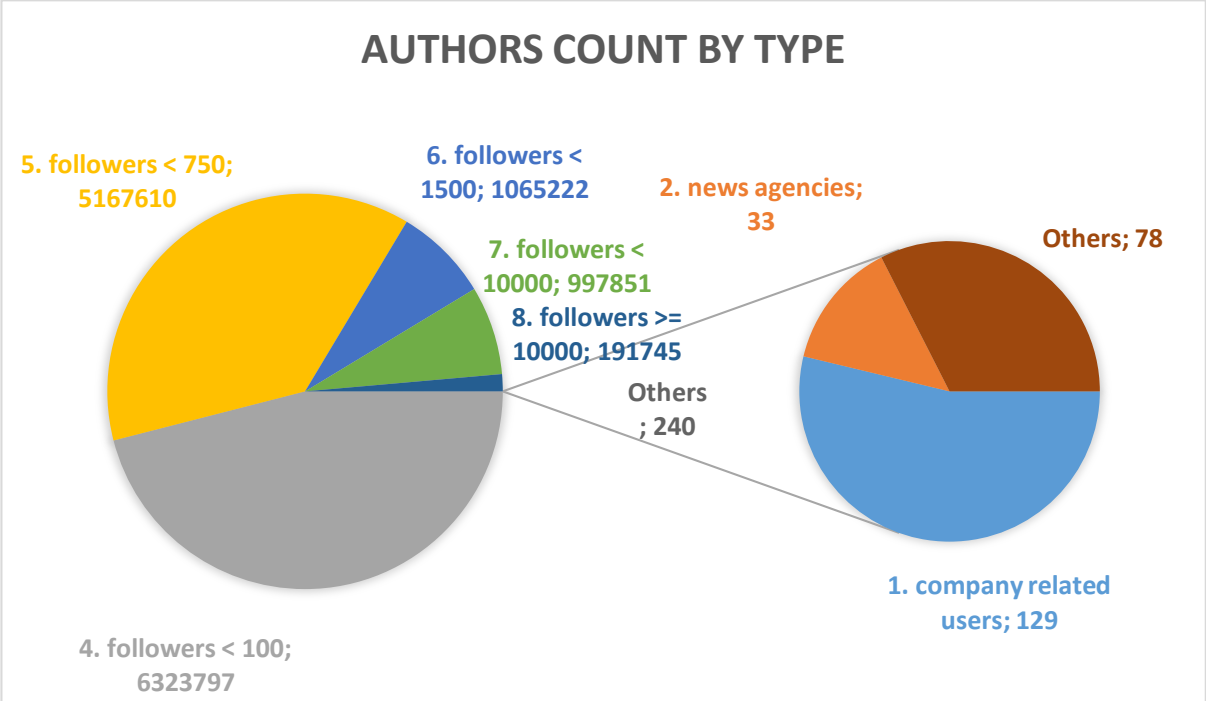


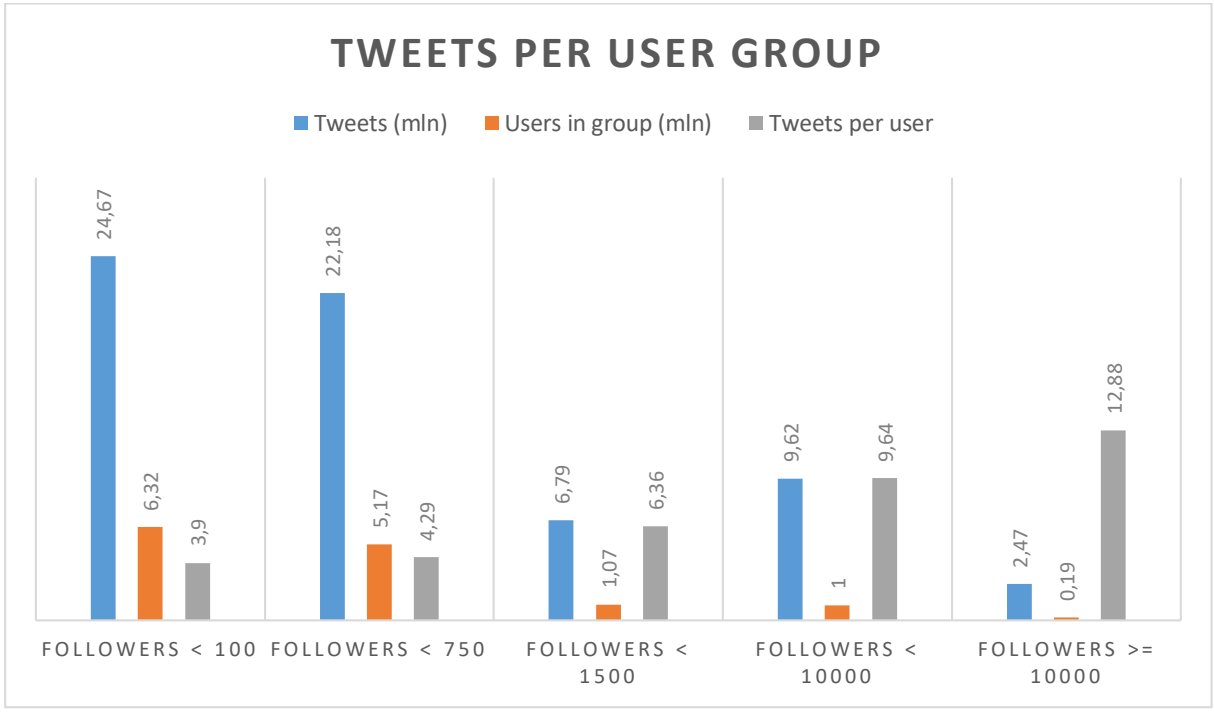
Figure 3: Author types

After discarding irrelevant tweets 47 million tweets remained, which were later grouped by author types. The authors were divided into 8 groups:

1. company-related official Twitter accounts
2. news agencies official Twitter accounts
3. users with less than 100 followers
4. users with less than 750 followers
5. users with less than 1500 followers
6. users with less than 10000 followers
7. users with more than 10000 followers

8. users mentioned by Forbes magazine to be the most influential finance-related Twitter accounts.

As can be seen in Figure 3 most of the tweet authors have relatively few followers. 84% of them belong to two groups containing users with less than 750 followers and only about 1% of them has more than 10 000 followers.



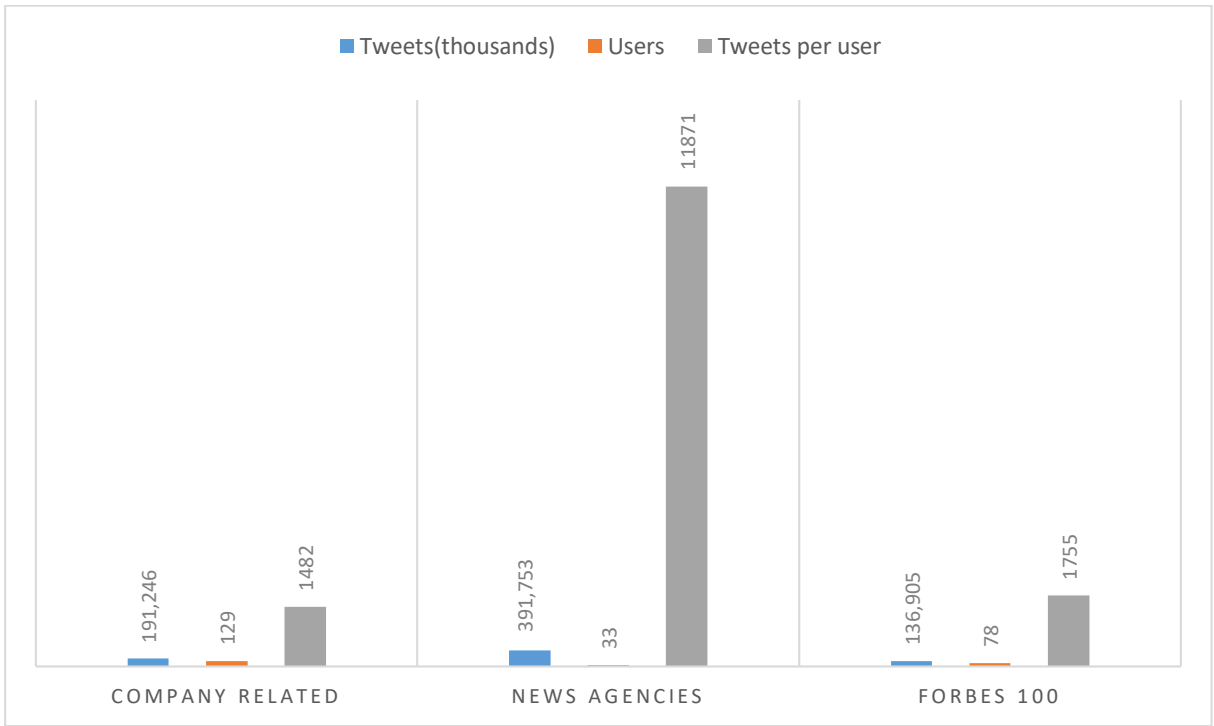


Figure 4: Author types

Figure 4 illustrates that even though the users with few followers tweeted relatively infrequently (on average 4 tweets during the study period) 70% of the tweets used in this study are posted by them. Company related accounts, news agencies and users mentioned by Forbes tweeted thousands of times during the study period. However since the amount of those users is small, less than a million tweets were initially collected.

3.2.1 Financial data

The financial data was collected from Alpha Vantage, from where price and volume data (timestamp, stock symbol, open price, minimum price, maximum price, close price, volume) were collected for the S&P 100 companies. Those stocks are traded on NYSE, which is open 5 days a week from 09:30 to 16:00 EEST. The price information is only available for opening hours and is updated once per minute. From this data, several new features are generated and two of them, maximum profit and maximum loss were used as a prediction target. Equation (1) was used for calculating maximum profit and equation (2) for calculating maximum loss. p_1 is the current price of the stock (the open price of this minute), p_2 is the maximum price and p_3 is the minimum price for the next five working days.

$$Y_1 = (\frac{p_2}{p_1} - 1) * 100 \tag{1}$$

$$Y_2 = (\frac{p_1}{p_3} - 1) * 100 \tag{2}$$

As can be seen in Figure 5 in late February there was a historic stock market crash after a record-long 11 years pull market. The crash was caused mainly by the COVID virus, but also at the same time the tensions between the US and China were escalating quickly and OPEC countries were arguing about oil production. [1]

Based on previous crises it was expected that recovery would take at least several years if not a decade. However, as can be seen from Figure 5 the market crash in March was soon followed by an extremely quick recovery. By the end of the study period in mid-May, the S&P100 index had already gained over 30% from the March lows and by early August record highs were reached. [4] In the meantime, volatility and trading activity was extremely high. On the one hand, the prices had dropped significantly which lured a large amount of new inexperienced investors to the market who were willing to make risky investments to make a quick profit. [24] On the other hand, there was much uncertainty about the scale and impact of the COVID virus. There was a strong fear that the stock market recovery had been too quick and another crash would follow soon. [2]

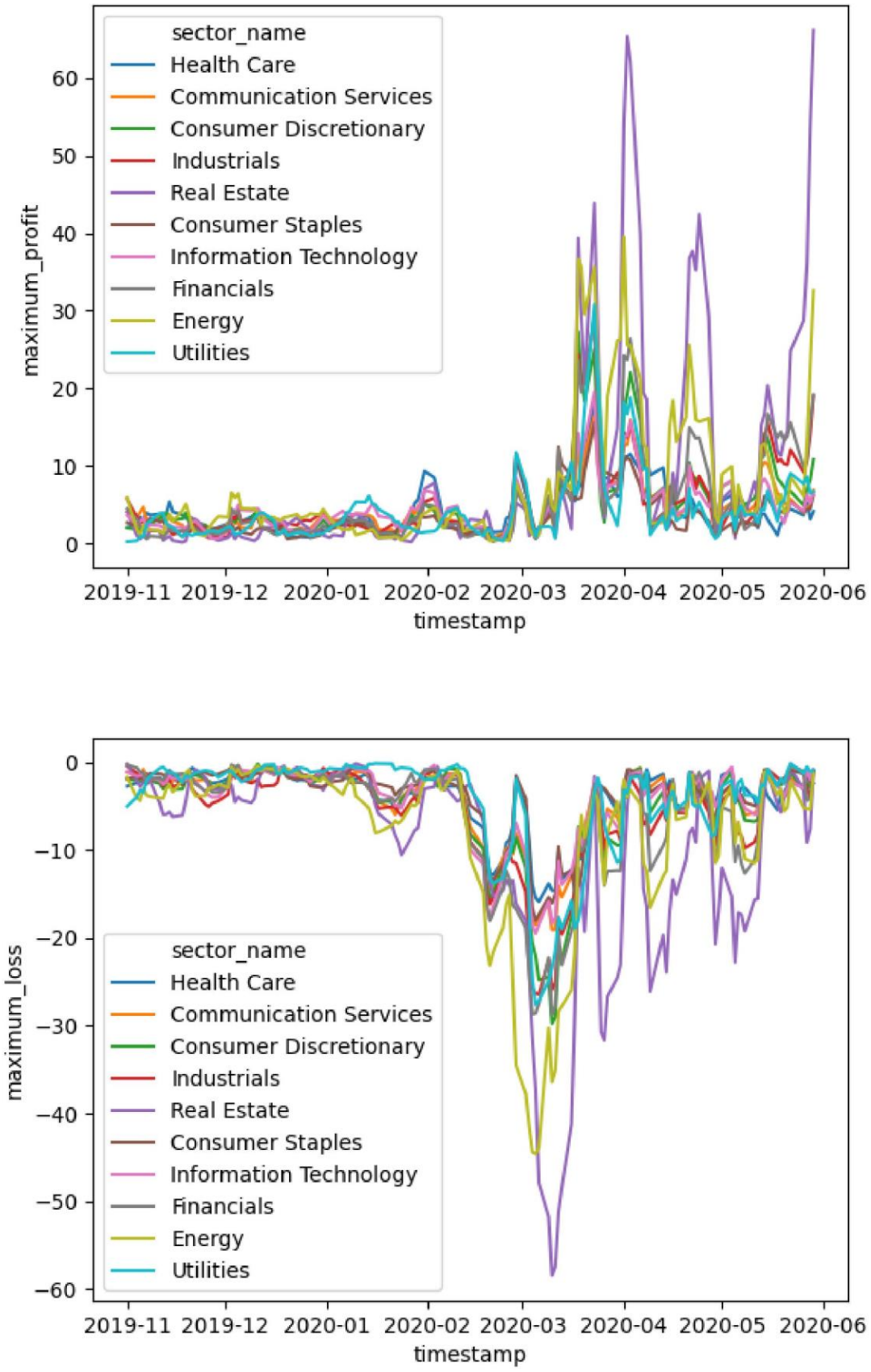


Figure 5: Maximum profit and maximum loss in 5 trading days (percentage)

4 Methodology

As described in the dataset section, the dataset can easily be categorized as BIG data. Thus, in this research, Apache Spark and MLlib were used to fully automate the process from data collection to making predictions. The main advantages of using distributed processing systems are that they are horizontally scalable, fault-tolerant and built for processing big data. All of those factors are crucial for creating trading programs. [9]

4.1 Data pipeline architecture

Since 2006, when Google introduced distributed file systems and distributed computation frameworks, there exist multiple tools and frameworks that can be used to collect and process large amounts of data parallelly. [11] Choosing which components to use for different processing steps is a complicated problem and depends on many factors: data throughput, latency, scalability, fault tolerance, etc. [9, 22] For some of the use cases it is sufficient if the data is collected and processed all at once or in relatively large batches, for example, once a week. On the other hand, there exist many scenarios where it would be important to collect and process the data with minimal latency (some milliseconds). There are three products (Apache Hadoop, Apache Spark, Apache Storm) that are capable to handle large amounts of data in a distributed process. They are all similar (horizontally scalable, fault-tolerant, distributed processing frameworks for analyzing big data), but all of them have different advantages. [22] In this research, predictions are made once a minute with a prediction horizon of 5 working days. Thus it is enough to process the data in small batches (sub-minute latency). For this research Apache Spark was picked, because it is about 100 times more powerful than Hadoop and it has the widest range of components (Spark Streaming, Spark SQL, MLlib for machine learning, GraphX for displaying the results). The latency of Apache Spark (several seconds) is enough for the task. Even though Apache Storm has even smaller latency (sub-second), it has higher requirements for individual servers and is more complex to develop. As illustrated in Figure 6 the process consists of 5 phases: data collection, data processing, machine learning, model evaluation, analysis.

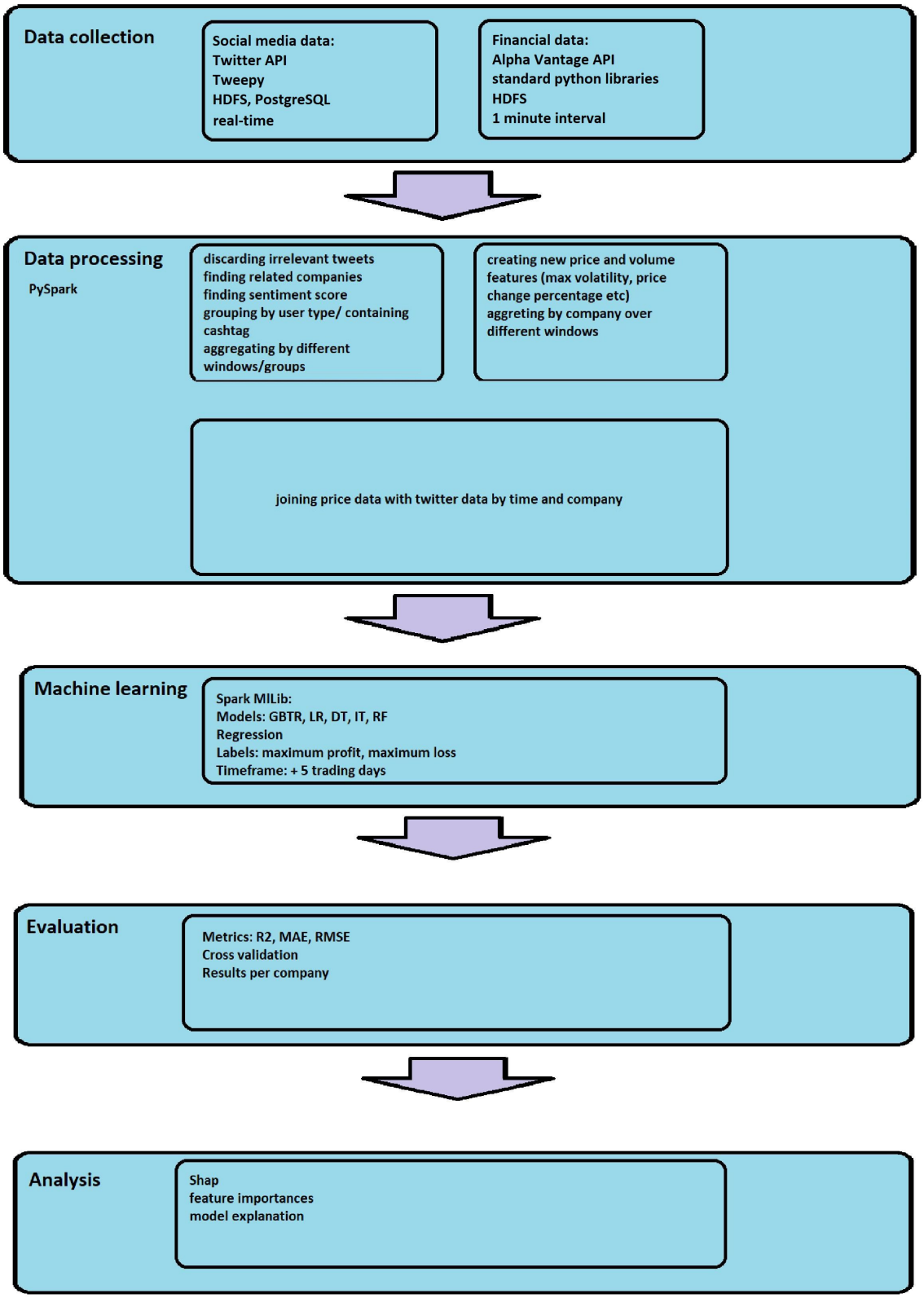


Figure 6: A process model

4.2 Pipeline phases

4.2.1 Data collection

The data is collected from two sources Twitter API and Alpha Vantage API. Twitter data was collected using the Tweepy library. Live Twitter feed was prefiltered using 400 keywords and following more than 300 users. Data was saved initially to Postgres database and later to Hadoop Distributed File System (HDFS). Financial data was collected through Alpha Vantage API, using standard python libraries. Price data was updated minutely during the opening hours of the market. The data was collected as soon as updated, but as opposed to Twitter data it could have been accessed later. Financial data was saved directly to HDFS. Please refer to Chapter 3 for more details.

4.2.2 Data cleaning

Even though the Twitter feed was pre-filtered before collecting the data, a large number of tweets were discarded later because they were not actually mentioning the S&P100 companies. For that reason, about 100 additional keywords were created and tweets containing any of them were discarded. This resulted in discarding 19 million tweets from the initial 66 million. For example, the keyword "football" was used to discard tweets mentioning football clubs (WBA - West Bromwich Albion) that have the same abbreviations as S&P 100 companies (WBA - Walgreens Boots Alliance Inc). Another example would be discarding tweets that are associated with popular bands ("V" - artist name of a popular singer, "BTS" - his band name) that use the same abbreviations as companies (V - Visa Inc). Another problem is cashtag collision, the same cashtags used in different markets for different companies. Evans et al. brought out the significance of the problem. They used 100 tickers for one month in their study and found out that over 64% of tweets contained colliding cashtags. [10]

4.2.3 Data processing

After cleaning the Twitter data, sentiment scores for the tweets were found using the TextBlob library. The range of sentiment scores was -1.0 to 1.0. The next step was to join the social media data with financial data. For each minute there was price information about every company. Sentiment scores from the previous one hour were aggregated by the company, by user group, by containing a cashtag or not, and by being a retweet or not.

For some companies, the sentiment score could always be relatively low. For example, if the company itself or the whole sector is not popular due to some reason (oil producers, because of the Greenpeace movement, etc). This doesn't mean that the stock price is always declining. [29] Thus to gain more information previous one-hour average sentiment scores were compared to sentiment scores of the same company for previous periods (4 hours, 1 day, 3 days, 7 days). Additionally, the tweet count of the previous hour for different groups and companies was

calculated and compared to previous periods. The current price and volume were compared to moving averages, minimums and maximums of different periods from 5 minutes to 7 days. Finally, the volatility of volume and price for different periods was calculated.

4.2.4 Employing predictive algorithms

Machine learning was done by using implementations of five different models (Linear regression, Decision tree regression, Random forest regression, Gradient-boosted tree regression, Isotonic regression) from Apache Spark MLLib. MLLib was picked as it is horizontally scalable, supports HDFS and it has relatively low requirements for individual servers. For every company, separate models were trained. Because the number of daily tweets per company varied significantly. Also, the average volatility of the companies stock price was very different.

4.2.5 Model evaluation

For evaluating the models the data was grouped by month. Then data from each month were randomly divided into five groups. Then in each iteration, one group was taken for evaluation and four were used for training the model. The following metrics for evaluation were used: MAE, RMSE, R2.

4.2.6 Model analysis

In the final step SHAP (SHapley Additive exPlanations) library was used to explain how the model made decisions. Also to find out which features were most important and to illustrate dependencies between different models. [17]

5 Evaluation

The main goal of this research was to find out whether including social media data based features would improve results for stock price prediction and if so then tweets from which type of users would give most information gain. How tweet authors were divided into groups is explained further in the Data section. For each company, five machine learning models were used:

1. Gradient Boosted Tree Regression (GBTR)
2. Linear Regression (LR)
3. Isotonic Regression (IT)
4. Decision Tree (DT)
5. Random Forest (RF)

Later the results from models containing social media based features were compared to models containing only financial data.

5.1 Metrics

As maximum profit prediction is a regression problem the following metrics were selected:

1. **Root Mean Squared Error (RMSE):** Root Mean Squared Error is a metric used for regression models that shows how much on average the prediction differs from the observed values (standard deviation of the residuals). Lower values mean better results, but the range of target values have to be also taken into account [13]
2. **Mean Absolute Error (MAE):** Mean Absolute Error indicates the average absolute difference between target and prediction. Because the errors are not squared it is less sensitive to outliers. Again the lower values are better and the target range has to be taken into account. [13]
3. **R-squared (R2):** R-squared is a statistical measure that shows the proportion of the variance for a dependent variable. The values are normally in the range of 0 to 1. The higher values indicate the model is performing better. [13]

In this study, Cross-validation was used, a model evaluation technique that is used for generalizing the results. It was implemented by firstly grouping data into six groups by month and then splitting it randomly again to five equal junctions. For each iteration, four parts from each month were used for training and one was used for evaluation.

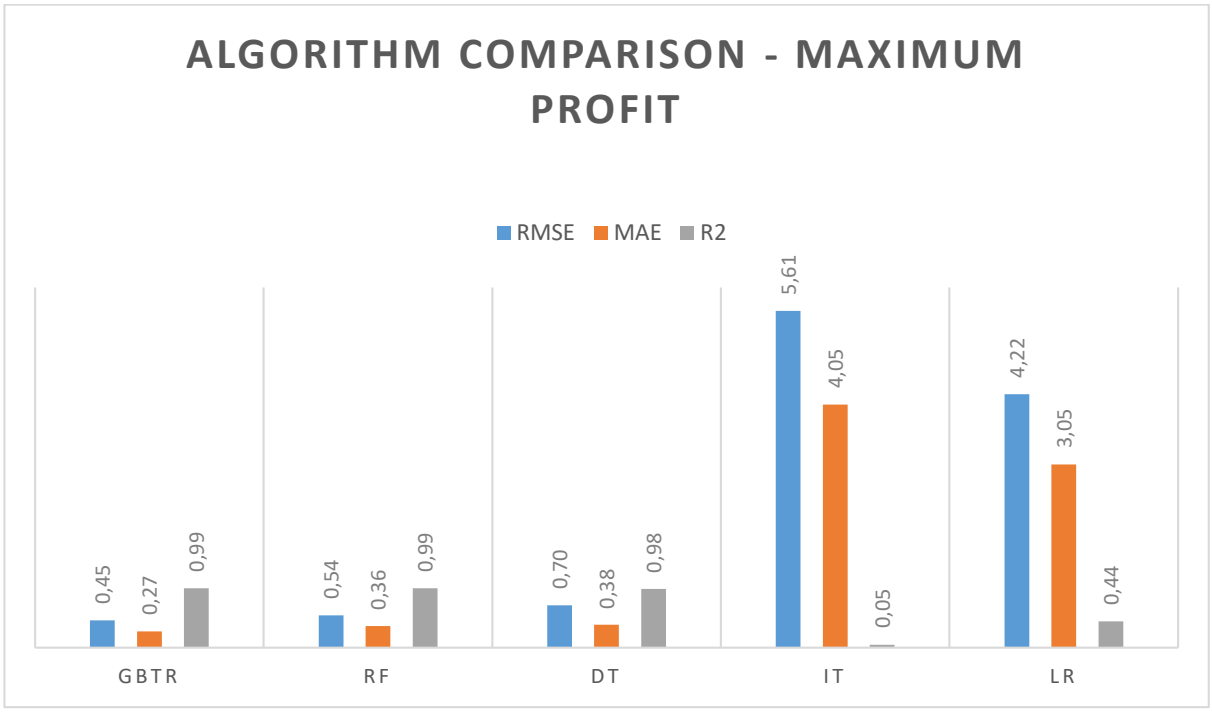
5.2 Results

Five Spark MlLib models were used on each company separately for training the model and predicting maximum profit and maximum loss. Figure 7 shows how different algorithms performed on an average using both social media data and financial data. The average five trading days maximum profit value was 4.91 for this study period. As can be seen from the results three models were performing well and the best results were obtained from the Gradient Boosted Tree Regression (GBTR) model. On the other hand, Isotonic Regression (IT) and Linear Regression (LR) models were not suitable for the task. The results for predicting maximum loss were similar, the best results were obtained by GBTR.

As best results were obtained from the GBTR model, it was picked to compare how the results differ by sector. Figure 8 illustrates the results for predicting maximum profit. The graph also compares the results for using different datasets:

1. using only financial data (PRICES)
2. using only social media data (TWEETS)
3. using both financial and social media data for training (ALL)

As can be seen from Figure 8 there were moderate differences between errors for different sectors, but they were in accordance with how maximum profit values per sector differed. For all sectors, the average errors per sector were best for using social media and financial data combined. It gave on average 10 percent gain from using only financial data. The errors for models using only Twitter data for input were about three times higher. However, there were 21 companies that didn't gain from adding social media based features. For most of them, the reason probably is that there were not enough tweets posted about the companies. On the other hand, they included a couple of very frequently mentioned companies also, for example, Netflix. Most of the tweets mentioning Netflix are about the different TV-series and movies that can be watched there and are not directly about the company. Even though in the long run it is important what people are thinking about the content it didn't seem to reflect in short-term price changes. In conclusion, when there is enough data about the company in social media, it is useful to add it as an extra data source for predicting maximum profit.



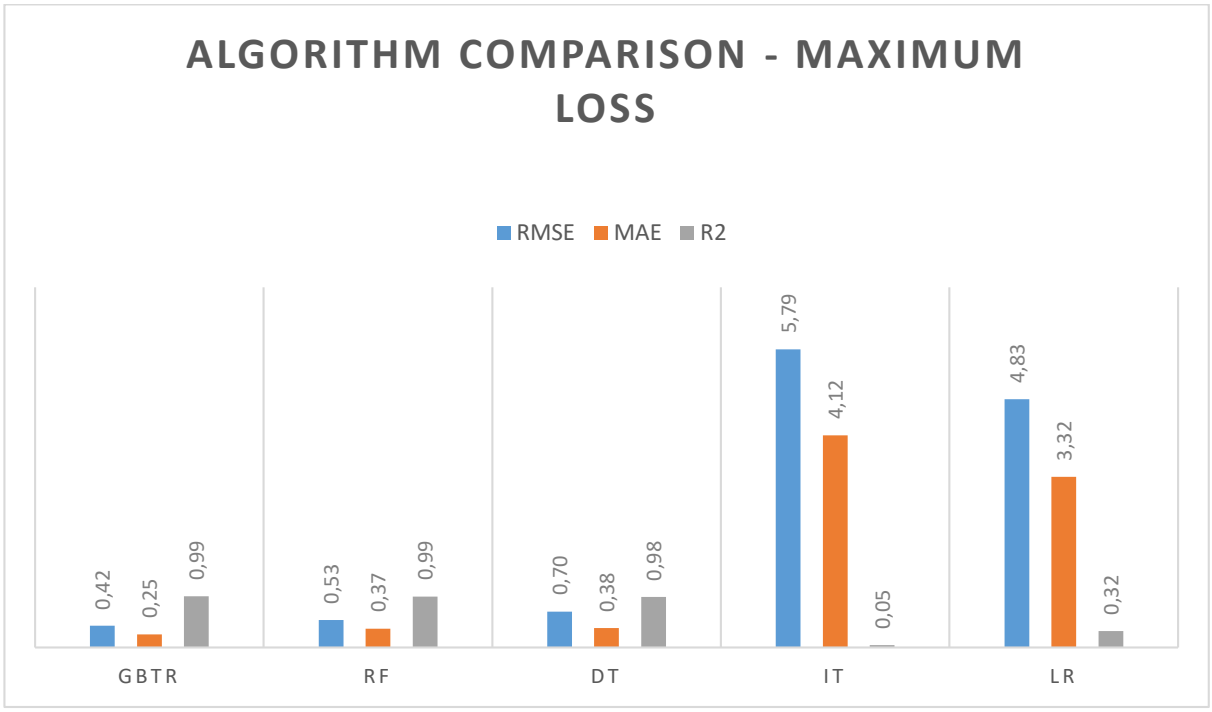


Figure 7: Results by the algorithm - maximum profit prediction

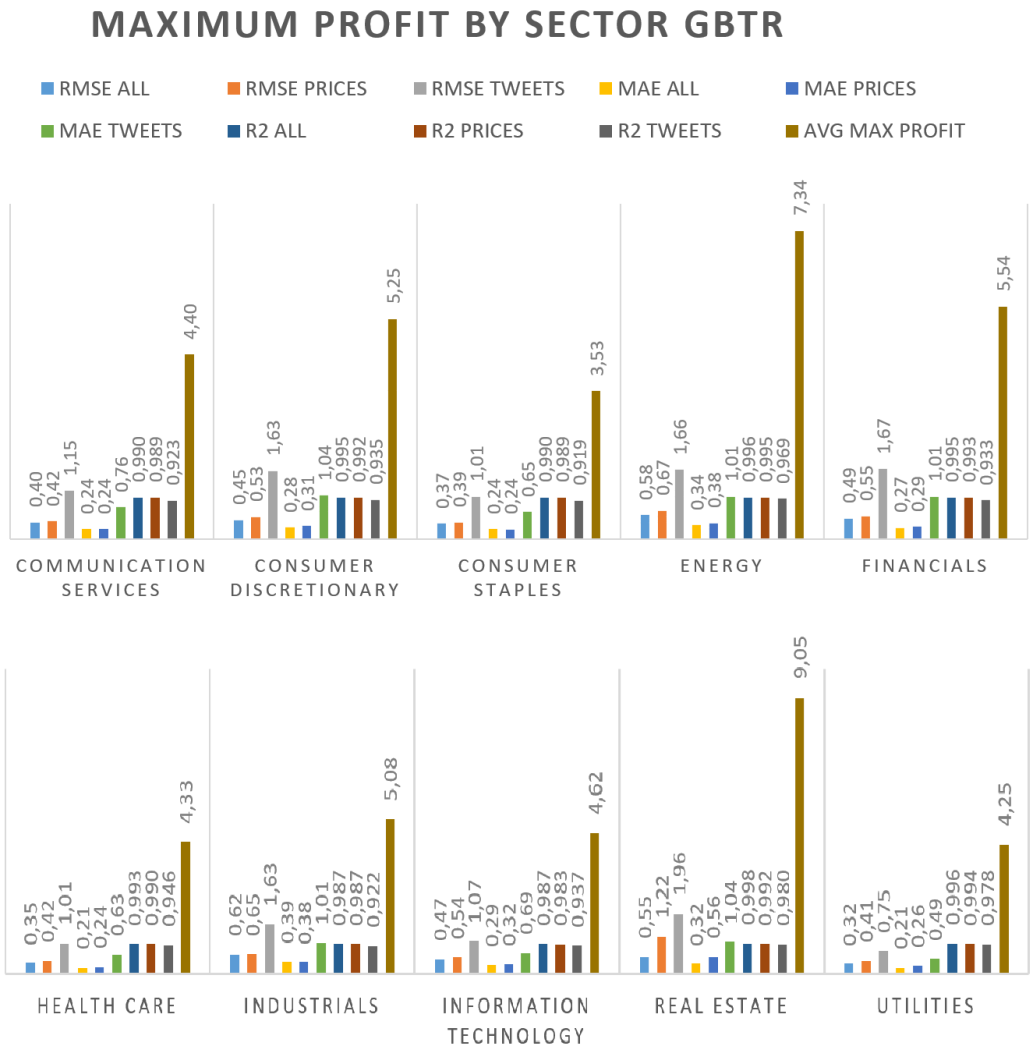


Figure 8: Maximum profit prediction results by sector (GBTR model)

Figure 9 shows the same metrics for predicting a maximum loss. The results were similar. The difference in sectors was small and was in accordance with how the maximum loss for companies differed. On average the models were able to obtain 11% better results from adding social media as a data source. Again some companies were not able to gain from adding social media data. The companies were mostly the same as seen from predicting maximum profit, but for predicting maximum loss there were more of them. In conclusion, it is beneficial to use social media data for predicting maximum loss as well, if enough data is collected about a company and the tweets mention mostly the company itself or its stock rather than some of its products.

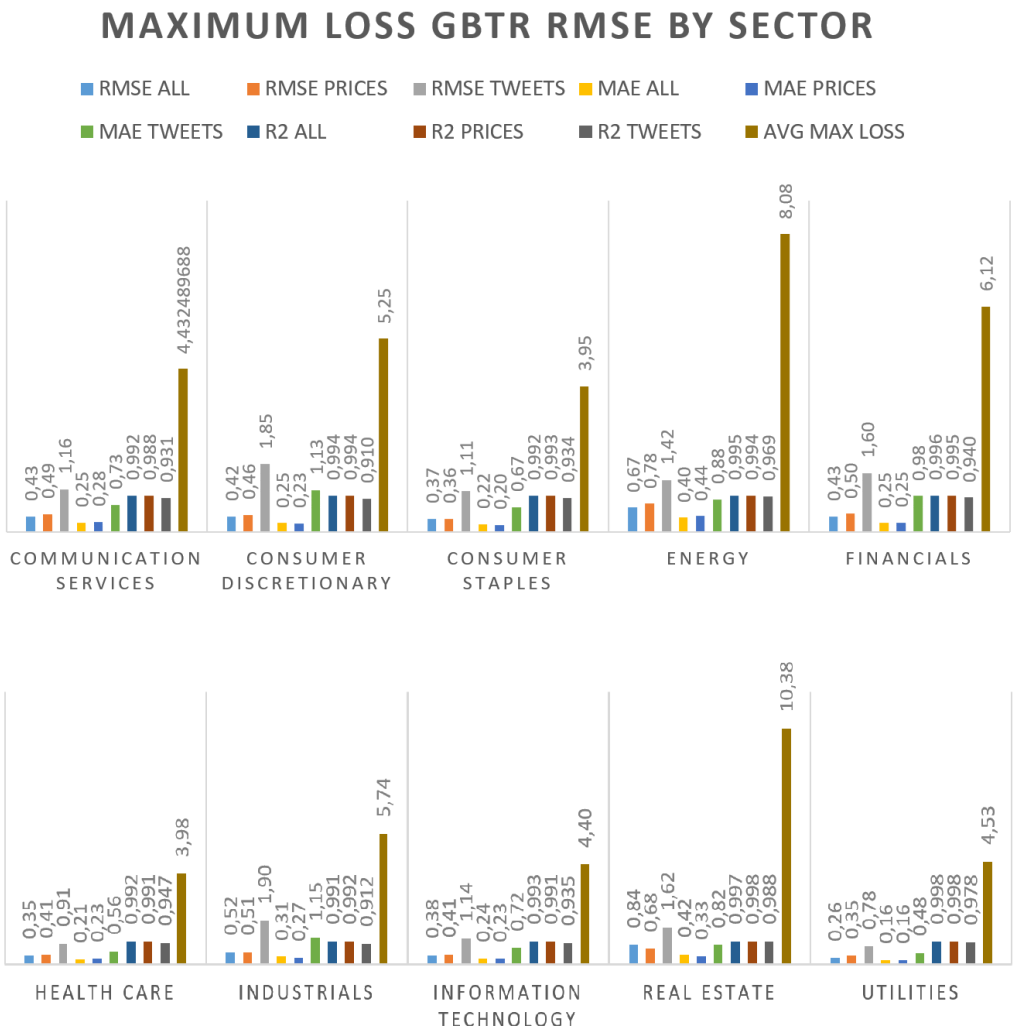


Figure 9: Results by the sector. Maximum loss.

5.2.1 Feature importance

In total there were 441 features used in the model. Apple Inc and GBTR model were picked as an example for Figure 10. The figure shows the top 30 features that had the highest mean SHAP value for predicting maximum profit. These are the features that had the most impact on model output. As can be seen from the graph the features extracted from financial data had more impact. The most important features (named "..._low_price_diff" and "..._high_price_diff") show how much the current price differs in percentage from the previous period maximum and minimum price. "..._high_low_diff" indicates the volatility of the previous periods. Interestingly the features showing how much the tweet counts for different groups changed was more important than the sentiment change. For tweet counts the user Group 2 containing the news agencies and Group 3 containing users that had less than 100 followers were most important. When a company is mentioned in news mostly something unexpected (positive or negative) have hap-

pened and this mainly causes quick price movement. After that, the price moves in the opposite direction until stabilizing relatively quickly on a new level. [29] During the crises majority of the news had negative sentiment. There were a lot of fears about the near future, reports that indicated losses or smaller profits, and people being laid off. [3] The worst-case scenario was quickly reflected in prices in March. Even though the news remained negative, soon the investors thought this had already been calculated into the price. [25] Thus even though news triggered larger price movements, not much information could be gained from their sentiment scores. On the other hand, it was important how the larger crowd reacted. How large tweeting activity change it caused for groups containing a large number of users and how their sentiment changed. As found out, Group 8, the users mentioned by Forbes magazine had the least impact on model output. It was probably because they had the least amount of tweets per group and many times fewer followers than for example news agencies. Separating tweets containing cashtag or not didn't have a large and similar effect for different companies. Cashtags should be used to mention that the tweet is about a certain company, but often they are used for other purposes and also many users prefer hashtags instead of cashtags, so the rule isn't followed strictly enough. Evans et al. brought out the fact that the same stock symbols refer to different companies in different markets amplifies the problem significantly. [10]

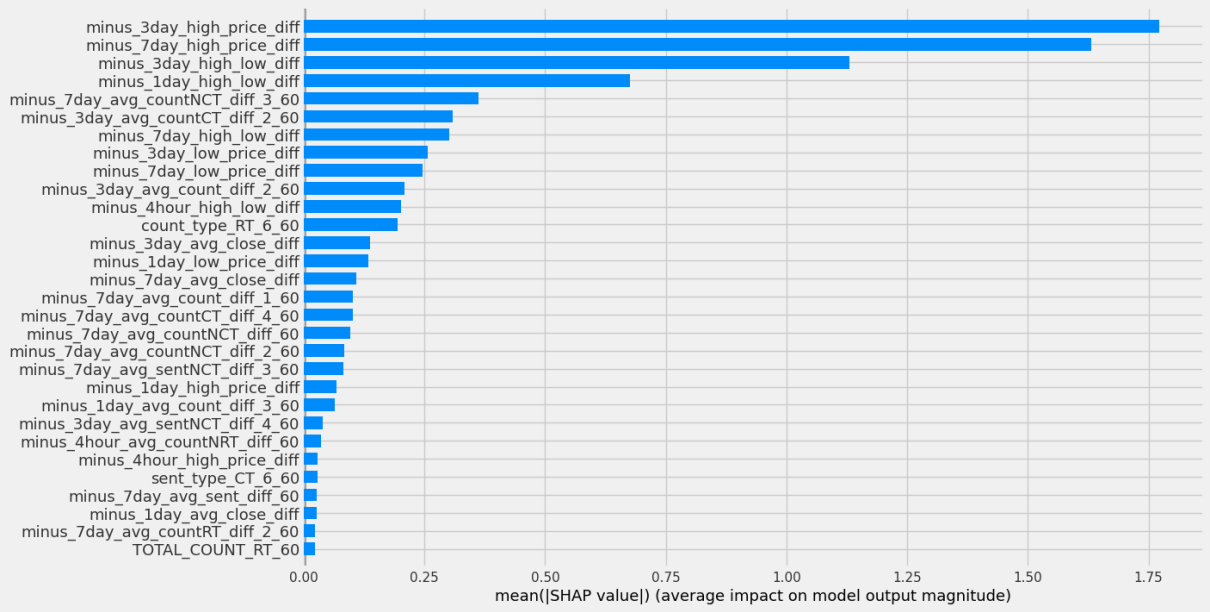


Figure 10: Shap - Feature impact for model output (company: Apple Inc, model: GBTR, target: maximum profit) model

Figure 11 shows the most important features for predicting the maximum loss. The feature importances are similar to the ones seen for predicting maximum profit, but the model is more dependent on a single feature. As seen from the graph the features indicating the magnitude of price volatility in the previous periods had the highest impact on model output. The tweet counts are again more important than sentiment scores. As for user groups, most value is gained

from Group 2 containing news agencies accounts and Group 3 containing users having less than 100 followers.

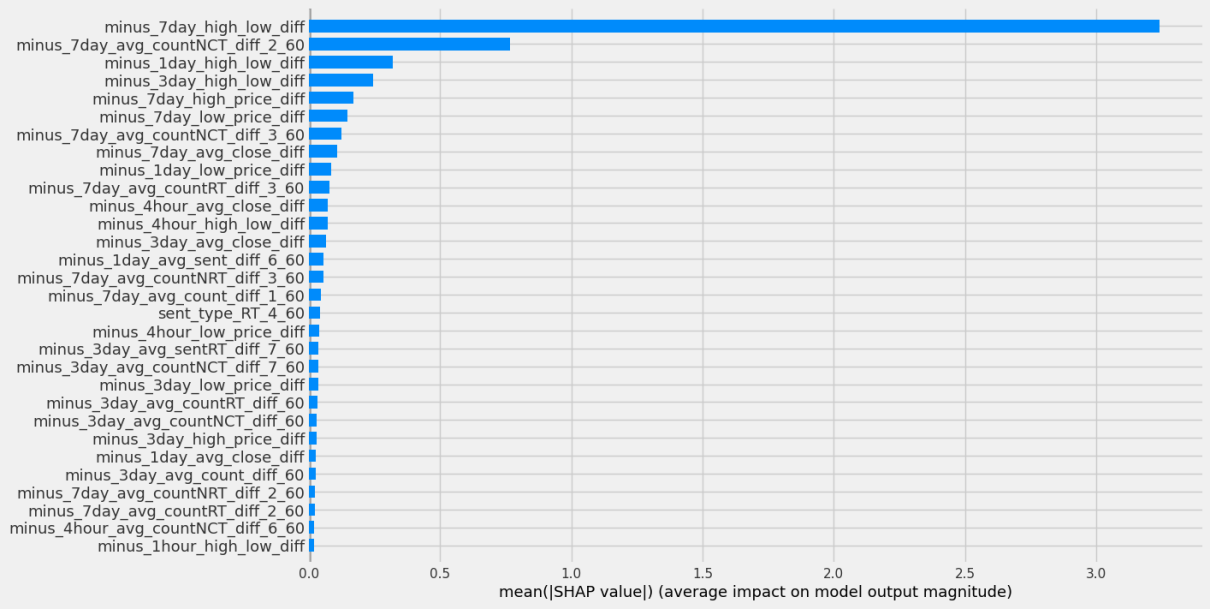


Figure 11: Shap - Feature impact for model output (company: IBM, model: GBTR, target: maximum loss) model

6 Conclusion

In this research social media and financial data were collected from a live stream for six months about S&P100 companies to predict maximum profit and maximum loss of the next five trading days. In total, 66 million tweets from 13,7 million different users were collected. After cleaning the data 47 million tweets remained and aggregated sentiment values were used for prediction. During the study period, the stock market had a historic crash due to the COVID-19 virus. It was amplified by tensions between the US and China and also between oil producers.[1] The crash was followed by a long period of extremely high volatility and millions of new inexperienced investors joining the market [24]. Despite the uncertainty about the impact of the virus the market recovered quickly.

For prediction, five different Spark MLLib models were used (Linear Regression - LR, Random Forest - RF, Decision Tree - DT, Isotonic Regression - IR, Gradient Boosted Tree Regression - GBTR). Although Linear regression and Isotonic regression were not suitable for the task, promising results were obtained from three models: GBTR, DT and RF.

The features extracted from financial data, showing the volatility of price and trading volume for different previous periods had the highest impact on model output. For social media, the most important features implicated how the tweets count changed for different user groups. Most information was gained from user groups containing news agencies accounts and from the groups having less than 750 followers. The news is usually triggering the larger price changes. On the other hand, the magnitude of the change depends on how the larger crowd interprets the news and how much reaction it causes. [29] 70% of the tweets collected in this study were posted by users having less than 750 followers. For sentiment-based features, the ones showing overall sentiment change were more important than the sentiment of any specific group.

Predicting price changes is a complicated problem and the price depends on a high amount of factors. [27] It was found out that adding social media as an extra data source is beneficial for both predicting maximum profit and maximum loss. The RMSE values were about 10% smaller for using combined data compared to using only financial data for input. However, the results didn't improve for all companies. Those companies were mostly the ones that were not tweeted frequently enough on social media. For others, the problem was that most of the tweets were mentioning some of the products rather than the company itself or its stock. What people think about a product might be important for a longer prediction horizon, but it didn't reflect in prices in the five day period, that was studied in this research.

During this research the market situation was extraordinary. There were a large stock crash and a rapid recovery. The weekly maximum profit and maximum loss were many times higher than they would be normally. In this situation, the emotions of the investors play a larger role than they would in a steady market and it made the market inefficient. [18] In those conditions, the models were able to gain valuable information from social media. In further researches, it would be interesting to find out if the same applies to different market phases as well. Addi-

tionally models trained on data from different social media platforms could be compared.

References

[1] Z. Allam. Oil, health equipment, and trade: Revisiting political economy and international relations during the covid-19 pandemic. Surveying the Covid-19 Pandemic and its Implications, page 119, 2020.

[2] D. Altig, S. Baker, J. M. Barrero, N. Bloom, P. Bunn, S. Chen, S. J. Davis, J. Leather, B. Meyer, E. Mihaylov, et al. Economic uncertainty before and during the covid-19 pandemic. Journal of Public Economics, 191:104274, 2020.

[3] F. Aslam, T. M. Awan, J. H. Syed, A. Kashif, and M. Parveen. Sentiments and emotions evoked by news headlines of coronavirus disease (covid-19) outbreak. Humanities and Social Sciences Communications, 7(1):1–9, 2020.

[4] G. Banerji. Why Did Stock Markets Rebound From Covid in Record Time? Here Are Five Reasons. https://www.wsj.com/articles/why-did-stock-markets-rebound-from-covid-in-record-time-here-are-five-reasons-11600182704, 2020. [Online; accessed 11-October-2020].

[5] C. Banton. An Introduction to U.S. Stock Market Indexes. https://www.investopedia.com/insights/introduction-to-stock-market-indices/, 2020. [Online; accessed 17-October-2020].

[6] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. Journal of computational science, 2(1):1–8, 2011.

[7] J. Breckenfelder. Competition among high-frequency traders, and market quality . https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2290~b5fec3a181.en.pdf, 2019. [Online; accessed 03-March-2020].

[8] M. Checkley, D. A. Higón, and H. Alles. The hasty wisdom of the mob: How market sentiment predicts stock market behavior. Expert Systems with applications, 77:256–263, 2017.

[9] S. Das, R. K. Behera, S. K. Rath, et al. Real-time sentiment analysis of twitter streaming data for stock prediction. Procedia computer science, 132:956–964, 2018.

[10] L. Evans, M. Owda, K. Crockett, and A. F. Vilas. A methodology for the resolution of cashtag collisions on twitter–a natural language processing & data fusion approach. Expert Systems with Applications, 127:353–369, 2019.

[11] D. Harris. The history of Hadoop: From 4 nodes to the future of data. https://gigaom.com/2013/03/04/the-history-of-hadoop-from-

4-nodes-to-the-future-of-data/, 2013. [Online; accessed 14-September-2020].

[12] B. Huang, Y. Huan, L. D. Xu, L. Zheng, and Z. Zou. Automated trading systems statistical and machine learning methods and hardware implementation: a survey. Enterprise Information Systems, 13(1):132–144, 2019.

[13] A. Kassambara. Regression Model Accuracy Metrics: R-square, AIC, BIC, Cp and more. http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/, 2018. [Online; accessed 21-September-2020].

[14] E. Kim. Twitter unveils 'cashtags' to track stock symbols. https://money.cnn.com/2012/07/31/technology/twitter-cashtag/index.htm, 2012. [Online; accessed 11-October-2019].

[15] M. Lachanski and S. Pav. Shy of the character limit:" twitter mood predicts the stock market" revisited. Econ Journal Watch, 14(3):302, 2017.

[16] B. Li, K. C. Chan, C. Ou, and S. Ruifeng. Discovering public sentiment in social media for predicting stock movement of publicly listed companies. Information Systems, 69: 81–92, 2017.

[17] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[18] Š. Lyócsa and P. Molnár. Stock market oscillations during the corona crash: The role of fear and uncertainty. Finance Research Letters, 36:101707, 2020.

[19] M. Makrehchi, S. Shah, and W. Liao. Stock prediction using event-based sentiment analysis. In 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), volume 1, pages 337–342. IEEE, 2013.

[20] I. Mäe. Milline Eestis tegutsev pank sobib sinu investeerimisstrateegiaga? https://www.aripaev.ee/edetabel/2016/06/03/milline-eestis-tegutsev-pank-sobib-sinu-investeerimisstrateegiaga, 2016. [Online; accessed 20-May-2020].

[21] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo. Text mining for market prediction: A systematic review. Expert Systems with Applications, 41(16):7653–7670, 2014.

[22] H. Nazeer, W. Iqbal, F. Bokhari, F. Bukhari, and S. U. R. Baig. Real-time text analytics pipeline using open-source big data tools. ArXiv, abs/1712.04344, 2017.

[23] N. Oliveira, P. Cortez, and N. Areal. The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with Applications, 73:125–144, 2017.

[24] S. Rega. How Robinhood and Covid opened the floodgates for 13 million amateur stock traders. https://www.cnbc.com/2020/10/07/how-robinhood-and-covid-introduced-millions-to-the-stock-market.html, 2020. [Online; accessed 10-October-2020].

[25] S. Salbrechter. Stock Price Prediction Based on a Sentiment Analysis of Financial News. PhD thesis, Wien, 2020.

[26] A. Shah. The 100 Best Finance Twitter Accounts You Should Be Following. https://www.forbes.com/sites/alapshah/2017/11/16/the-100-best-twitter-accounts-for-finance/#5e226c947ea0", 2017. [Online; accessed 1-June-2019].

[27] D. Shah, H. Isah, and F. Zulkernine. Stock market analysis: A review and taxonomy of prediction techniques. International Journal of Financial Studies, 7(2):26, 2019.

[28] I. L. Stats. Twitter Usage Statistics. https://www.internetlivestats.com/twitter-statistics/, 2020. [Online; accessed 10-October-2020].

[29] H. K. Sul, A. R. Dennis, and L. Yuan. Trading on twitter: Using social media sentiment to predict stock returns. Decision Sciences, 48(3):454–488, 2017.

[30] Wikipedia. SP 100. https://en.wikipedia.org/w/index.php?title=S%26P_100&oldid=867611884, 2018. [Online; accessed 12-November-2018].

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Ardi Aasmaa**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Predicting stock price based on media monitoring,

   supervised by Rajesh Sharma.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Ardi Aasmaa
12/11/2020