FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

# Sentiment analysis of public information to predict stock market movements

DIPLOMARBEIT

zur Erlangung des akademischen Grades

**Diplom-Ingenieur**

im Rahmen des Studiums

**Wirtschaftsinformatik**

eingereicht von

**Bc. Jiří Brom**
Matrikelnummer 1425300

an der Fakultät für Informatik
der Technischen Universität Wien
Betreuung: Ao.Univ.Prof. Dr. Dieter Merkl

Wien, 12. Februar 2018

| Jiří Brom | Dieter Merkl |
|---|---|

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

# Sentiment analysis of public information to predict stock market movements

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

**Diplom-Ingenieur**

in

**Business Informatics**

by

**Bc. Jiří Brom**
Registration Number 1425300

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Dr. Dieter Merkl

Vienna, 12th February, 2018

| Jiří Brom | Dieter Merkl |
|---|---|

# Erklärung zur Verfassung der Arbeit

Bc. Jiří Brom
Nova Ves 5, 375 01 Olesnik, Czech Republic

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 12. Februar 2018

Jiří Brom

# Acknowledgements

I would like to express a special thanks to my supervisor Dieter Merkl for helping me to come up with an interesting research topic and overseeing my progress. Secondly I would also like to thank my family and friends for supporting me during my whole studies abroad and helped me in finalizing this project.

# Kurzfassung

Der Schwerpunkt dieser Arbeit liegt auf der Vorhersage zukünftiger Bewegungen von Aktienkursen unter der Verwendung von Sentimentanalysen aus öffentlich zugänglichen Daten. Das Ziel ist es, verschiedene Datenquellen und Verarbeitungstechniken zu vergleichen, deren Vorteile und Nachteile zu identifizieren und daraus die beste Kombination zu finden, welche die höchste Vorhersagegenauigkeit bietet. Aus verschiedenen populären sozialen Netzwerken und Zeitschriften, welche als Repräsentant der öffentlichen Meinung herangezogen werden können, wurden die folgenden drei für die Analyse ausgewählt: Twitter, Stocktwits und die Suche nach Nachrichtenartikeln von Bing. Twitter stellt die allgemeinste Quelle der öffentlichen Meinung dar – von hier werden Daten, welche von einer Vielzahl von Usern ohne spezifische Beziehung zum Aktienmarkt bezogen. Stocktwits, ein anlageorientiertes soziales Netzwerk, liefert Beiträge von Internetnutzern, welche sich für Wirtschaft und Aktienhandel interessieren. Nachrichtenartikel dienen als Informationsquelle, welche die Meinungen der beiden genannten Benutzergruppen beeinflusst. Die Auswahl der genannten Quellen basierte auf mehreren technischen Faktoren, einschließlich der freien Verfügbarkeit der zugrundeliegenden API sowie der Menge der daraus täglich abrufbaren Daten. Als Quelle für die Prognose der Aktienkursbewegungen wurden die folgenden neun Aktiengesellschaften und Indizes ausgewählt: Coca-Cola, McDonald's, Microsoft, Netflix, Nike, Tesla, Dow Jones Industrial Average, NASDAQ, Standard & Poor's 500. Diese Auswahl wurde durch die Markenbekanntheit beeinflusst.

Verarbeitungstechniken, die auf die zugrundeliegenden Daten angewendet wurden, umfassen typische Textmanipulationsmethoden, wie z. B. Stemming, Stopwortentfernung, POS-Tagging oder Bigramm-Kollokationen. Der Sentimentanalyseprozess basiert auf einem allgemeinen Repräsentationsmodell, dem sogenannten „Bag of Words“ (Vektorraummodell). Für die Klassifizierungsaufgaben wurde eine Kombination gängiger Maschinenlernalgorithmen verwendet: Naive Bayes, Logistische Regression und Support Vector Machines. Die Ergebnisse wurden mittels Granger-Kausalität und binärer Klassifikation analysiert. Der Granger-Kausalitäts-Test untersucht die Korrelation zwischen täglichen Stimmungs- und Kursschwankungen der Daten-Serien. Der binäre Klassifikations-Test versucht, die zukünftige Auf- oder Abwärtspreisbewegung basierend auf der Sentimentanalyse der letzten drei Tage vorherzusagen. Als Hauptergebnis der Analyse wurde festgestellt, dass das soziale Netzwerk Stocktwits das größte Vorhersagepotenzial aufweist und dass eine starke Korrelation zwischen dem KGV eines Unternehmens und der Vorhersagbarkeit dessen Aktienkurses besteht.

# Abstract

The focus of this research is the phenomenon of predicting future movements of stock market prices using sentiment changes obtained from publicly available sources. The aim is to compare a subset of different input data sources and processing techniques in order to identify their benefits and shortcomings, and find their best combination that would provide the highest prediction accuracy. Out of many popular networks and journals which could be used as a source of public sentiment the following three subjects for the analysis were chosen: Twitter, Stocktwits and Bing search news articles. Twitter social network represents the most general source of public sentiment - data obtained from a big amount of Internet users with no specific relation to investment. Stocktwits, an investment oriented social network, provides input from Internet users interested in economy and stock market. News articles serve as a source of information which influences the opinions of both groups of users. The selection of the mentioned sources was based on multiple technical factors including free accessibility of API or the amount of daily available data. As the source of stock price movements which we try to predict the following nine stock market companies and indexes were chosen : Coca-Cola, Mcdonald's, Microsoft, Netflix, Nike, Tesla, Dow Jones Industrial Average, NASDAQ, Standard & Poor's 500. Their selection was influenced by the general brand recognition.

Processing techniques applied to the analyzed data include typical text manipulation methods such as stemming, stop words removal, POS tagging or bigrams collocations. The sentiment analysis process uses a very common data representation model called Bag of Words (Vector space model). For the classification tasks we used a combination of well known machine learning algorithms: Naive Bayes, Logistic Regression and Support Vector Machines. The results were analysed using Granger causality and binary classification. The Granger causality test identifies the correlation between two series of daily sentiment and stock changes. The binary classification test tries to predict the future up or down price movement based on the sentiment measured for the last three days. As the main result of the analysis it was discovered that the Stocktwits social network has the biggest prediction potential and that there exists a strong correlation between a company's P/E ratio and the predictability of its stock price movements.

# Contents

CHAPTER 1

# Introduction

## 1.1 Motivation

The usage of sentiment analysis to predict stock market movements has been a common practice for several years in many investment-oriented companies. Unfortunately, the vast majority of commercial products lack transparency about the data sources and techniques used for their analysis and prediction. Academic projects are obviously transparent enough but they mostly focus on one specific type of input data like Twitter posts or news articles from finance journals. There is no research focusing on the comparison of various data sources and their benefits or shortcomings for sentiment based stock prediction. It is unclear whether one data source is more suitable for stock prediction than other, and what are the best techniques to manipulate the input data. Finding the answers to these questions can enhance the known techniques of stock movement prediction and it can be beneficial for future studies in this field.

## 1.2 Problem Statement

Since the very beginning of the stock market existence there have been many people around the world asking the question whether it is possible to predict the future price of stocks. Two basic theories dealing with this problem provide a rather skeptical answer. The Efficient Market Hypothesis states that it is impossible to “beat the market” because the existing share prices always reflect all the relevant information and stocks are therefore always traded at their fair value [MF70]. Similarly, the Random Walk Theory says that stock price changes have the same distribution and are independent of each other, so the future movement cannot be predicted based on the past trend [Fam95].

Nevertheless, many economists actually managed to "outperform the market" and earned a lot of money by investing in the right assets [VW04]. Both hypotheses are nowadays

facing strong criticism and various papers opposing their validity appeared, such as [Shi03] and [Sto97]. Although there are many different approaches to stock price movement prediction, most of them have a common aspect and that is the influence of the public information [MM94].

The influential public information can be represented in numerous ways. A businessman's decision whether to invest money can be based on an article in an important economical newspaper, a message in a stock board discussion or simply on an overall feeling of the public attitude received from recent Twitter posts [BMZ11]. Although it is clearly impossible to collect all the existing information that influence business decisions, at least a subset of relevant data should be sufficient to give us a hint of the possible shift direction.

Well-chosen and precisely collected data can serve as an input for sentiment analysis algorithms which are able to provide an exact outcome describing the current global sentiment for a given topic. The result can supply important advice for stock market traders and enhance the complicated and time-consuming process of manual information retrieval.

Currently, there are many different sources of data as well as many possible methods of their processing. In order to get the most accurate prediction, we need to find the optimal combination of data sources and analysis techniques which will provide the highest precision outcome. Therefore, a detailed research of various sentiment analysis methods and a comparison of their performances is needed.

## 1.3 Aim of the work

The expected outcome of this work is an indication of the mutual influence between the stock price movements and the public sentiment. The result will be provided in a form of a comparative study focusing on various information sources and processing algorithms used to predict the movements of stock market prices. The outcome will tell us why some of the researched input data sources and processing algorithms are more suitable for the given task than others and it will highlight the aspects and properties influencing the precision.

The final conclusions will be based on real outcomes obtained from a program developed in the Python programming language. The program can download data published in various news journals and social networks during a given day. The collected information serves as an input for multiple sentiment analysis algorithms which result in specific ratios of positive versus negative sentiment. In case of a prevailing positive mood we expect the stock price to increase, in case of a negative result we expect the opposite. The download and consequent analysis of the data was repeated daily for 10 months in order to obtain a representative time series of public mood changes. In the Chapter 6 (Prediction) the results will be compared with real values of the observed stock variables and the accuracy of the prediction will be calculated.

The thesis will answer the following research questions:

- What is the best accuracy of the stock price movement predictions obtained for all the observed data and stock variables?
- For which combination of input data sources and preprocessing techniques do we obtain the best results and why?

CHAPTER 2

# State of the art

## 2.1 Efficient market hypothesis

As the cornerstone of research dealing with stock market prediction we could consider the Efficient market hypothesis which was formulated for the first time by Burton Malkiel and Eugene Fama in their famous paper *Efficient capital markets* [MF70]. According to the theory all stock markets fully reflect the public information made available to the market participants at any given time. Since no investor has the privilege to get new information earlier than anyone else, it is impossible for him or her to purchase undervalued stock or sell stocks for inflated prices and so the share has to be traded at its fair value. The only way an investor can possibly obtain higher returns is by purchasing riskier investments and being lucky.

Obviously, there have been many arguments against the validity of EMH in the real world and these are summarized in the paper *The Efficient Market Hypothesis and Its Critics* by Burton Malkiel [Mal03]. A common claim used by the EMH opponents is the existence of investors who have consistently beaten the market over long periods of time, as for example Warren Buffet. Another points of contention to the case include instances when one investor looks for undervalued market opportunities while another investor evaluates a stock on the basis of its growth potential. These two investors will already have arrived at a different assessment of the stock's fair market value. Therefore since investors value stocks differently, it is impossible to ascertain what a stock should be worth under an efficient market.

The supporters of EMH, however, remain calm against the opposition. Their defence is that the hypothesis does not dismiss the possibility of market anomalies that result in generating superior profits. In fact, market efficiency does not require prices to be equal to fair value all the time. Prices may be over- or undervalued only in random occurrences, so they eventually revert back to their mean values. As such, because the deviations from

a stock's fair price are in themselves random, investment strategies that result in beating the market cannot be consistent phenomena. Furthermore, the hypothesis argues that an investor who outperforms the market does so not because of their skill but because of luck. EMH followers say this is due to the laws of probability: at any given time in a market with a large number of investors, some will outperform while others will underperform [Hea17].

The EMH controversion is well illustrated by the contrasting opinions of Nobel Prize laureates in 2013. Eugene Fama was awarded Nobel Prize in Economics Sciences jointly with Robert Shiller and Lard Peter Hansen. While Fama's approach based on the Efficient market theory is considered as a "neoclassical approach", Shiller is known as a supporter of so called "behavioral school" which is based on the idea that some financial events can only be explained by recognizing that psychological imperfections often cause people to behave irrationally [She13]. His famous work includes studies such as *From efficient markets theory to behavioral finance* and other research papers focusing on so called "bubbles" in financial markets [Shi03].

## 2.2 Recent academic research

One of the most famous papers concentrating on stock market prediction based on sentiment values is a study called *Twitter mood predicts the stock market* from Johan Bollen et al [BMZ11]. In the research Bollen and his two colleagues used OpinionFinder and Google-Profiles of Mood States as lexica to construct seven different sentiment analyzers, each related to a different type of mood as for example "happy", "calm", "alert", etc. The sentiment values were extracted from non-topic related tweets collected during the year 2008 and used to predict the daily up and down changes in closing values of Dow Jones Industrial Average index. The results show incredible accuracy of 87% using a predictor trained on calm mood state sentiment.

The paper became very popular and often cited. In the time of writing this thesis the paper has had 2514 citations on Google Scholar and it has been used as a benchmark for many other studies. However, in 2012 an anonymous blogger drew attention to some misleading approaches used in the study [Fol14]. The main inaccuracy seems to be the fact that the test was conducted on a very small time window, namely only 15 business days (December 1 to December 19, 2008). Another problem is the total number of 49 tested hypotheses which decreases the importance of just a few significant results.

In 2010 Arshul Mittal and Arpit Goel from the Stanford University tried to reprocess Bollen's findings and reached 75% accuracy, which is a much lower number than the original outcome [MG12]. A similar study conducted by Chen and Lazer also relates to Bollen's work and concludes that Twitter data predates the market by about 3 days [CL13]. Given the above mentioned facts when considering the Bollen's study results, it is probably better not to take the specific outcome numbers too seriously, especially not as benchmarks for other studies. On the other hand, despite the controversy, the study

clearly provides some useful comparison outcomes, as for example the fact that the stock market reacts to negative sentiment much more significantly than to the positive one.

A slightly different approach to predicting the stock price movements was chosen by Robert Schumacher and Hsinchun Chen from University of Arizona. Instead of using sentiment analysis, they implemented an artificial intelligence system called AZFinText which learns directly by comparing input text from various news articles with stock price movements within 20 minutes after the article release. The best accuracy reaches approximately 58% in the price direction prediction when concentrating on noun phrases only. The system also includes a simulated trading engine which was measured to achieve 2.57% investment returns [Sch+12].

The paper from Zhang and Skiena performs a comparative study on how company related news variables reflect the company's stock trading volumes and financial returns. The results show a significant correlation between the variables and the stock price movements. It reveals that opinions in blogs are more persistent over time than news. Based on the findings a trading strategy was created providing investors with consistently favorable returns over a long run [ZS10].

Tetlock based his work on the theoretical models of noise and liquidity traders. His research concentrates purely on the data obtained from the Wall Street Journal which he uses to demonstrate that high media pessimism predicts the fall of the market prices followed by a reversion to fundamentals [Tet07].

There are a couple of other research projects such as [Din+14], [GK10] or [Lee+14] that employ very similar approaches and results.

## 2.3 Commercial products

Today we can find many products which provide online services for stock price prediction (SentimentTrader, SentiTrade), however, their methods and algorithms used for calculating the prediction are not publicly available and are thus useless for the purposes of this thesis. As M. Rechenthin et al [RSS13] mention, websites like DataMinr, Bloomberg or MarketWatch provide the sentiment analysis of stocks but they lack the transparency in how to calculate the stock sentiment. The reason for this "closed source" solution is mostly commercial motivation. Some products (SentimentTrader, SentiTrade) offer their services for a time-limited subscription fee, other websites (MarketWatch, DataMinr) allow free registration but their approach still remains a secret.

A little more transparent solution is the online sentiment analysis tool Sentdex.com maintained by Harrison Kinsley who also runs a YouTube channel providing tutorials on data analysis. Although the Sentdex product has not been described by any academic paper, the tutorial videos and the product homepage provide basic information about the used data and processes. The algorithm running behind is written using Python's Natural Language Toolkit and crawls input data from over 20 of the most famous American journals (Reuters, Bloomberg, WSJ, etc.) [Kin15].

CHAPTER 3

# Input data

To perform machine learning based sentiment analysis and compare its results with stock data prices we need to process three different types of input data, as it is depicted in Figure 3.1. Firstly, we need real stock data which we want to predict and which will be later used to evaluate the accuracy of our prediction. Secondly, it is necessary to define the source of sentiment which is supposed to be analyzed (Social media, etc). And lastly, for supervised machine learning we need sentiment-labeled data suitable for the corresponding source of sentiment. These three types of data sources are further described in the next three sections.

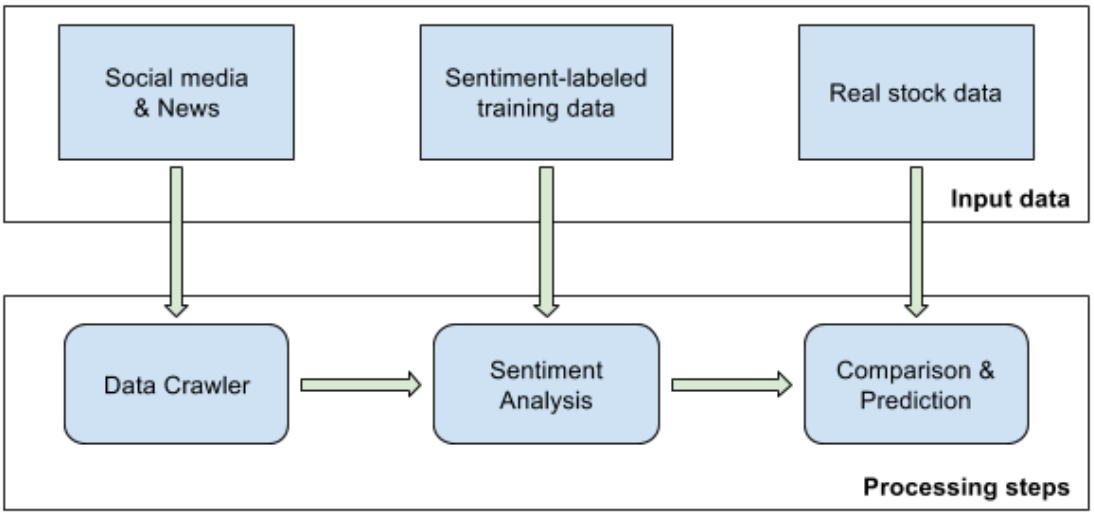


Figure 3.1: Data processing flow

## 3.1 Stock market data

The results of the sentiment analysis need to be compared to some real stock market data. Therefore, it is necessary to define stock variables which are going to be examined for correlation with the measured sentiment values. At the stock market we can either focus on some specific companies (e.g. Microsoft, Apple) or we can concentrate on a broader concept by analyzing various stock market indexes which usually represent a specific group of companies (e.g. Nasdaq, S&P 500). As it was examined in State of the art, both approaches have been taken in existing research works. However, none of these studies was analyzing both sources in order to compare their benefits and prediction possibilities. In this study we have selected several variables from both of the two mentioned groups.

### 3.1.1 Market indexes

Since the language of the analyzed data is English, it makes sense to concentrate only on the indexes from English speaking countries, especially from the biggest one – the United States. For the analysis the three stock market indexes listed below were chosen. When trying to predict their stock price movements we are going to concentrate on general non-topic related data. In other words, we will not search for any specific tweets or sentences containing for example names of the indexes or similar stock index identifiers. More specific information about the corresponding sentiment data crawling is provided in Section 3.2.

#### S&P 500

The so called Standard and Poor's 500 is an index referring to 500 large US companies listed on NYSE or NASDAQ. In comparison with the other two major indexes it is often considered as the one best representing the US stock market because of its large focus.

#### Dow Jones Industrial Average

The DJIA, founded in 1896 by the Wall Street Journal editor, Charles Dow, is a price-weighted average of 30 significant stocks traded on NYSE and NASDAQ. It includes well-known companies such as General Electric Company, Microsoft Corporation, Coca-Cola, etc.

#### Nasdaq

Unlike the previous two indexes, NASDAQ is not only an index but stock exchange itself. It is the second-largest exchange in the world by market capitalization, right after the New York Stock Exchange. It is mostly associated with technology companies like Microsoft, Apple, Google, Tesla, Netflix. All trading happens mostly online.

| official name | search-word | ticker | stock market |
|---|---|---|---|
| The Coca-Cola Company | coca-cola | KO | NYSE |
| McDonald's Corporation | mcdonalds / mcdonald's | MCD | NYSE |
| Microsoft Corporation | microsoft | MSFT | NASDAQ |
| Netflix, Inc | netflix | NFLX | NASDAQ |
| Nike, Inc. | nike | NKE | NYSE |
| Tesla Motors, Inc. | tesla | TSLA | NASDAQ |

Table 3.1: Companies chosen as targets for prediction

### 3.1.2 Companies

When selecting companies for this research, it was necessary to specify some basic criteria. At first, there was a need for unambiguity of the company's name. For instance, Apple is well-known and commonly referenced company at social networks but it is rather unsuitable for text-based search because of its ambiguous name. It is very complicated to identify the difference between "apple" as a fruit and "Apple" as a company, especially on social networks where we cannot rely on the use of capital letters. Another important criterion is the publicity of the company. In order to obtain a sufficient number of records for sentiment analysis, it is necessary to pick such companies which are often discussed in business news as well as among ordinary users of social media. This also requires that the company is somehow directly related to the end customer (B2C market), so we can obtain direct customer feedback in case of a positive or negative response to some product or service.

Based on the given criteria the companies listed in Table 3.1 were chosen. The column "search-word" represents the string used in the collection scrips used to identify the proper posts and sentences in Twitter and news articles. The column ticker serves the same purpose in case of the Stocktwits crawling. Official name and Stock market columns supply additional information which is not taken into consideration during the collection process.

#### Coca-Cola

The Coca-Cola Company is one of the most famous nonalcoholic beverage producers, as well as one of the world's most recognizable brands. It is a home to 20 billion-dollar-brands, including four of the top five soft drinks: Coca-Cola, Diet Coke, Fanta, and Sprite. Other top brands include Minute Maid, Powerade, and Vitaminwater. The company owns or licenses and markets more than 500 beverage brands, mainly sparkling drinks but also waters, juice drinks, energy and sports drinks, and ready-to-drink teas and coffees. With the world's largest beverage distribution system, The Coca-Cola Company reaches its consumers in more than 200 countries.[Mot13a]

Although Coca-Cola divides its Twitter activity into dedicated pages for various products and sub-brands, the main Coca-Cola Twitter feed has still nearly 3.5 million followers.

The social media team rarely posts any straightforward marketing messages and instead primarily uses Twitter to respond to daily mentions (@Coca-Cola) from customers. Such user posts usually include complaints, follow requests or compliments. [Mot13a]

### McDonald's

The McDonald's Corporation is the world's largest chain of hamburger fast food restaurants, employing 1.8 million people and serving more than 58 million customers daily. The business began in 1940 with a restaurant opened by brothers Richard and Maurice McDonald in San Bernardino, California. Nowadays, more than 80% of the McDonald's restaurants in over 100 countries are owned and operated by independent local business men and women in the form of franchising. [17c]

Being one of the most recognizable brands in the world, one can assume that its social accounts are extremely active. As of today, the McDonald's Twitter profile posts several updates every day to entertain its 3.5 million followers. Nevertheless, McDonald's has to battle a fair amount of negative publicity. For example, in 2012 McDonald's used the hashtag #McDStories to promote the video content of their suppliers talking about McDonald's ingredients. However, the campaign was hijacked by consumers complaining about the company's service and the quality of the food. [Mot13b]

### Microsoft

With annual revenues of more than $32 billion, the Microsoft Corporation is more than the largest software company in the world: it is a cultural phenomenon. The company's core business is based on developing, manufacturing, and licensing software products, including operating systems, server applications, business and consumer applications, and software development tools, as well as Internet software, technologies, and services. Led by Bill Gates, the world's wealthiest individual and the most famous businessman, Microsoft has succeeded in placing at least one of its products on virtually every personal computer in the world, setting industry standards and defining markets in the process. [17d]

Due to the scope of its product range and target markets, Microsoft has a huge amount of different social network accounts. The most popular feeds include those of products like Xbox, Bing, Office, etc, but it also attracts a decent number of followers for things like Microsoft Cloud, Security or SQL Server. The main Microsoft Twitter feed has over 8 million followers but it generally only retweets other official accounts or repurposes Facebook content and very rarely responds to mentions. [Mot13c]

### Netflix

Netflix is an American entertainment company specializing in online on-demand video streaming as well as a DVD-by-mail service in the United States. Founded in California in 1997, Netflix began its current subscription model in 1999. Now the company has over

93 million subscribers in 190 countries. The platform offers not only TV shows licensed from distribution but also its original content by investing into its own original series. Unlike traditional broadcasters, Netflix's goal is not to appeal to as broad an audience as possible, but to cater to niches and effectively give every slice of the population a show or movie they cannot live without. [17e]

The great job on social media is considered to be one of the key activities behind the recent success of Netflix. One of the largest waves Netflix causes on social media is due to a mass release of all episodes of a TV show at once. The social media reaction to that kind of marketing is phenomenal. One can see an enormous amount of conversation built up from that one show release which causes a wave of social support helping to drive and build a fan base. Netflix gives up a serialized, weekly drumbeat of social mentions that a television series usually gets. As a result of this, they often see a huge amount of conversation that slowly and steadily declines. To keep the conversation going after the first binge, Netflix releases show details on social media to provide additional spikes a few weeks after. [Moh17]

The relation between the information spread on social media and the stock price of Netflix can be seen on the following example from 2012. Netflix CEO Reed Hastings posted about the company's milestone on Facebook - Netflix's monthly viewing surpassed 1 billion hours for the first time. It was shared by a technology-focused blog an hour later and reached several news outlets within the next two hours. That seemingly simple post caused Netflix's stock to rise from $70.45 to $81.72 the following trading day. [Sim16]

#### Nike

Founded as an importer of Japanese shoes, Nike has grown to be the world's largest marketer of athletic footwear, holding a global market share of approximately 37 percent. In the United States, Nike products are sold through about 22,000 retail accounts. Worldwide, the company's products are sold in more than 160 countries. Both domestically and overseas Nike operates retail stores, including NikeTowns and factory outlets. Nearly all of the items are manufactured by independent contractors, primarily located overseas, with Nike involved in the design, development, and marketing. [17f]

Sports are inherently a social activity so brands like Nike are a natural fit when it comes to social media marketing. On both Facebook and Twitter Nike has individual feeds for its subsidiary brands, as for example golf, basketball, etc. For each of the feeds, the focus is very much on responding to mentions rather than pushing out marketing messages. For example, the Nike.com feed (4.5 million followers) responds to more than 100 tweets per day regarding order queries, stock information and product details. [Mot13d]

#### Tesla

Tesla Motors is an American company that designs and manufactures luxury electric vehicles which are among the world's top-selling models in this branch. It was founded in 2003 by 5 men including Elon Musk, who remains to be the CEO and face of the

company today. Tesla is not only well known for its innovative designs and technologies but also for its focus on the usage of social media marketing. Elon Musk himself writes frequent, lively Twitter posts about Tesla and its Model S electric sports car, along with observations about his professional and personal interests ranging from space travel to movies. All these activities help to build a very popular image of Tesla, especially among the young generation of internet users who support ecology and innovations [Lai16].

Although the company's stock price was on average rising over the past 6 months, the stock curve occasionally encounters some minor or major declines. These could be, among other influences, caused by technical problems which are from time to time reported in the news. As an example we could mention the car burning issues which reached viral attention during March 2017 [Lam17]. It can be interesting to analyze whether such events are measurable in the form of sentiment and how they influence the stock price movements.

## 3.2 Sources of sentiment carrying data

There are many sources of public information which can influence the movements of stock prices. As it was already mentioned in Chapter 2, other research works focus for example on historical economic data or 8-K reports, which are documents notifying the public about various changes in a company, such as acquisition, bankruptcy, resignation of directors, etc. The aim of this research is to focus on publicly available information which reflects the mood in the society. Therefore, when choosing the proper input data sources, the following requirements were taken into consideration:

1. The source is freely available to anybody on the Internet
2. The information is represented in a written form so it is possible to perform textual sentiment analysis
3. There is a possibility to extract topic-related information from the source (i.e. articles about a specific company)
4. The information is accessible on a daily basis.

Based on the presented criteria the following three data sources were chosen:

- Twitter
- Stocktwits
- News articles

For further details see the following Subsections.

### 3.2.1 Twitter

#### Data specification

Twitter is among the top ten most widely used social networks in the world [17g]. It was founded in 2006 in San Francisco and it rapidly gained worldwide popularity reaching 100 million of monthly active users in 2011. At the time of writing this thesis, it has gained a total number of 313 million of monthly active users and this number continues to increase annually by millions of new users [17h]. The subject of our analysis are so called "tweets", 280-character long messages posted by registered users. Although Twitter contains tweets in many languages, our research focuses only on the tweets written in English, which is clearly the most commonly used language on this network [13].

Although there exist other very popular social networks which could possibly provide similarly useful source of information, Twitter was chosen because of its several technical and practical aspects. Unlike Facebook, it allows us to perform wide key-word searches and obtain tweets from any users in the network, not only those who are in our friend list. All can be done in a very simple way by using one of Twitter's REST API which also provides the possibility to specify a time range and filter the results by a keyword. Another benefit is the fact that in other similar research projects it is very common to use Twitter as input data source. Therefore, we have some references of knowledge and inspiration.

Twitter is a rather general social network without any specific topic or target audience which gives it the highest rate of ambiguity out of the three chosen sources. Many of the tweets that are used as the input for our analysis can thus be just some sort of a personal feeling, advertisement or a joke, such as the one in Figure 3.2. They can be written by a user with zero knowledge about the given subject as well as by an expert in a given field. Equally uncertain is the target audience. Some tweets can be economically influential when read by investors or salesmen, while other readers could have no relation to economy or investment whatsoever. However, as the task of this work is to measure the global mood of the given resource, we will not try to distinguish between amateurish and professional posts but we will aim to obtain the overall Twitter sentiment from as many tweets as possible.


Figure 3.2: An example of topic-related post.

| Subject | Working days | Weekends | All days |
|---|---|---|---|
| Non-topic (general) | 13767 | 13550 | 13709 |
| Coca-Cola | 1683 | 1365 | 1598 |
| McDonald's | 6432 | 5710 | 6240 |
| Microsoft | 12121 | 10751 | 11756 |
| Netflix | 9136 | 9007 | 9101 |
| Nike | 11603 | 10996 | 11441 |
| Tesla | 8066 | 5516 | 7386 |

Table 3.2: Average numbers of daily collected tweets during June 2017

#### Collection process

Our collection algorithm uses Search API because unlike Streaming API the collecting script does not need to be running all the time and after several tests it appears to be more stable and reliable. All the company related data are downloaded using keywords which were previously specified in Table 3.1. In order to gather non-company related data (further referenced as "general" data) that will be used to predict the chosen stock indexes, we used a universal key-word "the" because it is not possible to perform the search without a keyword. Search API limits users to download maximum amount of 18000 tweets per a 15 minute window. The need to download more data can be therefore easily solved by implementing a cycled collector with 15 minutes sleep time. Another major limitation is the ability to search only one week back in the tweets' history. Since our data collector runs every day, it should not cause any problems for this research.

The collection script runs daily after midnight and downloads company-related tweets as well as general tweets. The algorithm performs only one iteration of the 15 minute windows, thus downloading maximum of 18000 tweets. This is, however, sufficient for the calculation of daily sentiment because the majority of the observed companies usually do not reach such a number of tweets per day anyway. Table 3.2 shows the average numbers of daily collected tweets in June 2017. In order to make sure that the amount of daily tweets is consistent over the whole week, the table compares values for working days, weekends and all the calendar days. Since all the subjects exceed 1000 records per day, this data should be sufficient for the purposes of sentiment analysis.

### 3.2.2 Stocktwits

#### Data specification

Stocktwits is a social network designed for exchange of ideas and opinions about various economic aspects, mostly stock trends. In a simple way it could be said that Stocktwits is something like Twitter for investors or economy enthusiasts. The network was founded in 2008 and nowadays its total number of users has reached 300 000 [17a]. The vast majority of posts are written in English and the subjects discussed are mostly stock variables or companies listed at some of the biggest US stock markets (NYSE, NASDAQ).

Although the users of Stocktwits are mostly anonymous, we can presume that the vast majority of them are interested in investments and stock market. Why else would they join purely business oriented social network? As a result, when comparing the expertise of its authors and target audience with Twitter, we can expect it to be slightly higher and the stock prediction relevance of each single crawled tweet is expected to be higher as well.

Stocktwits still limits each tweet to be maximum 140 characters long but shares with Twitter features such as referencing to another tweet using "at" sign or allowing to attach various images or videos. On the other hand, Stocktwits differs in a few unique features related to investing. Firstly, the vast majority of tweets contain one or more ticker symbols starting with a dollar sign which references to the corresponding company or stock market index. It can be understood as an ordinary hashtag used in Twitter but referencing only to predefined stock variables. Secondly, each user has the possibility to mark his/her post as "bullish" or "bearish" depending on the current feeling about how the referenced variable might develop in the near future. "Bullish" is a tag of a person who is positive about the future price situation of a specific company or the whole market. In other words, a Bullish person expects the prices to go up. A "Bearish" person is the exact opposite, he/she expects the prices to go down.


Figure 3.3: An example of a Stocktwits topic-related bullish post.

### Collection process

Although the API provided by StockTwits is not as well developed and widely documented as Twitter's API, it provides basic features like key-word search or time-based search, which are necessary for our purposes. Company related data are collected using ticker search. As it was previously mentioned, a user can reference a tweet to a specific company or index by a ticker symbol. Using this feature in the API, it is possible to filter all the posts containing the requested ticker within a specified time period. E.g. "$TSLA" returns all the Tesla-related posts. General (not company related) data can be downloaded from a "suggested" stream. The social network documentation is not very transparent about the logic of marking posts as "suggested". As it is explained in the API documentation, the stream returns messages from "a curated list of quality StockTwits contributors".

Similarly to Twitter, also StockTwits introduces a request limit per time window, which is 6000 posts per hour. As the average number of total posts per day over all categories is much lower, there is no need to implement iterating cycles with sleep time. Although

| Subject | Working days | Weekends | All days |
|---|---|---|---|
| Non-topic (general) | 2660 | 602 | 2111 |
| Coca-Cola | 18 | 5 | 14 |
| McDonald's | 32 | 8 | 25 |
| Microsoft | 126 | 39 | 102 |
| Netflix | 190 | 42 | 150 |
| Nike | 139 | 25 | 108 |
| Tesla | 1277 | 190 | 987 |

Table 3.3: Average number of daily collected Stocktwits posts during February 2017

there is no limit of searching posts in history, this feature was not used to crawl any data before the 1st of November 2016 in order to keep the research consistent for all the three data sources.

Table 3.3 summarizes the average amount of daily collected data in June 2017. As we can see, the numbers are much lower compared to those obtained for Twitter. However, as it was mentioned before, the relevance of Stocktwits posts is expected to be much higher than Twitter tweets. It is hence assumed that even such a small number of input data can provide reliable results. Another noticeable specialty of Stocktwits is a significant decrease of the average amount of posts during weekends. The reason for this is probably the fact that the network is purely stock market oriented. The users are used to react to the newest events happening in the stock market, therefore, the network is active mostly during working days. In the further sections there will be presented various methods of manipulating with weekend values. Although we are not going to remove any low input data sources or processing methods from the analysis, it can be assumed that the low amount of input data is going to have a negative influence on prediction methods using weekend values.

### 3.2.3 News articles

#### Data specification

The third source of information we are going to use are news articles. In comparison with the two previously described sources, this one is not written by random Internet users but mostly by professional journalists. That affects our analysis in several ways. Firstly, there should be less noisy data in the analyzed text because professionally written articles are likely not to contain slang words or grammar mistakes. Secondly, while some posts on social networks can be only jokes or other examples of irrelevant information, the content of news articles should be more serious and relevant to the given topic. Even though the opinion value obtained from media is not directly representing the mood of the ordinary Internet users, the majority of people get their information from news anyway. Moreover, the subjects of news are usually the most current topics resonating in the society, therefore the outcome from news analysis should be equivalently relevant sentiment measure as those obtained from Twitter or Stocktwits.

Unlike tweets, which usually contain one simple message, a single news article can express multiple various information with completely different subjects and sentiment. It would be inappropriate to examine the opinion of the article as a whole because the information relevant to our subject can only constitute a small part of a long article. The analysis will thus rather concentrate on sentiment of each sentence. In case of company based prediction we will take into account only those sentences which contain the given key word.

#### Collection process

Although it is quite simple to write a so called "spider" which downloads news articles by crawling HTML code of chosen websites, this approach is suitable only for downloading general data. When searching for specific company related articles, it would be necessary to crawl huge amount of different websites in order to obtain a sufficient number of relevant articles. A much easier solution is to use the service of an external news search engine. There exist many of such engines but, unfortunately, the vast majority of them are paid or somehow limited. For the purposes of this research we used the free version of Bing Search because its monthly limits are high enough to obtain all the necessary data for free. The engine limits users to perform maximum 1000 API requests per month; each request can search for a specific keyword and it returns maximum 100 articles per query. This amount is sufficient for our needs because, as it was tested, for a given day the first 100 topic related articles contain the most accurate information while the second 100 return already irrelevant results and off-topic articles. It is hence sufficient to perform one request per company (6) per day. Except for the keyword based search, Bing enables a user to search for a pre-defined category, as for example business related articles. This feature is used to obtain the general data.

Table 3.4 shows the average amounts of daily collected sentences from news articles in February 2017. Although the amounts are much lower than those obtained from Twitter, it should be sufficient for the sentiment analysis of all the stock entities. As we can see, there is no significant drop in the data amount for weekends, we should not therefore encounter a problem when using the weekend values in prediction methods, unlike in the case of Stocktwits.

## 3.3 Training data for sentiment analysis

The two most common types of sentiment analysis are machine learning approaches and a lexicon approach. The first one uses labeled data sets as input of supervised learning algorithms while the second approach assigns a polarity score to each word that occurred in the analyzed record based on values contained in a lexicon. Both methods have been used in various sentiment analysis studies. For example, OpinionFinder 2.0, used in famous research from Bollen et al, contains a predefined lexicon assigning each word a polarity score from -1 to +1. [BMZ11]

| Subject | Working days | Weekends | All days |
|---|---|---|---|
| Non-topic (general) | 1018 | 1216 | 1070 |
| Coca-Cola | 86 | 59 | 79 |
| McDonald's | 60 | 56 | 59 |
| Microsoft | 170 | 141 | 162 |
| Netflix | 109 | 121 | 112 |
| Nike | 134 | 81 | 120 |
| Tesla | 192 | 197 | 193 |

Table 3.4: Average number of daily collected sentences from news articles during February 2017

There are many possible variants and setups of both approaches or their combination. The preference of one over the other usually depends on the target domain of the analysis. For the purpose of this research we chose the machine learning approach because of the following reasons:

- According to the comparative study from Kolchyna et al., the machine learning approach on average outperforms the lexicon based approach. [Kol+15]
- When we train a machine learning classifier on data similar to the target domain, the classifier can learn and adapt to the specific features of the domain language. This can be effective for the informal language used on social networks.
- When working with labeled data sets, we can measure and compare the performance of the classifier, as for example accuracy, precision and recall.
- We can compare and combine various machine learning algorithms, as for example Naive Bayes, Maximum Entropy and Support Vector Machine.

When searching for the suitable datasets to be used in the machine learning process, we have to take into consideration the target domain of our analysis. Obviously, it would not be optimal to use just some general sentiment-labeled dataset for all the three domains together. For example, the IMDB movie reviews dataset is very popular in many sentiment analysis tutorials and it is even included in the default NLTK library. However, because of its domain language the dataset is very inconvenient for business oriented sentiment analysis. While the most informative features of the IMDB based trainer are probably words like "star", "boring" or "waste", our algorithm should rather focus on more general or business related features, as for example "great", "success" or "fall". The best approach is that each of our three data sources has its own training dataset consisting of a subset of records which were extracted from the source itself or which are very similar to the target domain.

### 3.3.1 Twitter

Being one of the top used social networks, Twitter posts are very popular subject of various sentiment analyses. There is a big choice of existing datasets with labeled sentiment values. Some of them are manually labeled by individuals or a group of people, while other datasets are automatically generated using so called "Distant supervision". It means that the positive versus negative sentiment of a given tweet is decided based on some heuristic rules, as for example the presence of happy or sad emoticons. For our study, we have used a publicly available dataset from a project called Sentiment140. It contains 16 million tweets marked with values "0" as negative and "4" as positive. As explained on the web Sentiment140, the training set was automatically created using the assumption that any tweets with positive emoticons are positive, while those with negative emoticons are negative. [GBH09] The following bullet list presents five examples of a negative training dataset.

- *I hate when I cant sleep*
- *the weekends over*
- *Going to veterinarian. My cat is sick*
- *Ate so much today. ...fatty!*
- *i dont feel too good*

As the study [GBH09] proposes, it is better to clean the dataset from noisy data before the training process. The most common operations are the removal of user references and the removal of URLs. As the dataset contains 1.6 million records, we do not need to remove anything but we can simply filter out such tweets which do not contain any links and references. For the training maximally 50 000 of input records are needed.

### 3.3.2 Stocktwits

Stocktwits is a rather small and not widely known social network so it is not very surprising that an Internet search for a sentiment labeled Stocktwits dataset was unsuccessful. Another option, using the same training data as for Twitter, seems possible but not an optimal solution. When we compare the content of tweets from both networks, we realize that their mood expressing words and the format of sentences highly differ. While very common positive words used on Twitter include "love" or "cool", we would hardly search for such occurrences on Stocktwits because the majority of positive posts there use rather business oriented slang phrases, such as "going up" or "bullish".

Fortunately, the social network itself offers another possible solution. Because the network is investment oriented, it enables users to express their "buy" vs "sell" attitude by marking their post with "Bullish" vs "Bearish" label. Since this data are available to download over API together with the posts, it enables us to build our own sentiment-labeled dataset with original Stocktwits posts. Despite the small portion of posts which are labeled

this way (roughly only 5%), it is not a problem to obtain a sufficient dataset of several thousands records by connecting the data over a long period of time. The following bullet list shows five examples of a “Bullish” training set.

- *Tesla Stock Is Showing Signs of Hope $TSLA*
- *$ACIA among other strong signals, we have a new CCI buy signal.*
- *$ARLZ moving higher on volume spike*
- *$AMAT now at 16-year highs*
- *$DAL flights to Havana starting soon*

Although the described dataset creating approach seems simple and effective, it raises a question whether we can rely on the sentiment labels created by mostly anonymous Stocktwits users. To tackle this problem, we will take a look at the results of a study from Michael Rechenthin called "Stock chatter: Using stock sentiment to predict price direction" [RSS13]. The paper describes research conducted on data obtained from Yahoo stock message boards which used to have a very similar structure and content language to today’s Stocktwits. The message boards contributors were also mostly anonymous users interested in investments, who discussed their opinions and strategies in a message thread belonging to a specific stock market entity. Moreover, the users had a similar possibility to publish an explicit value of their message sentiment, in this case they had a choice of five options: “strong buy”, “buy”, “hold”, “sell”, and “strong sell”.

The "Stock chatter" paper, another example of stock prediction research, questions the existence of dishonest posters who are capitalizing on the popularity of the boards by writing sentiment in line with their trading goals as a means of influencing others and therefore undermining the reliability of the boards. In order to identify these dishonest posters, M. Rechenthin performed a test using unbiased sentiment analyzer. The test identified some of so called "pump and dumpers", however, their number was so small that after eliminating their posts from the testing data, there was no discernible difference to be found. Given these results, we conclude to not try to identify or eliminate any dishonest Stocktwits posters because we assume their number would be similarly insignificant as it has been reported in case of Yahoo message boards.

In October 2016 Yahoo cancelled the old message boards and came up with so called "conversations" instead. The new structure was probably not very welcome by its users because the usage of yahoo conversations is very small nowadays. Although there is no evidence, it could be assumed that part of the former Yahoo message boards users switched to Stocktwits.

#### News articles

For the previous two data sources it was very easy to either download an existing dataset or to create one using extracted mood labels. However, in case of business-related news

articles there is not such a simple solution. Our target data are sentences describing some company related information, as for example “Tesla has been at the lead of innovation” or “Tesla is facing serious development issues”. An optimal training set should therefore contain most informative features, as for example “success”, “problem”, “increase”, “fall”, etc. Unfortunately, after conducting an extensive search through the Internet, we did not find any suitable dataset. The majority of publicly available datasets are extracted from social networks or user reviews and contain many rather informal and emotional features, as for instance “love’, “hate”, “cool” or “crazy”. These words are, however, very seldom used in official texts used in public media. Thus, the only suitable solution for this problem is to create a new labeled dataset from a subset of the target data.

There are several possible approaches to create our own dataset from the existing unlabeled data. The most basic one is to create one manually, however, in order to obtain a sufficient number of records, it would be necessary to read through approximately 10000 sentences because many of them do not have any significant sentiment and cannot be used as training data. Another option is to generate a dataset from the existing data programmatically. This can include another sentiment analyzer or a lexicon with sentiment labeled words. Usage of Twitter or Stocktwits training data is again facing the problem of the domain language difference between news articles and Twitter or Stocktwits posts. Given these arguments, the chosen approach for generating a new dataset was decided to be the usage of an existing lexicon containing pre-labeled positive and negative words. The lexicon was obtained from a study of B. Liu et al. who defined 2000 positive and 4000 negative words. [LHC05] Unlike the majority of other publicly available lexicons, this one does not limit the lists to adjectives or adverbs only but contains sentiment significant verbs and nouns as well.

Using a simple Python script, we can loop through the existing sentences and divide a subset of these into positive and negative groups. A sentence is labeled as positive when it contains one or more words from the positive lexicon and zero words from negative lexicon and vice versa. In case there is a negation (“not” or “n’t”) in the sentence, it is stored to a special set of records which are then manually assigned to the correct sentiment group. Similarly, the occurrences of both positive and negative words in one sentence are manually processed as well. Fortunately, the percentage of records which needed to be processed manually was rather low, approximately 10% of the total amount. Sentences containing no words from either of the lexicons were ignored. The following bullet list shows some examples of the resulting positive training set.

- *The company said that was its "best result in 10 years."*
- *Yet other signs are encouraging.*
- *The stock gained $3.81 to $120.79.*
- *But there’s a happy ending, anyway.*
- *It’s wonderful that they are contributing to the affordability.*

CHAPTER 4

# Sentiment analysis

In the following sections we will describe the text representation model and examine the influence of various data preprocessing techniques which are commonly used in natural language processing tasks. For each technique there will be performed an accuracy test considering different data manipulation scenarios. Based on the test results we will discuss and evaluate the influence and suitability of the given technique on our sentiment classification. The performance will be tested independently for all the three data sources, i.e. Twitter, Stocktwits and News, as described in the Chapter 3.

## 4.1 Text representation model

When modeling text with machine learning algorithms, we have to define the model used to represent the input data. For the purposes of this study there was used a very common model called "Bag-of-Words" complemented by "Vector space model". The Bag-of-Words model is simple to understand and implement, and has experienced success in solving problems such as language modeling or document classification. The idea is that for a given sentence or tweet we extract the unigram words (terms) only to create an unordered list of words representing the document. After this step, there follows some further text processing, as for example POS tagging, stemming, bigrams, etc. [Bro17]

Using the bag of words that we extracted and modified in the previous step we create a vocabulary of all the known words. This is further used to define a vector where each feature is a word (term) and the feature's value is either 1 or 0 indicating whether the term occurs in the input record or not. In the end of this process, we end up with a set of vectors where one vector represents exactly one input record (tweet or sentence). The aim of the whole process is that the input data representation changes from the textual data to an ordered list of binary values (vectors).

In order to better explain the above described steps, we will provide an example of the whole process. In the following bullet list there are three example records.

- A new product is on the market.
- The new product is very popular.
- The old product has low quality.

All the three records together form the following vocabulary:
*"a", "new", "product", "is", "on", "the", "market", "very", "popular", "old", "has", "low", "quality"*

The third sentence "The old product has low quality." forms the following feature vector:
*"a" = 0, "new" = 0, "product" = 1, "is" = 0, "on" = 0, "the" = 1, "market" = 0, "very" = 0, "popular" = 0, "old" = 1, "has" = 1, "low" = 1, "quality" = 1*

"The old product has low quality." can be therefore expressed in binary values as: [0,0,1,0,0,1,0,0,0,1,1,1,1]

## 4.2 Performance evaluation metrics

Using the training datasets described in Section 3.3, we created three distinct datasets, each containing 5000 sentiment labelled records. Although the Twitter dataset contains much higher number of records, there was a need to decrease the tedious processing time which was caused by a high number of total test runs. Before each accuracy test, the records were randomly divided into a training set and a testing set using the percent ratio 75:25. All the tests were performed using the NLTK's Naive Bayes classifier. In order to obtain representative outcomes, the accuracy measures were calculated as mean values over 10 independent test iterations. Each test run results in precision, recall and accuracy which are calculated using the following four measures:

- **True Positives (TP):** number of positive records, labeled as positive.
- **False Positives (FP):** number of negative records, labeled as positive.
- **True Positives (TN):** number of negative records, labeled as negative.
- **False Negatives (FN):** number of positive records, labeled as negative.

**Precision**: The "exactness" of the classification tells us what percentage of records that the classifier labeled as positive are actually positive and vice versa. It can be expressed with a reference to either positive or negative records. For instance, as it is shown in Formula 4.1, positive precision is calculated as the number of true positives over the number of true positives plus the number of false positives.

$$\frac{TP}{TP + FP} \tag{4.1}$$

**Recall**: The "completeness" of the classification tells what percentage of positive records did the classifier label as positive and vice versa. Similarly to precision, it can be expressed with a reference to either positive or negative records. As we can see in Formula 4.2, positive recall is calculated as the number of true positives over the number of true positives plus the number of false negatives.

$$\frac{TP}{TP + FN} \tag{4.2}$$

**Accuracy**: Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. As the Formula 4.3 shows, the calculation can be expressed as the sum of correctly classified records (TP + TN) over the sum of all the records (TP + TN + FP + FN).

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{4.3}$$

## 4.3 Data preprocessing

### 4.3.1 Part-of-speech tagging

The process of part-of-speech tagging marks up each word in the text as corresponding to a particular part of speech, such as nouns, adjectives, verbs, etc. Knowing the correct POS tag can be beneficial for many language processing tasks, including sentiment analysis. The idea is based on the assumption that some words in a sentence have zero information gain of sentiment and, therefore, they could be simply removed from the input. That forces the classifier to concentrate only on those words which are believed to be the sentiment holders. The most important sentiment carrying parts of speech are usually expected to be adjectives, adverbs, nouns and verbs. However, depending on the target domain, also interjections or other groups can be influential. Another benefit of POS tagging can be a decrease of the feature set size which can improve the computational time but, on the other hand, the POS tagging itself is quite memory consuming as well.

As it has been already discussed in other sentiment analysis studies [Kol+15], there is no strong evidence of POS removal having some significant influence on the accuracy of classification. While some researchers report positive influence of POS tagging [BF10], others obtained neutral or even negative impact [PLV02]. Similar outcomes were obtained during the development of our classifiers. Table 4.1 lists eight word classes which were used to create four different POS testing combinations. These combinations and their test results are shown in Table 4.2. When looking at the accuracy column, we see that

only Twitter classifier significantly reacts to the use of POS. In this case, the accuracy measured when using part-of-speech tagging dropped by approximately 1-2 % over the test run with no POS tagging used. On the other hand, Stocktwits and News classifiers do not show any significant accuracy changes in any of the four combinations.

| shortcut | word class |
|---|---|
| E | existential there ("there is") |
| F | foreign word |
| J | adjective |
| N | noun |
| P | pronoun |
| R | adverb |
| U | interjection |
| V | verb |

Table 4.1: NLTK word classes used for POS tests

| **POS combination** | **accuracy** | **precision-pos** | **precision-neg** | **recall-pos** | **recall-neg** |
|---|---|---|---|---|---|
| **Twitter** | | | | | |
| JVNR | 70.94 | 74.92 | 68.12 | 63.09 | 78.83 |
| JVNREU | 71.09 | 75.08 | 68.28 | 63.28 | 78.95 |
| JVNREUFP | 71.3 | 76.44 | 67.87 | 61.66 | 80.95 |
| -ALL- | 72.14 | 77.4 | 68.64 | 62.62 | 81.65 |
| **Stocktwits** | | | | | |
| JVNR | 77.92 | 81.7 | 74.94 | 71.95 | 83.89 |
| JVNREU | 77.94 | 81.71 | 74.97 | 71.99 | 83.88 |
| JVNREUFP | 77.93 | 82.28 | 74.62 | 71.21 | 84.66 |
| -ALL- | 77.87 | 85.11 | 73.12 | 67.58 | 88.14 |
| **News** | | | | | |
| JVNR | 81.87 | 81.32 | 82.47 | 82.76 | 80.97 |
| JVNREU | 81.82 | 81.26 | 82.4 | 82.68 | 80.91 |
| JVNREUFP | 81.88 | 81.05 | 82.77 | 83.22 | 80.51 |
| -ALL- | 81.77 | 81.35 | 82.23 | 82.47 | 81.07 |

Table 4.2: Accuracy for specific POS combinations

Other neutral results were obtained for all the precision and recall variables in News outcomes. In contrast, Stocktwits show a positive improvement in the precision-recall balance for all the POS combination. Without POS tagging, the Stocktwits classifier apparently inclines to use a rather negative sentiment label which causes decrease of negative precision (67.58) and increases its recall (88.14), having a similar but opposite effect on the other two variables. The combinations "JVNR" and "JVNREU" resulted in the best balanced outcomes. POS tagging, therefore, seems to have a positive influence on the Stocktwits classifier. Similarly, better balanced variables were measured for the

Twitter classifier but in this case, we are dealing solely with the decrease in positive precision and negative recall, obtaining no real performance improvement.

### 4.3.2 Case insensitivity

Ignoring the difference between uppercase and lowercase letters is a very common approach in sentiment analysis. It is assumed that two words differentiating from each other only in the letter case represent the same meaning and carry the same sentiment value. Case insensitive analysis then enhances the correctness of feature matching and increases the accuracy. An exception could occur in the case of the words with multiple meanings. Here, it is necessary to distinguish proper nouns, as for instance company names, from their original meaning. To give an example, we can mention "Apple" as a company name vs. "apple" as a fruit. Fortunately, for our analysis we chose only such companies whose names are unambiguous and we are thus not bounded by this problem.

As it is visible in Table 4.3, both Twitter and Stocktwits classifiers react positively to the usage of case insensitivity. All the five variables show some increase in their values by approximately 1%. The News classifier does not show any significant change of performance but the technique of case insensitivity is beneficial even in such a case because it decreases the length of the feature set and therefore speeds up the processing time.

| **case sensitive** | **accuracy** | **precision-pos** | **precision-neg** | **recall-pos** | **recall-neg** |
|---|---|---|---|---|---|
| | | | **Twitter** | | |
| True | 70.78 | 76.49 | 67.18 | 60.1 | 81.5 |
| False | 71.97 | 77.31 | 68.45 | 62.37 | 81.59 |
| | | | **Stocktwits** | | |
| True | 77.39 | 84.01 | 72.98 | 67.65 | 87.12 |
| False | 78.27 | 85.72 | 73.48 | 67.9 | 88.67 |
| | | | **News** | | |
| True | 82.38 | 82.36 | 82.41 | 82.4 | 82.34 |
| False | 82.33 | 82.18 | 82.51 | 82.6 | 82.06 |

Table 4.3: Accuracy measured for usage of case (in)sensitivity

### 4.3.3 Stop words

The term stop words denotes a predefined group of words which are considered to have no information gain for sentiment analysis. The idea is similar to the part-of-speech removal but in this case we do not concentrate on the whole word classes but we group all unwanted words of various word classes into one data set. There exist multiple different sets of stop words but their content is usually quite similar, mostly differentiating only by its size. For this analysis, we used the one available in Python's NLTK library which contains approximately 150 stop words and mostly consists of prepositions (on, at),

articles (a, the) and some of the most common, short functional words (is, were) etc. Like in the case of POS tagging, the benefits of stop words removal are the focus on real sentiment holders and decrease of the feature set length. The advantage of stop words over POS tagging is that the process of removing a predefined word list is much easier and faster in comparison with the memory consuming word labelling.

The test results are shown in Table 4.4. Stop words removal caused a significant improvement of Twitter's precision and recall by making the variables better balanced, yet the classifier's accuracy declined by approximately 1%. The Stocktwits classifier encounters not only a similar improvement in balance of precision and recall but it also shows a small increase in the accuracy performance. This makes stop words removal even more beneficial to the Stocktwits classification than the similar technique, POS tagging, which had previously encountered similar results without any accuracy gain. It seems that stop words are causing some unwanted bias in the Stocktwits classfication. The News classifier outcome, on the other hand, does not show any interesting changes.

| **stop words removed** | **accuracy** | **precision-pos** | **precision-neg** | **recall-pos** | **recall-neg** |
|---|---|---|---|---|---|
| | | | **Twitter** | | |
| True | 71.03 | 71.67 | 70.49 | 69.62 | 72.46 |
| False | 72.1 | 76.78 | 68.88 | 63.41 | 80.79 |
| | | | **Stocktwits** | | |
| True | 78.85 | 82.43 | 76.0 | 73.37 | 84.34 |
| False | 78.11 | 85.43 | 73.34 | 67.85 | 88.41 |
| | | | **News** | | |
| True | 81.03 | 80.57 | 81.49 | 81.76 | 80.28 |
| False | 81.32 | 80.98 | 81.66 | 81.83 | 80.8 |

Table 4.4: Accuracy measured for removal of stop words

### 4.3.4 Stemming and lemmatization

As the research from Christopher Manning et al. [MRS08] explains, the goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance: "am", "are", is transformed to "be", etc. The difference between lemmatization and stemming is that a stemmer operates on a single word without knowing the context, and therefore cannot discriminate between words which have different meanings, depending on the part of speech. For instance, the word "better" has "good" as its lemma but this link is missed by stemming, as it requires a dictionary look-up. On the other hand, stemmers are typically easier to implement and run faster, and the reduced accuracy may not matter for some applications. Another benefit of both stemming and lemmatization is a decrease of the feature set length and thus a decrease of the processing time.

For the testing purposes, we have chosen the Porter's stemmer which is one of the oldest and still very popular stemming algorithms [MRS08]. The WordNet lemmatizer and usage of no stemming or lemmatizing technique were the other two tested options. The results are shown in Table 4.5. While the effect in the Stocktwits data set is just very small, both Twitter and News classifiers denote a significant improvement in all the variables when using the Porter stemmer compared to usage of no stemming techniques. The WordNet lemmatizer shows a positive impact as well but the effect of the Porter stemmer is slightly higher.

| technique | accuracy | precision-pos | precision-neg | recall-pos | recall-neg |
|---|---|---|---|---|---|
| **Twitter** | | | | | |
| None | 71.46 | 75.81 | 68.47 | 63.16 | 79.81 |
| Lemma | 71.7 | 76.22 | 68.56 | 63.13 | 80.3 |
| Stemmer | 73.02 | 77.58 | 69.8 | 64.77 | 81.25 |
| **Stocktwits** | | | | | |
| None | 77.71 | 84.73 | 73.11 | 67.65 | 87.79 |
| Lemma | 77.48 | 84.64 | 72.82 | 67.17 | 87.8 |
| Stemmer | 77.75 | 84.98 | 73.04 | 67.36 | 88.1 |
| **News** | | | | | |
| None | 81.95 | 81.5 | 82.42 | 82.67 | 81.22 |
| Lemma | 82.87 | 82.35 | 83.43 | 83.68 | 82.07 |
| Stemmer | 83.37 | 82.87 | 83.94 | 84.18 | 82.53 |

Table 4.5: Accuracy measured for usage of stemming and lemmatization

### 4.3.5 Bigrams

The term "bigram" denotes a tuple of two neighboring words from the analyzed text. This means that when parsing the sentence "Company's debt increased", we obtain tuple features "(company, debt), (debt, increased)". The importance of bigrams lies in the fact that some words show different sentiment when coupled with their neighbors. For instance, the word "increased" would be probably classified as positive when analyzed as unigram, however, in tuple with the word "debt" it changes sentiment to negative. A similar function works in the case of negation ("not", "n't"). The disadvantage of using bigrams is the double size of the feature set length which dramatically slows down the processing time, especially when using high amount of training data. On the other hand, this shortcoming can be effectively solved by implementing some feature sorting and limiting techniques.

Table 4.6 shows the accuracy measured for bigrams testing. While the usage of bigrams in Twitter and Stocktwits classifiers denotes a significant increase in accuracy and some precision and recall variables, the News analysis shows the exact opposite. This can be caused by the domain language differences. Both Twitter and Stocktwits posts are written by ordinary Internet users and therefore they contain many phrases commonly

used in informal language, as for example "miss my", "I wish", "can't sleep" or "look good", which can be better identified using bigrams. News articles are usually denoting sentiment by using simple expressive words, such as "progress", "fail", "support" or "debt". Bigrams are therefore apparently unsuitable for our News analysis.

| bigrams | accuracy | precision-pos | precision-neg | recall-pos | recall-neg |
|---|---|---|---|---|---|
| **Twitter** | | | | | |
| True | 72.53 | 77.17 | 69.28 | 64.07 | 80.98 |
| False | 71.84 | 76.49 | 68.7 | 63.24 | 80.44 |
| **Stocktwits** | | | | | |
| True | 79.39 | 85.4 | 75.11 | 70.9 | 87.85 |
| False | 77.82 | 85.6 | 72.87 | 66.93 | 88.71 |
| **News** | | | | | |
| True | 80.33 | 80.02 | 80.63 | 80.8 | 79.8 |
| False | 82.03 | 81.88 | 82.24 | 82.33 | 81.74 |

Table 4.6: Accuracy measured for usage of bigrams

### 4.3.6 Optimal combination of preprocessing techniques

Now, when all the preprocessing techniques were tested and evaluated, it is necessary to combine them in such a way as to obtain the best possible performance. As it was realized, our three input data domains do not react identically to the tested techniques. As a result, each classifier design has to be discussed individually.

The first source domain, Twitter, showed very clear performance boost when using case insensitivity, Porter stemming and bigrams. All the three techniques increase all the variables by approximately 1%. Although none of them helps to balance the irregular values of precision and recall, they do not add any negative influence either. POS tagging and stop words removal both improve precision and recall balance but decrease accuracy. Since the loss in accuracy (1%) in comparison with the difference in positive and negative precision (10-20%) is rather insignificant, it would be better to prioritize the balanced state over accuracy gain. The problem is that both of these techniques have a similar function and combining them might be useless. Therefore, in order to find out which one to use, it is necessary to compare all the three possible combinations: Stop words only, POS tagging only and a combination of both (shown in Table 4.7). It is clear that Stop words perform significantly better then POS tagging and their combination does not bring any further improvement. Given these facts, only Stop words removal, case insensitivity, Porter stemming and bigrams will be added to the final Twitter classifier.

Stocktwits outcomes are similar to the Twitter ones. Both case insensitivity and bigrams clearly increase all the values, bigrams even improve the variable balance. Although Porter stemming does not show any significant changes, it can be beneficial to include it because of its feature length reduction. Similarly to the Twitter classifier, also Stocktwits shows a tendency of improving precision and recall balance when including POS tagging

| technique | accuracy | precision-pos | precision-neg | recall-pos | recall-neg |
|---|---|---|---|---|---|
| Stop words | 71.03 | 71.67 | 70.49 | 69.62 | 72.46 |
| JVNR POS | 70.94 | 74.92 | 68.12 | 63.09 | 78.83 |
| Both | 70.88 | 71.35 | 70.62 | 70.05 | 71.72 |

Table 4.7: Twitter: Combination of Stop words and POS tagging

or Stop words removal. In contrast, accuracy in this case stays the same (POS), or even improves (Stop). Table 4.8 shows comparison of Stop words and POS tagging. Although Stop words removal clearly outperforms POS tagging, the combination of both techniques certainly provides the best balanced state. Since the decrease of accuracy compared to the increase of negative precision is rather insignificant, we choose to include both techniques. The final Stocktwits classifier will therefore consist of all the preprocessing techniques.

| technique | accuracy | precision-pos | precision-neg | recall-pos | recall-neg |
|---|---|---|---|---|---|
| Stop words | 78.85 | 82.43 | 76.0 | 73.37 | 84.34 |
| JVNR POS | 77.92 | 81.7 | 74.94 | 71.95 | 83.89 |
| Both | 77.67 | 77.59 | 77.84 | 77.9 | 77.44 |

Table 4.8: Stocktwits: Combination of Stop words and POS tagging

Unlike Twitter and Stocktwits, the News domain does not show such sensitive reactions to input data manipulation. The only significant results are a positive influence of the Porter stemmer and a negative influence of Bigrams, both changing all the variables by approximately 1-2%. The precision-recall balance is in a very good state in all the testing results and we thus do not need to apply any corrections, as it was the case with the other classifiers. The News react neutrally to POS tagging, Stop words removal and case sensitivity. Out of these, it makes sense to include only case insensitivity because it reduces the feature set length. The final combinations of all the previously analyzed preprocessing steps are summarized in Table 4.9. Their resulting accuracy performance is shown in Table 4.10.

| Data source | POS | lower case | stop removal | stemming | bigrams |
|---|---|---|---|---|---|
| Twitter | N | Y | N | Porter | Y |
| Stocktwits | N | Y | Y | Porter | Y |
| News | N | Y | N | Porter | N |

Table 4.9: Optimal combinations of preprocessing tasks

### 4.3.7 Size of training set

As it was previously mentioned, in all the tests above the training set of 3750 records was used (three quarters of 5000). The reason to use such amount of data was mainly

| Data source | accuracy | precision-pos | precision-neg | recall-pos | recall-neg |
|---|---|---|---|---|---|
| Twitter | 72.09 | 73.79 | 70.69 | 68.61 | 75.61 |
| Stocktwits | 79.64 | 80.82 | 78.73 | 77.2 | 82.06 |
| News | 83.2 | 83.08 | 83.34 | 83.38 | 83.02 |

Table 4.10: Accuracy after the optimal combination

caused by the need to perform a high number of tests in a reasonable time. Now, when we already know the influence of various preprocessing operations, we can try to boost the classifier accuracy by increasing the training set size.

In the case of News and Stocktwits, classifiers were limited by the data set size which was already used at its maximum. But the classifiers are reaching pretty high performance numbers (approximately 80% and higher) already. According to the agreement study of Wilson et al., different people agree on sentiment value of text only in 82% and it thus makes no sense to seek for higher outcomes of sentiment classification.

Much bigger space for improvement is offered by the Twitter classifier. The highest accuracy reached during the preprocessing tests was still under 75%. Fortunately, the dataset of 5000 tweets used for the test is just a small part of the huge dataset provided by the Sentiment140 research. That enables us to boost the accuracy by increasing the training set size. Obviously, the performance improvement is not infinite. There needs to be found a specific size of the training set at which it reaches its peak and does not increase anymore. Such a point will offer the optimal compromise between high accuracy and low processing time.

After performing several tests iterating through a number of training set sizes reaching from 1000 till 40000, we have obtained the results depicted in Figure 4.1. The results clearly show that the sought border of optimal data size is at the amount of 20000 records in the training set.

### 4.3.8 Filtering high informative features

In the previous paragraphs we have analyzed and specified the optimal preprocessing tasks and the best (or best available) sizes of the training sets for all the input domains. When combining all the best performing properties, as it was summarized in Table 4.9, we are facing the problem of a very extensive feature set size. As Table 4.11 shows, in case of the Twitter classifier, we end up with 130 thousand features. Although the usage of stemming or stop words removal provided some reduction of the feature set size, the usage of bigrams made the numbers doubled. The feature sets now contain all the words and tuples which occurred in the training data. That includes many low information features which do not increase the prediction accuracy and only increase the processing time. Eliminating low information features can give our model clarity by removing noisy data and save it from over-fitting. When we use only the features with higher information,

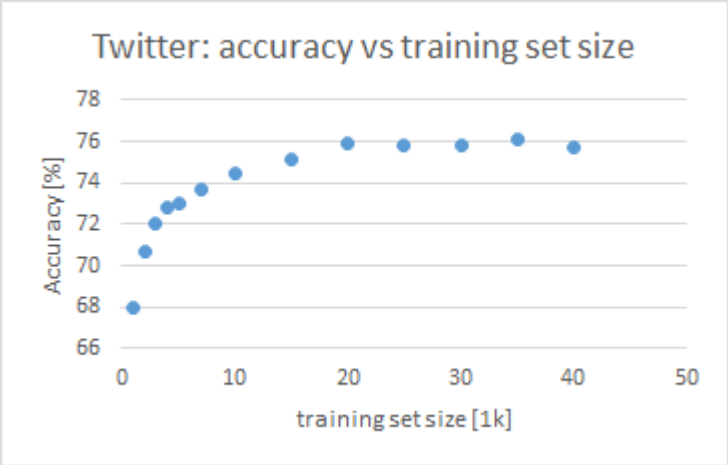


Figure 4.1: Twitter: Accuracy dependence on training set size

we can increase the performance while also decreasing the size of the model. This will result in smaller memory usage along with faster training and classification.

| Data source | training set [records] | unlimited feature set | optimal feature set |
|---|---|---|---|
| Twitter | 20000 | 130198 | 15000 |
| Stocktwits | 5000 | 35830 | 25000 |
| News | 5000 | 32171 | 25000 |

Table 4.11: Data set and feature set lengths

| Data source | accuracy | precision-pos | precision-neg | recall-pos | recall-neg |
|---|---|---|---|---|---|
| Twitter | 76.03 | 78.03 | 74.32 | 72.47 | 79.63 |
| Stocktwits | 79.78 | 81.15 | 78.55 | 77.6 | 81.97 |
| News | 83.25 | 81.72 | 84.95 | 85.68 | 80.82 |

Table 4.12: Accuracy after training set and feature set modifications

To find the highest information features, we need to calculate information gain for each word. Information gain for classification is a measure of how common a feature is in a particular class, compared to how common it is in all other classes. A word that occurs primarily in positive sentences and rarely in negative sentences has a high information gain. For example, the presence of the word “great” in a tweet is a strong indicator that the tweet is positive. That makes “great” a high information word.

One of the best metrics for the information gain is chi square. NLTK library includes this measure in the BigramAssocMeasures class in the metrics package. To use it we need to calculate three frequencies for each word: its overall frequency and its frequency within

both classes (positive/negative). Using these frequencies, we can calculate the observed value (Oi) and Expected value (Ei) to obtain the Chi-Square statistic from Formula 4.4. A high value of chi-square means there is a high information gain for the given word.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{4.4}$$

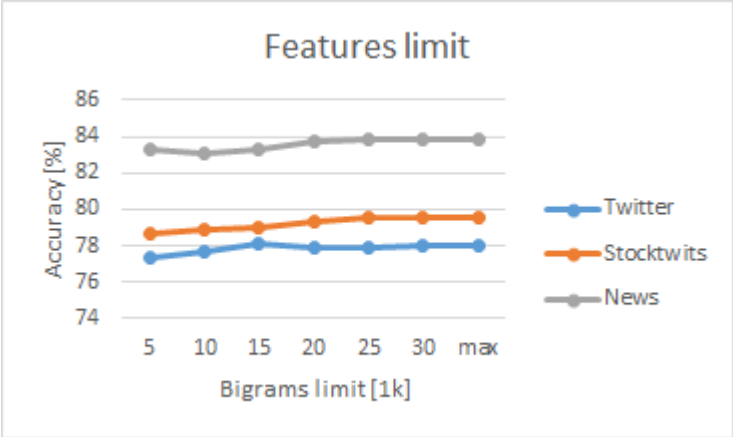


Figure 4.2: Most informative features

By sorting the feature set according to its information gain and iterating over multiple feature set sizes we have found the optimal limit of features at which the accuracy does not increase anymore. Figure 4.2 shows the results of the iterative test measuring the accuracy for most informative features limited from 5000 till 30000. The label "max" represents the unlimited size from Table 4.11. Based on the given results the values presented in the column "optimal feature set" of Table 4.11 were specified. Table 4.12 shows the resulting accuracy values after the feature set optimization.

## 4.4 Machine learning classifiers

The implementation of supervised machine learning algorithms was inspired by the NLTK tutorial provided by Sentdex.com and StreamHacker.com.[Per12] The choice of the five machine learning algorithms mentioned in the next sections was also supported by several existing studies, for example Go et al implement their Sentiment140 study using Naive Bayes, SVM and Maximum Entropy classifiers [GBH09].

The optimal combination of preprocessing tasks obtained in the previous steps was used. Similarly to the outputs of the preprocessing method tests, also these outcomes represent average of 10 independent runs.

### 4.4.1 Naive Bayes (Multinomial and Bernoulli)

In the most of machine learning or sentiment analysis tutorials Naive Bayes is the very first option to start with. It is one of the easiest classifiers which can be used because of its simplicity to be implemented in various programming languages (Python, Java, etc). The algorithm is based on the probabilistic theory of the Bayes theorem with strong and naive independence assumptions. However, despite this almost elementary properties, Naive Bayes performs well in various applications including text classification, sentiment detection, spam filtering or topic categorization. Although the classifier is often outperformed by more complicated techniques, one of its main advantages is low computational requirements. It can be therefore used in applications where higher CPU performance is not available.

There are multiple implementations of the Naive Bayes classification. Based on the information from the above mentioned sources [Per12] Multinomial NB and Bernoulli NB were chosen as suitable for sentiment analysis. Multinomial NB implements the naive Bayes algorithm for multinomial distributed data. It is suitable for classification with discrete features (e.g., word counts for text classification) and for tasks where the multiple occurrences of the grams matter. Bernoulli NB, on the other hand, does not take the number of word occurrences into consideration, which means that it works with binary-valued features. Another dissimilarity lies in the way of processing non-occurring terms. While Multinomial NB takes into account only the occurring words, Bernoulli NB assigns a boolean term to all the terms in the vocabulary, therefore also the absence of words is taken into consideration.

### 4.4.2 Logistic Regression (Maximum Entropy)

Logistic Regression, sometimes referred as Maximum Entropy, is a powerful statistical way of modeling a binomial outcome. Similarly to Naive Bayes, it works by extracting some set of weighted features from the input and combining them linearly (meaning that each feature is multiplied by a weight and then added up). In comparison with the Naive Bayes classifier, it does not assume that the features are conditionally independent of each other. It is based on the Principle of Maximum Entropy stating that the model best representing the current state of knowledge is the one with the largest entropy. The classifier is commonly used in a large variety of textual analysis tasks, such as language detection, topic classification or sentiment extraction. When compared to Naive Bayes, it requires higher CPU performance to train, mostly because of optimization processes needed to estimate the parameters of the model.

### 4.4.3 Support Vector Classification (Nu and Linear)

The Support Vector Machine algorithm is a simple linear classification/regression algorithm which tries to find a hyperplane which separates data into two classes as optimally as possible. The advantages of SVM are effectiveness in high dimensional spaces and memory efficiency. On the other hand, the method could result in a poor performance

when the number of features is much greater than the number of samples. Nu SVC is a similar method to the ordinary SVC but accepts slightly different sets of parameters. Linear SVC is another implementation of Support Vector Classification for the case of a linear kernel. The implementation of both methods are based on different libraries.

| **Algorithm** | **accuracy** | **precision-pos** | **precision-neg** | **recall-pos** | **recall-neg** |
|---|---|---|---|---|---|
| | | | **Twitter** | | |
| MNB | 77.78 | 78.88 | 76.77 | 75.85 | 79.7 |
| BNB | 77.43 | 76.39 | 78.58 | 79.39 | 75.45 |
| LR | 78.25 | 77.24 | 79.31 | 80.05 | 76.45 |
| LSVC | 75.24 | 74.48 | 76.07 | 76.78 | 73.71 |
| NuSVC | 78.06 | 76.31 | 80.04 | 81.34 | 74.76 |
| | | | **Stocktwits** | | |
| MNB | 79.1 | 81.84 | 76.8 | 74.85 | 83.34 |
| BNB | 76.63 | 85.58 | 71.34 | 64.15 | 89.14 |
| LR | 79.59 | 79.4 | 79.82 | 79.97 | 79.22 |
| LSVC | 78.77 | 80.82 | 77.08 | 75.55 | 81.97 |
| NuSVC | 79.18 | 77.44 | 81.23 | 82.48 | 75.9 |
| | | | **News** | | |
| MNB | 83.26 | 82.97 | 83.58 | 83.72 | 82.8 |
| BNB | 83.3 | 84.04 | 82.6 | 82.18 | 84.42 |
| LR | 84.54 | 84.83 | 84.28 | 84.12 | 84.96 |
| LSVC | 84.33 | 84.83 | 83.92 | 83.68 | 84.98 |
| NuSVC | 86.23 | 87.15 | 85.38 | 85.0 | 87.46 |

Table 4.13: Accuracy of machine learning classifiers
MNB - Multinomial Naive Bayes
BNB - Bernoulli Naive Bayes
LR - Logistic Regression
LSCV - Linear Support Vector Classifier
NuSVC - Nu-Support Vector Classifier

## 4.5 Voted classifier - final classification performance

The classification performance can be improved by combining all the five above described classifiers into one. As proposed by Sentdex.com [**kinsley**], we can implement a new classifier which counts the results (votes) from all the five previously described algorithms. Based on a given confidence threshold, it makes its own decision. The usage of 5 voting algorithms gives us the possibility to choose three different confidence thresholds because the final outcome (positive/negative) can be calculated as a majority of 3, 4 or all 5 votes. This can be interpreted as confidence levels of 60%, 80% and 100%. Confidence of 60% represents a simple voting majority (3 of 5) which assigns a positive or negative outcome to each single analyzed input record and therefore it provides similar but slightly

improved outcomes compared to the performance of voting algorithms individually. On the other hand, the confidence levels of 80% and 100% exclude records which were classified with lower confidence and they thus assign sentiment value to just a part of the input records. The skipped data can be considered as unclassified or neutral. The expected benefit of higher confidence level is an increase of classification precision. As it was already mentioned, the shortcoming is a presence of unclassified records, however, this can be neglected in case of a high amount of input data.

There are two approaches to measure the performance metrics of a voted classifier. We can either measure the performance of the complete input data set including the records classified as neutral, or there is the possibility to exclude the neutral records from the testing set metrics. Table 4.14 shows the results for both of these methods. Outcomes measured when including neutral records are in the parentheses, cells without parentheses contain results which are the same for both classes. Such a case is for example a 60% confidence level which outputs only positive or negative records. Precision is also not affected by this problem, as its calculation focuses solely on the correctness of records labeled as positive or negative.

When looking at the outputs in the parentheses, we see the problem arising from including the neutral class in the calculation: accuracy and recall measures decline with the increasing vote level. This is because all the records classified as neutral are considered as a mistake. Both accuracy and recall significantly decrease with each single confidence step for all the data sources. The results without parentheses are therefore much easier to understand when comparing the performance. The Twitter and News classifiers encounter increase in all the metrics by approximately 3% per each step of a confidence level. In both cases, the negative and positive precisions grow together and keep in almost perfect balance. That is a very good result because we can rely on there being no tendency to output one result over another. Unlike the Twitter and News classifiers, the Stocktwits classifier does not provide a very balanced outcome. With an increasing confidence level, the positive precision outgrows the negative one by approximately 10%, which causes bias in the classification outcome. In the case of Stocktwits, we can hence anticipate prediction problems for other than 60% confidence level.

Table 4.15 shows the percentage of records which were classified as neutral when using 80% and 100% confidence level. As expected, the number increases with higher confidence. In thea case of the Stocktwits classifier, we reach a very high number of 42.3% of neutral records, which means that we ignore almost half of the total input records. When taking into consideration the low daily amount of Stocktwits input data, this again confirms the previously made assumptions about problems with other than 60% confidence level used for the Stocktwits classification. Since the Twitter and News data sources reach hundreds or thousands of input data every day, the presented percentage of ignored records should not cause any negative performance effect. The influence of confidence levels will be further analyzed in the following chapters.

| votes | accuracy | prec.-pos | prec.-neg | recall-pos | recall-neg |
|---|---|---|---|---|---|
| **Twitter** | | | | | |
| 60% | 78.53 | 77.6 | 79.53 | 80.21 | 76.86 |
| 80% | 81.23 (72.67) | 80.49 | 82.0 | 83.0 (74.95) | 79.42 (70.4) |
| 100% | 84.73 (64.21) | 84.23 | 85.33 | 86.3 (66.34) | 83.13 (62.08) |
| **Stocktwits** | | | | | |
| 60% | 80.15 | 81.78 | 78.75 | 77.64 | 82.62 |
| 80% | 84.54 (72.04) | 91.64 | 79.82 | 75.16 (63.01) | 93.48 (81.08) |
| 100% | 87.8 (60.32) | 94.08 | 84.28 | 76.86 (45.97) | 96.24 (74.67) |
| **News** | | | | | |
| 60% | 85.08 | 85.48 | 84.72 | 84.52 | 85.64 |
| 80% | 88.32 (80.58) | 88.64 | 88.06 | 88.12 (81.3) | 88.48 (79.86) |
| 100% | 90.3 (76.43) | 91.08 | 89.58 | 89.62 (76.76) | 91.0 (76.1) |

Table 4.14: Voted classifier performance - (results including neutral records are in parentheses)

| votes | neutral records [%] |
|---|---|
| **Twitter** | |
| 80% | 12.5 |
| 100% | 30.5 |
| **Stocktwits** | |
| 80% | 14.3 |
| 100% | 42.3 |
| **News** | |
| 80% | 8.9 |
| 100% | 17.2 |

Table 4.15: Percentage of records considered as neutral

## 4.6 Sentiment aggregation

Before measuring the correlation between stock and sentiment data, we have to define the time unit we are going to work with (e.g. hours, days, etc.). The sentiment analysis outcomes which occurred within the specified time window will be then aggregated to obtain one single sentiment value representing the given time unit. As it is shown in Figure 4.3, the aggregated sentiment value will be expressed as the percentage of positive records out of all the analyzed records. In the following paragraphs we will discuss the different possibilities of the time frame usage and choose the optimal one for our study.

The aggregation of a time window of sentiment should match the aggregation of stock market prices which we are going to predict. For instance, when we choose a time window of one hour, it will be used as a time unit for both sentiment movements and stock price movements.

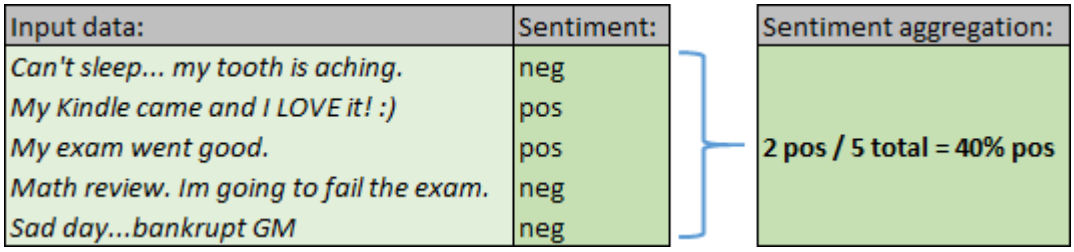


Figure 4.3: Aggregation of Twitter sentiment

### 4.6.1 Event driven prediction (no aggregation)

The shortest possible stock prices prediction uses a so called event-driven approach. In this method, we do not use any time window to aggregate the values but we simply make the stock movement prediction immediately when a significant event occurs. Such an approach is examined in the study of Schumaker and Chen which predicts up/down movements of stock prices 20 minutes after a release of a news article. [Sch+12] Obviously, this approach is suitable only for data sources where even a single post or article can make a big influence, as for example articles in famous world-wide media (NY Times, CNN) or tweets of influential people (politicians, CEOs). Since our data collection process does not follow any specific newspapers or Twitter accounts, this approach is not very suitable for our study. Furthermore, as we are not aware of any influential users of Stocktwits network, this method would be technically impossible to implement for this social platform.

### 4.6.2 One day aggregation

A very common form of sentiment aggregation is using the time frame of one day. Unlike in the previous approach, in this case we do not rely on any significant events but we examine "the power of the crowd". Since all our three chosen platforms provide number of input data on daily basis, it should not make any problems to implement this approach. In such a case, we can watch the predicted stock price change in the movements of daily closing prices. An example of the daily aggregated sentiment data for Tesla company is shown in Table 4.4.

### 4.6.3 One week aggregation

The idea behind this rather uncommon method is very similar to the previously described one- day aggregation but with much longer time window. The main advantage is that using one week as a time unit enables us to concentrate on longer term predictions which can be more interesting for some investors. On the other hand, this method bears the risk of "hiding" short-time deviations and therefore loosing prediction accuracy. Another disadvantage which disables us from using this method is the long time series of input data needed to obtain enough weeks for a trustworthy analysis.

| Date | Pos. sentiment [%] |
|---|---|
| 2017-05-01 | 48.58 |
| 2017-05-02 | 45.91 |
| 2017-05-03 | 34.27 |
| 2017-05-04 | 28.26 |
| 2017-05-05 | 48.11 |
| 2017-05-06 | 52.18 |
| 2017-05-07 | 50.02 |
| 2017-05-08 | 46.26 |
| ... | ... |

Figure 4.4: Example of sentiment aggregated per day (Tesla, Stocktwits)

Given the mentioned arguments, it was decided to use the one-day time frame for sentiment data aggregation in this study. Daily stock price values will be represented by the trading day closing prices.

## 4.7 Day-off handling

The sentiment analysis aggregation provides daily values of a positive sentiment percentage. However, the fact that e.g. on Tuesday the positive sentiment reached 70% gives us no useful information. What is important, is the daily change from one day to another, e.g. from Monday to Tuesday the positive sentiment increased by 20%. The same operation can be performed on the obtained stock market data. This gives us two time series of day-to-day differences (sentiment data and stock data). The only problem is the discrepancy in the lengths of both series. While the sentiment values are obtained for each single day, the stock data are missing weekends and bank holidays of the respective country, as it is shown in Figure 4.5. In the following paragraph we define three methods of handling this issue. They will be further referenced as "Differentiation" methods.

| day of week | Mon | Tue | Wed | Thu | Fri | Sat | Sun | Mon |
|---|---|---|---|---|---|---|---|---|
| date (May 2017) | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th |
| stock price | 322 | 318 | 311 | 295 | 308 | N/A | | 307 |
| calculation | - | 318-322 | 311-318 | 295-311 | 308-295 | - | - | 307-308 |
| stock change | - | -5 | -7 | -16 | +13 | - | - | -2 |

Figure 4.5: Stock price differentiation

### Trading days difference (Skipped)

Probably the easiest approach to handle the day-offs issue is to take into account only the sentiment values measured during stock market trading days. By doing so we ignore all the mood changes which occurred during weekends or public holidays. The shortcoming of this approach can arise e.g. when there happens a major negative event on Friday followed by heated on-line discussions during Saturday and Sunday. However, while the interest in the topic slowly decreases reaching almost its neutral value on Sunday evening, the investors' reaction on Monday morning can be still remarkable. An example of such a sentiment change calculation is shown in Figure 4.6. This approach is further referenced as 'Skipped'.

| | | | | | | ignored | | |
|---|---|---|---|---|---|---|---|---|
| day of week | Mon | Tue | Wed | Thu | Fri | Sat | Sun | Mon |
| date (May 2017) | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th |
| pos. sentiment [%] | 48 | 45 | 34 | 28 | 48 | 52 | 50 | 46 |
| calculation | - | 45-48 | 34-45 | 28-34 | 48-28 | - | - | 46-48 |
| sentiment change | - | -3 | -11 | -6 | +20 | - | - | -2 |

Figure 4.6: Skipped Differentiation method

### Calendar days difference (Natural)

In contrast to the previous approach, there is the possibility to naturally calculate the difference of every two consecutive calendar days. In this case, the sentiment shift on Monday is calculated by subtracting the Sunday's value, as it is illustrated in Figure 4.7. Using this method, we take into account the sentiment of the previous day, no matter whether it was trading day or not. The results obtained for non-trading days can be used in case a predictor focuses on more than one previous day. For example, if a predictor makes a forecast based on three previous days, the price movement on Monday is predicted based on the sentiment values measured on Friday, Saturday and Sunday. This approach is further referenced as 'Natural'.

### Averaged weekends difference (Averaged)

The last introduced method is an aggregation over the day-off periods. As it is shown in Table 4.8, this approach adds the sentiment values of non-trading days to the nearest previous working day and calculates the mean value. After that, the same steps as in Skipped approach are applied. This way all the measured sentiments are utilized and the problem of skipping weekends can be mitigated. The disadvantage of the mean value can appear when there are two contradictory events happening shortly one after another on a weekend. It can be presumed that the stock movements of the following days are rather influenced by the later affair while the previous one is already forgotten, nonetheless, the

| | | | | | | ignored | | |
|---|---|---|---|---|---|---|---|---|
| day of week | Mon | Tue | Wed | Thu | Fri | Sat | Sun | Mon |
| date (May 2017) | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th |
| pos. sentiment [%] | 48 | 45 | 34 | 28 | 48 | 52 | 50 | 46 |
| calculation | - | 45-48 | 34-45 | 28-34 | 48-28 | - | - | 46-50 |
| sentiment change | - | -3 | -11 | -6 | +20 | - | - | -2 |

Figure 4.7: Natural Differentiation method

mean value will erase the significance of both events. This approach is further referenced as 'Averaged'.

| day of week | Mon | Tue | Wed | Thu | Fri | Sat | Sun | Mon |
|---|---|---|---|---|---|---|---|---|
| date (May 2017) | 1st | 2nd | 3rd | 4th | 5th | 6th | 7th | 8th |
| pos. sentiment [%] | 48 | 45 | 34 | 28 | 48 | 52 | 50 | 46 |
| | | | | | average = 50 | | | |
| calculation | - | 45-48 | 34-45 | 28-34 | 48-28 | - | - | 46-50 |
| sentiment change | - | -3 | -11 | -6 | +20 | - | - | -2 |

Figure 4.8: Averaged Differentiation method

CHAPTER 5

# Correlation

In this chapter we will analyze the movements of the previously obtained stock and sentiment time series in order to measure their correlation. We aim to find out whether the sentiment changes cause similar shifts of the stock prices. The following sections describe the methods used for the measurement and present the obtained outcomes.

## 5.1 Granger causality

For measuring the correlation between calculated sentiment data and downloaded stock data, we used a common statistical method called Granger causality. The same approach was used in the studies of Bollen et al [BMZ11] or Mittal and Goel [MG12]. The Granger causality test is a statistical hypothesis test for determining whether one time series is useful in forecasting another. According to Granger causality, if a signal X1 "Granger-causes" (or "G-causes") a signal X2, then past values of X1 should contain information that helps predict X2 above and beyond the information contained in past values of X2 alone. Its mathematical formulation is based on linear regression modeling of stochastic processes [Set].

### Null and alternate hypotheses

As was previously mentioned, Granger causality is a statistical hypothesis test, therefore it is necessary to state the null and alternate hypotheses. The standard approach is a so called "bottom-up" procedure, where the assumption is that the analyzed time series are independent variables. Then the data sets are analyzed to see if they are correlated in case of rejecting the null hypotheses. In order to implement the described process we state the following null and alternate hypotheses:

**Null hypothesis:** Sentiment time series does not Granger-cause stock time series
**Alternate hypothesis:** Sentiment time series does Granger-cause stock time series

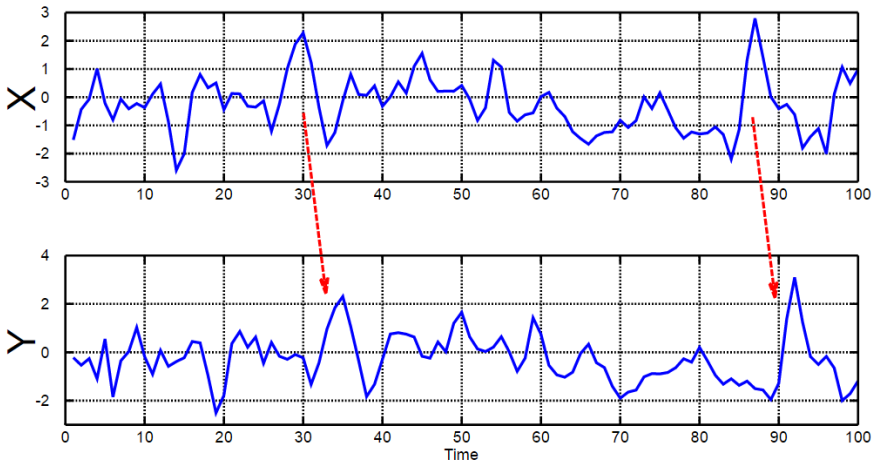


Figure 5.1: Granger causality visualization [17b].

### Choice of lags

When measuring the causality of two time series, it is necessary to specify the expected time lags denoting the "shift" between the data, as it is shown in Figure 5.1. Our time series consist of aggregated daily values, therefore, in our case one lag represents one day. According to the knowledge obtained from the existing literature, the changes in public sentiment can influence the stock values of several following days but usually not more than 3 days. [BMZ11] Our test will therefore include lags from 1 to 3 days.

### Alpha value (Significance level)

After executing the test with the details mentioned above, we obtain a p-value which tells us the probability of successfully rejecting the null hypothesis. In statistics it became a common habit to use 5% as a significant threshold which we will use as well.

## 5.2 Granger causality results

Since our data are already transformed into stationary time series of the same length (Sections 4.6 and 4.7) we do not need to do any further transformations and we can start with the Granger causality analysis. For each single test we have to provide two time series together with desired number of lags for which we expect one series to “Granger-cause” the other. The lag defines the number of days we expect to pass between the measured change of sentiment in media and the predicted price movement on the stock market.

The outcome of the Granger causality test is a p-value which states the probability of incorrectly rejecting the null hypotheses. Therefore, the lower the p-value, the higher the causality of the measured series. In our case, the p-value needs to be under 5% to be considered as significant.

The number of different Granger causality tests we have to perform for each single entity is given by the number of various sentiment analysis methods and precisions which were defined in the previous chapters. They include three Differentiation methods (Natural, Skipped, Averaged) and three possible precisions of a voted classifier (60%, 80%, 100%), all presented in Chapter 3. Combining these two variables results in 9 different time series of sentiment values. Each of these series then needs to be tested for three lags. All in all, we have three variables, each consisting of three values. This gives us 27 different combinations of the Granger causality tests for each single entity.

In the three following sections we will individually analyze the results of Twitter, Stocktwits and News articles. At first, we will present all the significant results and show them in tables summarizing outcomes obtained for all the possible combinations of lag, Differentiation method and precision. Then we will try to identify whether there exist any common features which apply to the given data source as a whole. For this purpose, we will calculate average p-values for lag, Differentiation method and precision using all the tested entities. Such an approach will allow us to compare the Granger causality results in a global aspect and identify the best performing setup.

### 5.2.1 Twitter

**Significant causality (>95%)**

Out of 9 observed entities (6 companies and 3 stock indexes) there was obtained at least one significant p-value in 4 of them. Table 5.1 summarizes results of all the tests performed for the Microsoft company. As it can be seen, the only significant result comes from the measurement combination of Averaged diff. method, 60% precision of sentiment analyzer and two lags shift between the time series. One significant p-value out of 27 test is not a very convincing proof, on the other hand, many of the other results are still very close to the border of 5%. For example, all the numbers obtained for the 1st and 2nd lag of the "averaged" method still keep under 0.1. The correlation of both stock and sentiment curves measured by the best performing combination is shown in Figure 5.1. The sentiment curve (blue) is shifted two days to the right in order to prove the causality results telling us that 2 is the most precise lag value.

| diff. method: | Natural | | | Skipped | | | Averaged | | |
|---|---|---|---|---|---|---|---|---|---|
| Lag / Prec. | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 60% | 0.15 | 0.40 | 0.98 | 0.09 | 0.08 | 0.18 | 0.07 | **0.039** | 0.10 |
| 80% | 0.08 | 0.30 | 0.91 | 0.08 | 0.13 | 0.24 | 0.053 | 0.052 | 0.13 |
| 100% | 0.08 | 0.27 | 0.84 | 0.12 | 0.14 | 0.27 | 0.08 | 0.06 | 0.14 |

Table 5.1: Twitter - Microsoft causality

The Netflix causality results from Table 5.2 are in some aspects very similar to the previous outcomes of Microsoft. Also here, the Averaged diff. method clearly outperforms the other two and provides only one significant value. Unlike the previous results where both 1st and 2nd lag resulted in similarly good values, here the 1st lag achieves a clearly higher correlation in all the three precision levels. Another similarity to the Microsoft results is the fact that in both cases the increase of precision does not improve the correlation but also does not show any clear traits of decreasing the performance. The results distribution over different precision levels seems to be purely random.

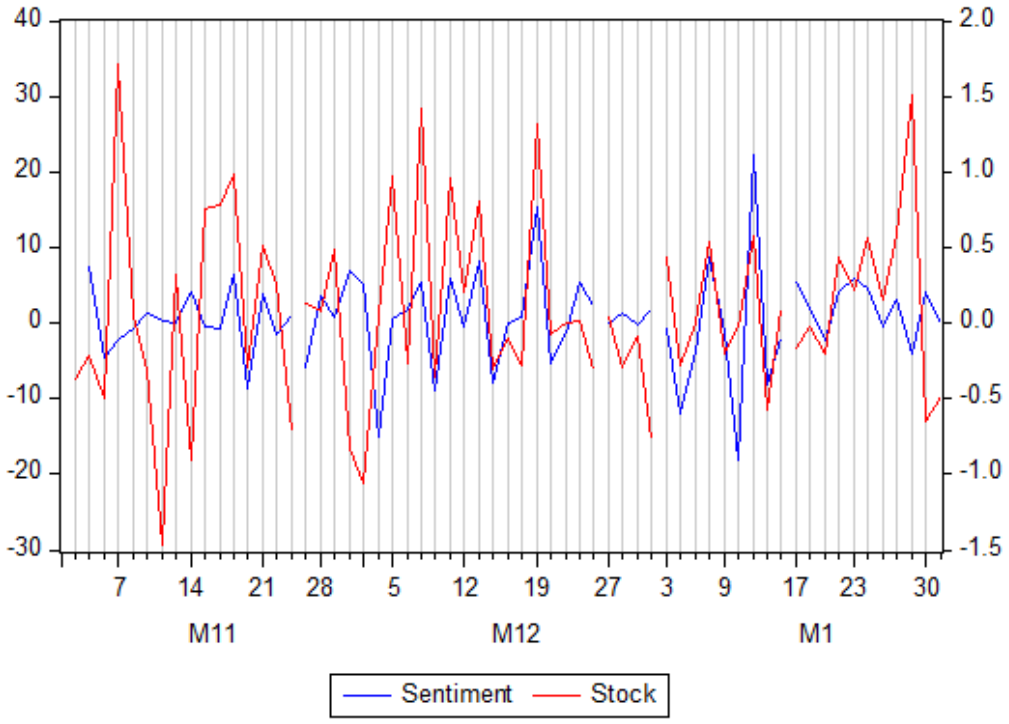


Figure 5.2: Twitter - Microsoft: Overlapping curves of stock and sentiment values lagged by two days. (Blue curve shifted 2 points to the right)

While Microsoft and Netflix encountered a significant correlation for the average Differentiation method, both DJIA and SPX stock indexes show values very far from significant in this sector. On the other hand, they both perform very well in case of Natural diff. method. This method also shows a clearly positive influence of higher sentiment analyzer precision. The only difference between both indexes is the significant lag. While DJIA correlates with the distance of only one lag, SPX shows the lowest p-values for the third lag. Their outcomes are summarized in Tables 5.3 and 5.4. The SPX correlation between its stock value and the sentiment curve is visualized in Figure 5.3.

| diff. method: | Natural | | | Skipped | | | Averaged | | |
|---|---|---|---|---|---|---|---|---|---|
| Lag: / Prec.: | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 60% | 0.17 | 0.60 | 0.72 | 0.12 | 0.10 | 0.21 | 0.07 | 0.15 | 0.30 |
| 80% | 0.21 | 0.62 | 0.82 | 0.09 | 0.09 | 0.19 | **0.043** | 0.11 | 0.24 |
| 100% | 0.34 | 0.87 | 0.97 | 0.12 | 0.16 | 0.33 | 0.07 | 0.20 | 0.38 |

Table 5.2: Twitter - Netflix causality

| diff. method: | Natural | | | Skipped | | | Averaged | | |
|---|---|---|---|---|---|---|---|---|---|
| Lag / Prec. | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 60% | 0.09 | 0.50 | 0.23 | 0.59 | 0.58 | 0.75 | 0.98 | 0.94 | 0.95 |
| 80% | **0.044** | 0.45 | 0.18 | 0.38 | 0.69 | 0.69 | 0.75 | 0.79 | 0.91 |
| 100% | **0.021** | 0.41 | 0.13 | 0.38 | 0.69 | 0.61 | 0.82 | 0.82 | 0.94 |

Table 5.3: Twitter - DJIA causality

| diff. method: | Natural | | | Skipped | | | Averaged | | |
|---|---|---|---|---|---|---|---|---|---|
| Lag / Prec. | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 60% | 0.20 | 0.11 | **0.026** | 0.55 | 0.68 | 0.71 | 0.88 | 1.00 | 1.00 |
| 80% | 0.14 | 0.09 | **0.016** | 0.76 | 0.71 | 0.77 | 0.86 | 0.41 | 0.45 |
| 100% | 0.07 | 0.08 | **0.010** | 0.52 | 0.46 | 0.61 | 1.00 | 0.91 | 0.77 |

Table 5.4: Twitter - SPX causality

#### Not significant causality (<95%)

We did not obtain any significant causality for the companies Coca-Cola, Nike, Samsung, Tesla and stock index Nasdaq. The closest to significant outcomes were retrieved for Coca-Cola which showed p-values under 0.1 for the highest precision and the 1st lag in all the three Differentiation methods. This company also showed results with a clearly positive influence of higher sentiment analysis precision. The other entities show neither significant results, nor a clear influence of sentiment analysis precision. In the case of Nike, Tesla and Nasdaq, slightly better results are noticeable for the 1st lag in all the three methods and precisions.

#### Common features

Table 5.5 shows the average p-values for all the Twitter outcomes. From the values we can identify that with increasing lags, also the p-value increases. This tells us that on

average the most precise prediction can be obtained when forecasting only one day in advance. The exactly opposite effect is visible in the last row of the table which shows a positive influence of higher sentiment analyzer precision. The voted classifier is therefore proved to be useful for the Twitter data source. The middle row contains the average of diff. methods. Unlike the previous two variables, this one shows almost identical results for all the three values. This leads us to conclusion that for the Twitter data source, there cannot be identified any single diff. method more suitable than the others.

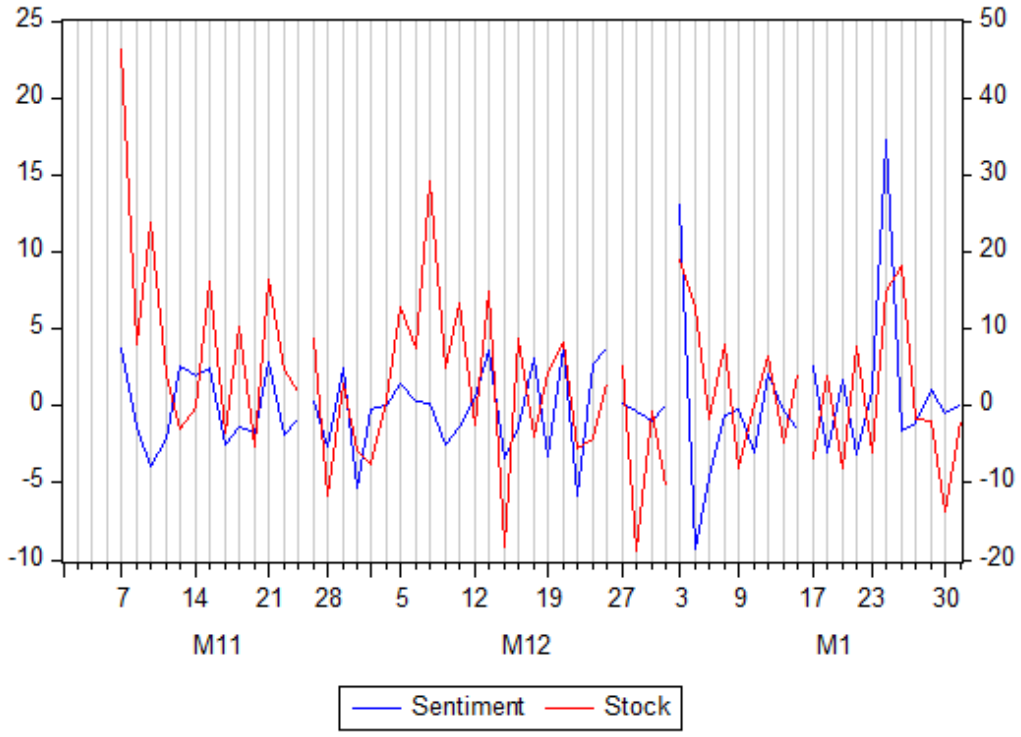


Figure 5.3: Twitter - SPX: Overlapping curves of stock and sentiment values lagged by two days. (Blue curve shifted 3 points to the right)

| **Lag:** | 1 | 2 | 3 |
|---|---|---|---|
| Total average: | **0.44** | 0.54 | 0.61 |
| **Diff. method:** | Natural | Skipped | Averaged |
| Total average | 0.528 | 0.524 | 0.535 |
| **Precision:** | 60% | 80% | 100% |
| Total average | 0.57 | 0.53 | **0.49** |

Table 5.5: Twitter: average p-values per variable

### 5.2.2 Stocktwits

#### Significant causality (>95%)

Out of 10 observed entities a significant p-value was found in only 2 cases, Microsoft and Tesla. Microsoft provided significant results already in the previous Twitter analysis, however, in case of the Stocktwits data the causality reaches much more significant outcomes. As we can see in Table 5.6, all the three lags of Skipped and Averaged diff. methods show p-values lower than 0.001 which is usually considered as "very significant causality". Movements of both time series are plotted in Figure 5.4 where the Sentiment curve was again shifted by 1 day to the right.

The second significant outcomes were obtained from Tesla. This company showed some close to significant results already in the Twitter analysis but the border of p-value 0.05 was crossed only for the Stocktwits data. Similarly to Microsoft, also Tesla reaches the highest correlation for Skipped and Averaged diff. methods but not for all the three lags. Tesla shows clear tendency of an increasing p-value with a higher lag. The best outcomes were therefore obtained for the lag value 1, which is also plotted in Figure 5.5 by shifting blue curve one day to the right.

| diff. method: | Natural | | | Skipped | | | Averaged | | |
|---|---|---|---|---|---|---|---|---|---|
| Lag / Prec. | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 60% | 0.11 | 0.12 | 0.20 | **.0003** | **.0001** | **.0001** | **.0007** | **.0002** | **.0002** |
| 80% | 0.92 | 0.59 | 0.98 | 0.08 | 0.23 | 0.16 | 0.09 | 0.18 | 0.24 |
| 100% | 0.27 | 0.58 | 0.45 | 0.04 | 0.16 | 0.29 | 0.07 | 0.20 | 0.38 |

Table 5.6: Stocktwits - Microsoft causality

| diff. method: | Natural | | | Skipped | | | Averaged | | |
|---|---|---|---|---|---|---|---|---|---|
| Lag / Prec. | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 60% | 0.52 | 0.93 | 0.94 | **0.013** | **0.047** | 0.13 | **0.046** | 0.14 | 0.22 |
| 80% | 0.77 | 0.98 | 0.46 | 0.06 | 0.17 | 0.33 | 0.12 | 0.30 | 0.32 |
| 100% | 0.60 | 0.96 | 0.63 | 0.12 | 0.32 | 0.56 | 0.21 | 0.31 | 0.38 |

Table 5.7: Stocktwits - Tesla causality

#### Common features

The average values from Table 5.8 calculated for all the nine entities only reflect the outputs of both significant results. The p-value depending on the lag shows direct proportion and increases together with a higher lag. The best performing diff. method was shown to be the Skipped method. This correlates with the assumption which was

made earlier in Chapter 3 about the expected drop of accuracy when using the Stocktwits data obtained during weekends.

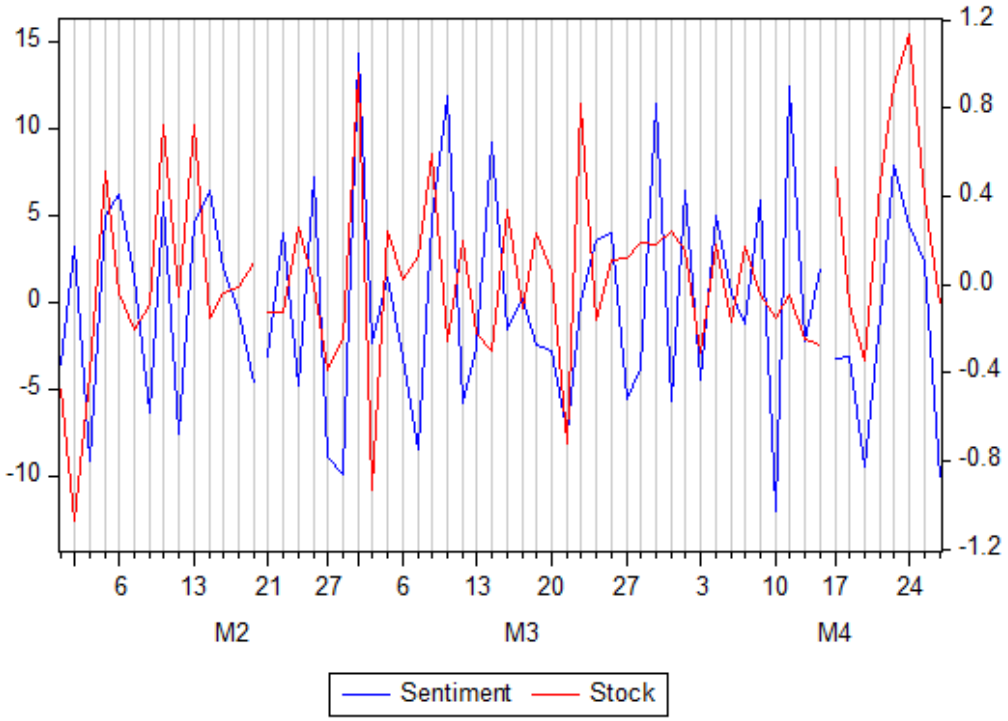


Figure 5.4: Stocktwits - Microsoft: Overlapping curves of stock and sentiment values lagged by two days. (Blue curve shifted 1 point to the right)

| **Lag:** | 1 | 2 | 3 |
|---|---|---|---|
| Total average: | **0.45** | 0.48 | 0.59 |
| **Diff. method:** | natural | skipped | averaged |
| Total average | 0.60 | **0.42** | 0.50 |
| **Precision:** | 60% | 80% | 100% |
| Total average | **0.46** | 0.54 | 0.52 |

Table 5.8: Stocktwits: average p-values per variable

### 5.2.3 News

Out of all the observed entities and all the variables combinations, there was not any significant p-value for the causality of the News sentiment. The closest to significant outcomes were obtained for McDonald's which showed 0.07 p-value for both Skipped

and Averaged diff. methods using 2 lags and 60% precision. However, as the 0.05 border was not crossed, we still cannot reject the null hypotheses. When taking a look at the average p-values summarized in Table 5.9, we can notice that the correlation tends to decrease with a higher lag and increase with higher precision. Such results correspond to those obtained for the Twitter data source. However, since we have not obtained any significant results for any entity, the relevancy of the measured values is not guaranteed.

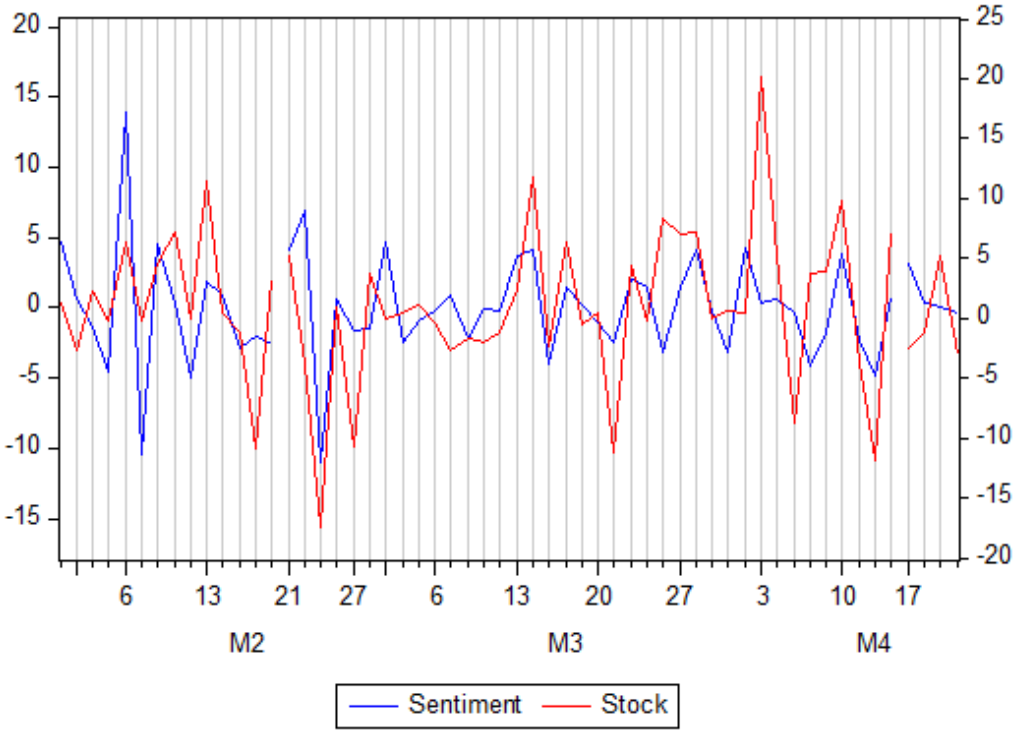


Figure 5.5: Stocktwits - Tesla: Overlapping curves of stock and sentiment values lagged by two days. (Blue curve shifted 1 point to the right)

| **Lag:** | 1 | 2 | 3 |
|---|---|---|---|
| Total average: | 0.52 | 0.58 | 0.61 |
| **Diff. method:** | Natural | Skipped | Averaged |
| Total average | 0.58 | 0.58 | 0.54 |
| **Precision:** | 60% | 80% | 100% |
| Total average | 0.59 | 0.57 | 0.54 |

Table 5.9: News: average p-values per variable

## 5.3 Comparison with real events

### 5.3.1 Twitter

Very interesting results were obtained for the McDonald's Granger causality and Twitter sentiment. Both Skipped and Averaged diff. methods exhibit very significant correlation results, especially in the 2nd lag where both methods reach p-values under 0.01. However, after short research of the McDonald's recent history, it was found out that these results are biased.

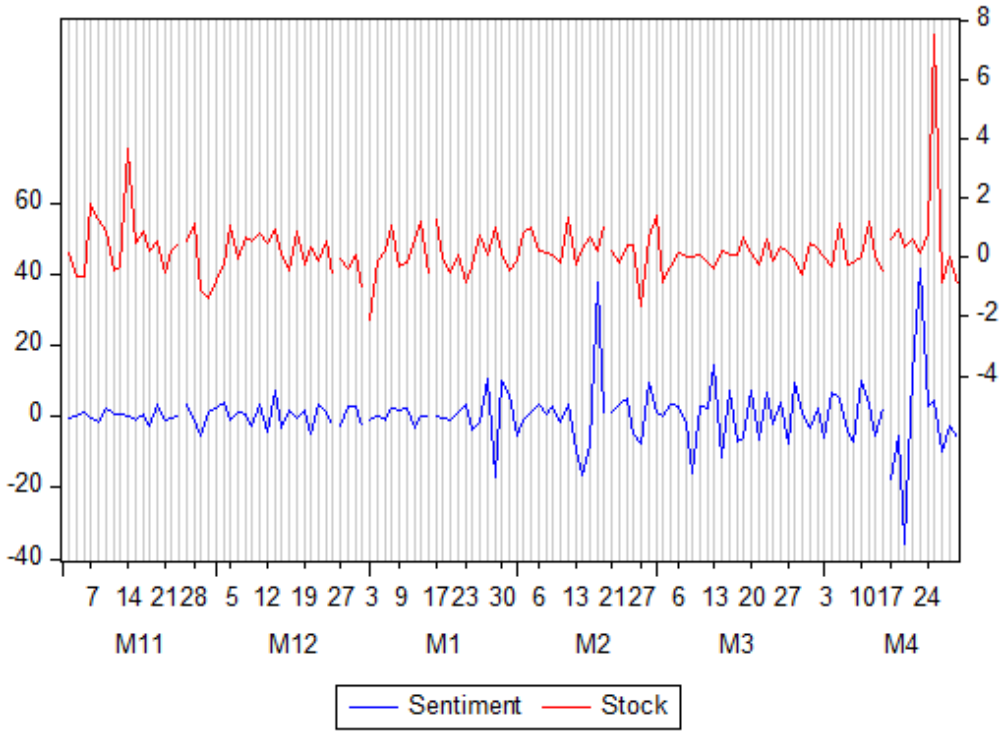


Figure 5.6: McDonald's: Comparison of stock and sentiment curves (no shifting)

As we can see in Figure 5.6, the correlation between both curves is supported by a very noticeable swing in the end of April. On the first sight, it looks like the increase of McDonald's share price was perfectly detected four days in advance by a sentiment analyzer but in reality, it is just an accidental occurrence of two important events within a short period of time. On the 21st of April the fast food chain unveiled pictures of its new staff uniforms, which immediately raised high public attention because of its surprising gray-scale design. The message got extremely viral on social networks where people started to share the content in various forms of jokes referring to the visual similarity of the new gray design to the one of the Star Wars imperial uniforms. Such

jokes were obviously interpreted by the sentiment analyzer as mostly positive, which caused a rapid growth of the sentiment curve. Four days later, on the 25th of April, the McDonald's company published its report of earnings for the first quarter which were globally accepted as very positive. This caused an immediate increase of the McDonald's stock price.

Given the fact that 21st of April 2017 was Friday and the 25th of April was Tuesday there was a weekend between the events which made it possible for Skipped and Averaged methods to count two days as one and influence the causality results. Natural diff. method counts each single day of a weekend without any aggregation and therefore the results of this method remained insignificant. Thus, the very interesting results of McDonald's causality are biased and there is actually no significant correlation after removing the high deviation events.

For other analyzed companies, we identified no such relation between the sentiment change and a real event. The majority of company related tweets seem to be independent of business oriented events, as for example release of quarterly reports or similar events which are important for investors.

### 5.3.2 Stocktwits

Unlike for Twitter, the posts in Stocktwits social network are clearly very dependent on the major business related events connected with the given company. As an example we can present the sharp sentiment and stock price movements which followed the release of Tesla's fourth quarter report of the year 2016.

The report was published on the 22nd of February 2017. Given the company's grand ambitions and billions in capital investments required to achieve them, the losses were expected. However, the main spotlight remained on the production of the long-awaited Model 3. Although the company assured the public that the delivery will meet the schedule, according to the Stocktwits posts listed below, it seems like many people have doubts about it. Further distrust in Tesla was also spread by announcing the departure of the current Chief Financial Officer. The following bullet list shows some examples of the skeptical Stocktwits posts:

- "$TSLA Model 3 will be delayed"
- "$TSLA Looks like actual deliveries of Model 3 will be in Q1 2018"
- "$TSLA thats bad news if the CFO is leaving."
- "$TSLA How much longer can Musk keep this Ponzi scheme running"
- "$TSLA 250 by Friday... wider loss, burning cash like crazy, CFO has alot to say at CC. Model 3 late again..."

- "$TSLA CFO is leaving because it's getting too hard to keep hiding those overwhelming losses..."
- "$TSLA Bad numbers, Not an impressive CC, Elon sounds depressed, CFO Leaving, Downgrades to come..."

Figure 5.7 shows the movements of Tesla's stock price and sentiment changes. As we can see, the skepticism which appeared on the social network right after the release of the quarterly report was followed by a sharp decline of Tesla's stock price. Therefore, in this case we can say that the sentiment was correctly preceding the stock market reaction.

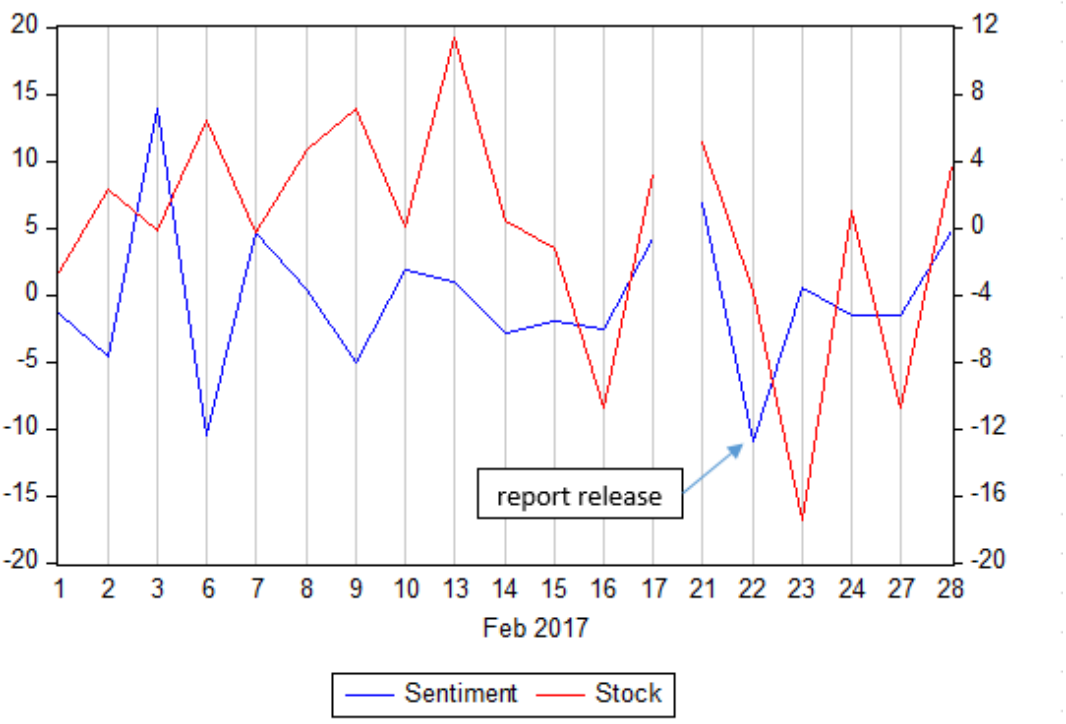


Figure 5.7: Tesla: Comparison of stock and sentiment curves after Q4 report release (no shifting)

CHAPTER 6

# Prediction

In this chapter we are going to investigate the possibilities of predicting stock changes based on sentiment changes. As it was already defined in the previous chapter, we assume that the stock price at a given day is not influenced by sentiment changes older than three previous days. While the Granger causality test was measuring the influence of all the three days individually, the predictor will take the three days together as one input record and let the classification algorithm learn their importance by itself. In the following sections, we will describe various possibilities of input data manipulation and evaluate the prediction results.

## 6.1 Input data preparation

Since the beginning of October 2016, when the input data collection started, until the end of August 2017, there were 220 working days (all of our companies and indexes are located in the US). As we work with day to day differences in both sentiment and stock values, it gives us a series of 219 daily changes. These need to be further aggregated into sets of three consecutive sentiment changes and one stock change (four days in one record), as it is depicted in Figure 6.1. Hence, in total we obtain 216 records which are ready to be used as an input for machine learning algorithms.

### 6.1.1 Records arrangement

The smaller is the amount of input records for machine learning, the higher is the probability of obtaining biased results. As the number of our input records is rather small, we have to introduce some further measures in order to minimize the thread of biased results. As an example of input data causing wrong classification tendencies, we could mention the inequality of positive vs. negative stock price changes. If a company's stock price was rather growing over the whole period of our measurement, the machine

| day | 1 | 2 | 3 | 4 | 5 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| sentiment change | up | down | down | up | down | up | up |
| stock change | down | up | down | down | down | down | up |
| | | | | | | | |
| record 1 | up | down | down | down | | | |
| record 2 | | down | down | up | down | | |
| record 3 | | | down | up | down | down | |
| record 4 | | | | up | down | up | up |

Figure 6.1: Aggregation of input data into records of four consecutive days

learning algorithm can learn that the probability of growth is higher than the opposite. As a result, we might obtain good prediction results just because both training and testing sets used data from the growth period. This problem can be easily solved by a manual preparation of both training and testing sets with an equal number of positive and negative records. In other words, if a company has e.g. 90 negative records and 116 positive records, then in each single machine learning run we use only 180 (90 positive and 90 negative) records. In order to obtain even more accurate results, we cycle each company measurement 50 times and before each of the cycles we shuffle the records to make sure that all of the total 216 records are used in various training and testing positions.

### 6.1.2 Sentiment binning

Binning is a very common data preprocessing method used in machine learning. Its purpose is to divide continuous input into discrete values which can be easily processed by classification algorithms. Our input values are daily changes of sentiment expressed in percentage. The range of the input variables therefore goes from -100 to +100. Even when ignoring the decimal numbers, 200 classes are still too many for a classification of only 216 records. As a result, the following three binning methods were introduced:

1. **2-classes binning:** Divides the input into only 2 classes around the middle value of 0. All the input values lower than 0 are considered as negative while the rest of the values are considered as positive. There is no neutral class present here. The same approach is used for all the stock data.

2. **3-classes binning:** Divides the input into 3 classes: Negative, Neutral and Positive. The aim is to obtain 3 classes of an approximately same size. The problem is that we cannot use any firm threshold value of a sentiment change which would suit for all the companies. While some companies have a very high dispersion of the sentiment, other keep the measured daily changes very close to the middle point. Hence, the best approach here is the usage of standard deviation. The optimal

threshold value for dividing the input into 3 classes was measured to be 1/4 of standard deviation.

3. **5-classes binning:** Divides the input into 5 classes. The approach here is the same as in the low binning, just with the difference that we add one more threshold to obtain 5 classes. The thresholds are created using one standard deviation and 1/4 of standard deviation.

## 6.2 Machine learning

The core step of the whole prediction process is machine learning classification. This step uses the preprocessed data to train the classifier and to obtain the final prediction accuracy. Given the small number of input records, we are forced to abandon the usage of some powerful but data demanding tools, as for example lately very popular neural networks. Instead, we will use the same algorithms which were previously used for the sentiment analysis, mostly because of its simplicity and high efficiency. The set of algorithms comprises of the following classifiers:

- Multinomial Naive Bayes
- Bernoulli Naive Bayes
- Logistic Regression
- Linear Support Vector Classifier
- Nu-Support Vector Classifier
- Voted Classifier (combination of all above)

Each prediction run calculates the accuracy, precision, recall and the order of the most informative features. The calculation of the most informative features tells us which one of the three input days was the most valuable one for making the prediction. The final values presented later are calculated out of 50 independent training runs. All the accuracy, precision and recall values are therefore mean values of all the 50 results. In the case of the most informative features, we cannot calculate the mean value so we use a special scoring system: Every training run assigns 3 points to the first most informative feature, 2 points to the second one and 1 point to the least informative one. These points are later summed up for all the 50 cycles so the highest number represents the highest importance.

## 6.3 Margin of error

Since the prediction results are calculated as the mean values of 50 cycles, it is necessary to define the margin of error for the sample mean. It will help us outline a border between

outcomes with significant prediction power and outcomes too close to 50% accuracy. The general formula for the margin of error for the sample mean is shown below.

$$ME = z * \frac{\sigma}{\sqrt{n}} \tag{6.1}$$

The variables represent:

- ME - margin of error
- z* - value for desired level of confidence from the Table 6.1
- $\sigma$ - standard deviation of the sample
- n - sample size

| Confidence [%] | z*-value |
|---|---|
| 80 | 1.28 |
| 90 | 1.645 |
| 95 | 1.96 |
| 98 | 2.33 |
| 99 | 2.58 |

Table 6.1: z*-values for selected confidence levels

When running the prediction tests for various combinations of input data and preprocessing methods, we obtain a unique data sample in each run. Thus, in theory, we should calculate a unique margin of error for every single input data combination. However, as the standard deviations of all the obtained samples are mostly very close to each other, it makes no sense to define a specific margin of error for each sample because the difference in the calculated values would be very small. Since the vast majority of the calculated standard deviations lies in the range (7,8) we will simply use the value 8 for all the samples. We also want to make sure that the results are not just accidents and that they are really far from 50% accuracy so we will use the highest confidence level of 99%. When putting these values into Formula 6.1, we obtain the margin of error 2.94 (rounded to 3). This means that for the results presented later we consider as significant outcomes only such mean values which are higher than 53%.

## 6.4 Prediction results

After running the tests we are going to examine the prediction results measured for various combinations of the previously described preprocessing steps. Some of these input variables were analyzed already in Chapter 5, however, for example the comparison of differentiating methods (Natural, Skipping, Averaged) did not show any clear results suggesting to prioritize one method over another. As a result, all the three approaches

will be analyzed in the prediction test as well. On the other hand, the positive influence of the higher sentiment analysis precision (60%, 80%, 100%) was already proven to be sensitive to the amount of input data records. In other words, the high limit of voted classifier precision had a positive effect on the Twitter and News data sources but it showed the exactly opposite tendency for Stocktwits. Therefore, in the following analysis we will use 100% voted classifier precision for Twitter and News while Stocktwits will use only 60%. Except for the best performing differentiating method and the most informative day-lag there will be also presented the comparison of binning performance (2, 3 and 5 classes).

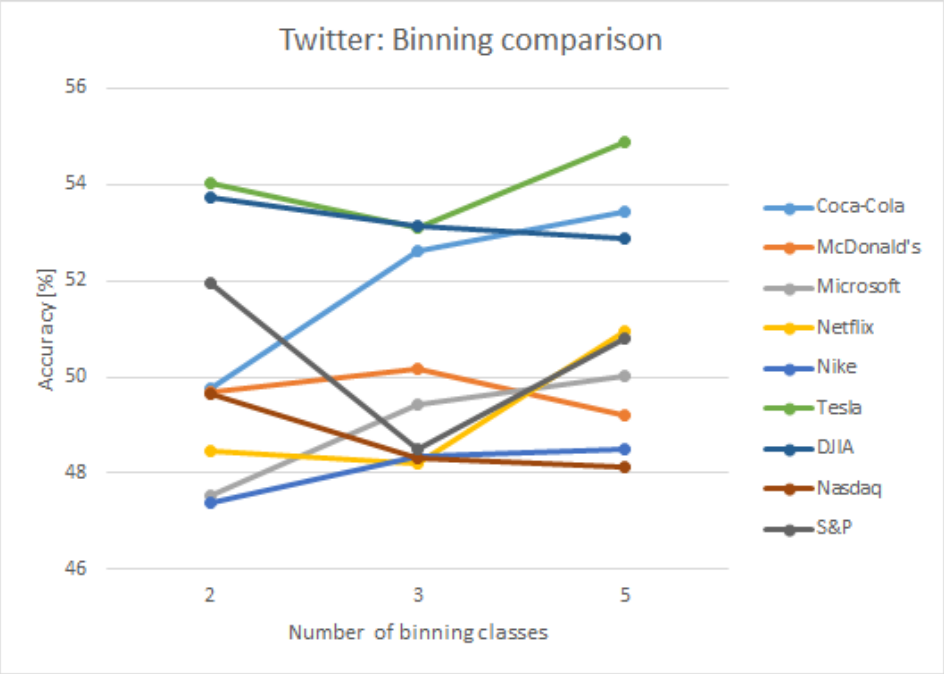


Figure 6.2: Twitter: Influence of input data binning on classification accuracy

### 6.4.1 Binning

At first, we are going to examine the influence of input data binning which was presented in Section 6.1.2. The fact that we have three Differentiation methods and we cannot use any of them as a default one, gives us also three prediction results for every single binning method. However, comparing nine results which represent three different groups of data would not be effective so the numbers presented in the following graphs show only average values for a given binning method.

Figure 6.2 shows the accuracy values obtained for the Twitter classifier. Although the chart visually seems to depict higher values for 5-classes binning, the improvement over 2-classes binning is very small and valid only for approximately half of the observed entities. Therefore, we cannot conclude any significant influence of input binning.

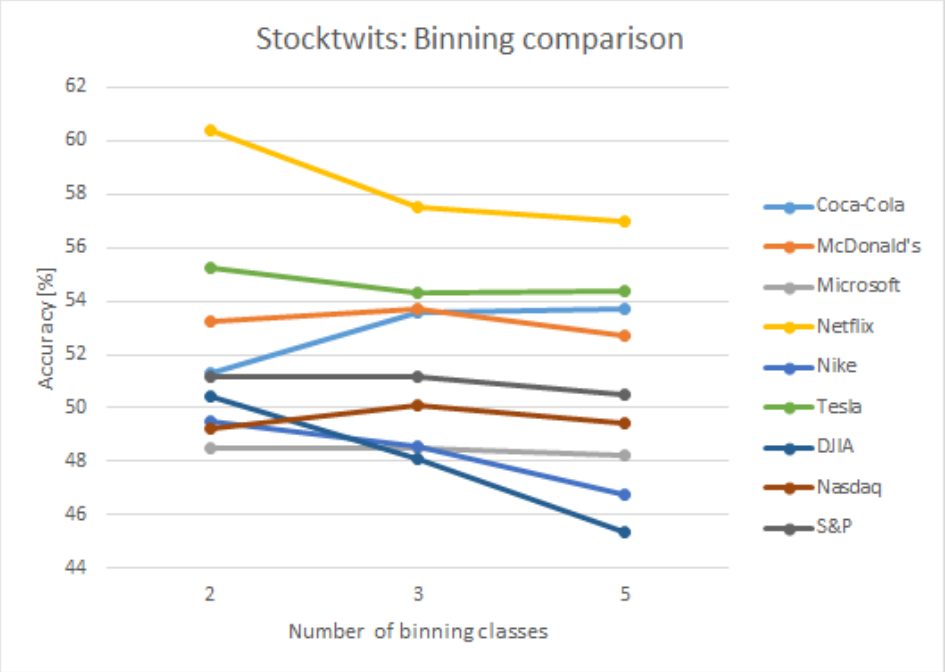


Figure 6.3: Stocktwits: Influence of input data binning on classification accuracy

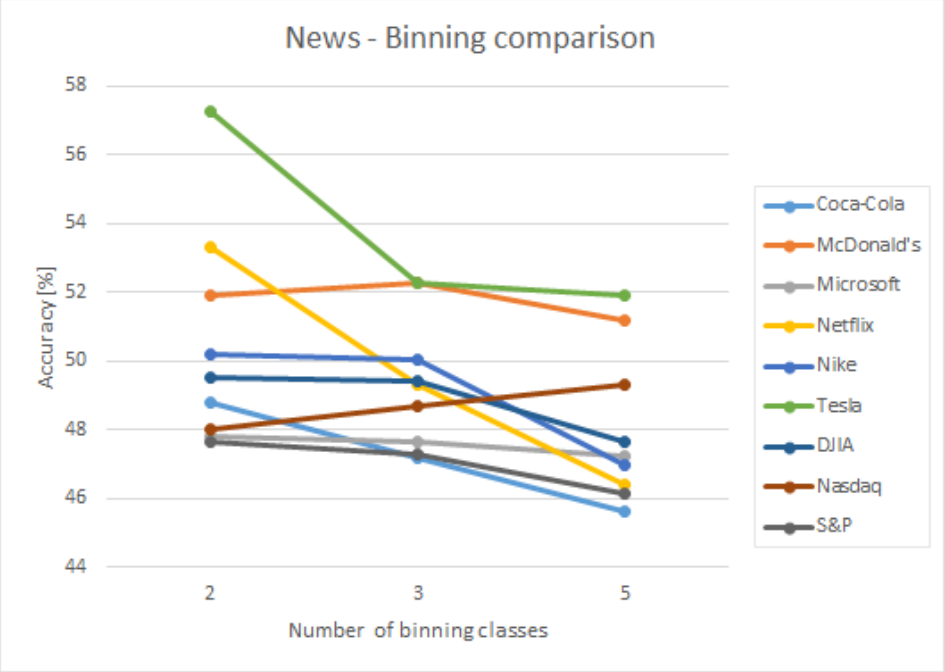


Figure 6.4: News: Influence of input data binning on classification accuracy

Even a less noticeable tendency is presented in Figure 6.3 which depicts the binning comparison for Stocktwits. Even though three of the lines exhibit a visible upwards or downwards direction, all the remaining measurements seem to be rather unresponsive to the binning class change. On the other hand, the binning dependency results measured for the News articles data source present very significant outcomes. As we can see in Figure 6.4, eight out of nine stock entities show a decrease of accuracy for higher numbers of binning classes.

In summary, the introduction of binning into classification provides no performance improvement for neither Twitter or Stocktwits data sources and in the case of News articles causes even a performance decline. Given these test results, we can conclude that the usage of more binning classes does not bring any benefit to our prediction model and in the further tests we will work only with the basic binary binning (2 classes).

### 6.4.2 Differentiation methods

In this section, we are going to investigate the dependency of our prediction model on the Differentiation methods which were defined in Chapter 4. Since we already know the most suitable binning method from the previous section, there is no need to calculate any average accuracy values, and we can simply use the Differentiation method outcomes which were obtained for 2-classes binning.

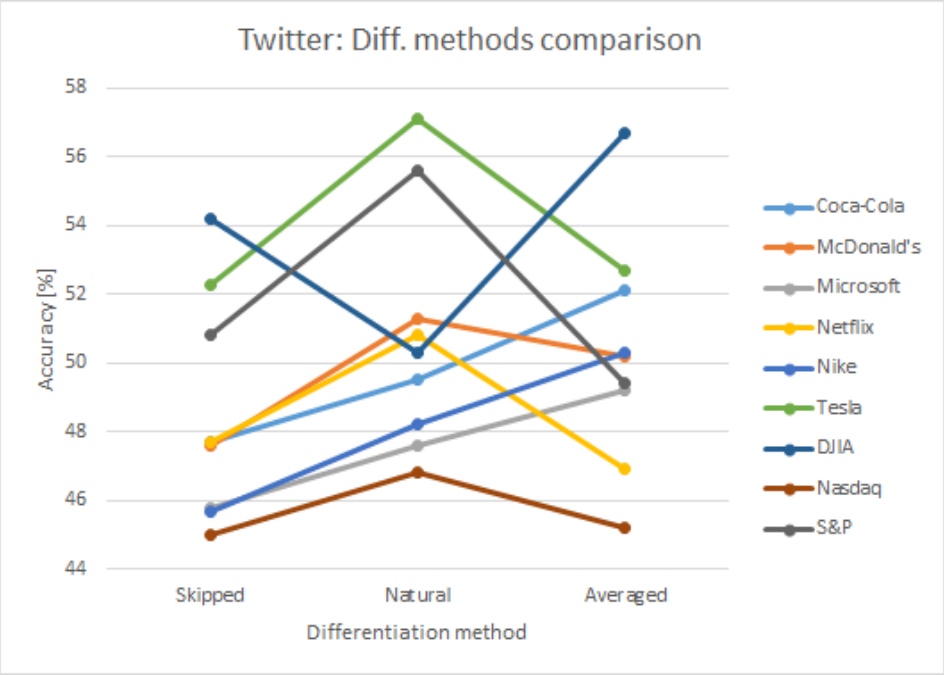


Figure 6.5: Twitter: Influence of Differentiation method on classification accuracy

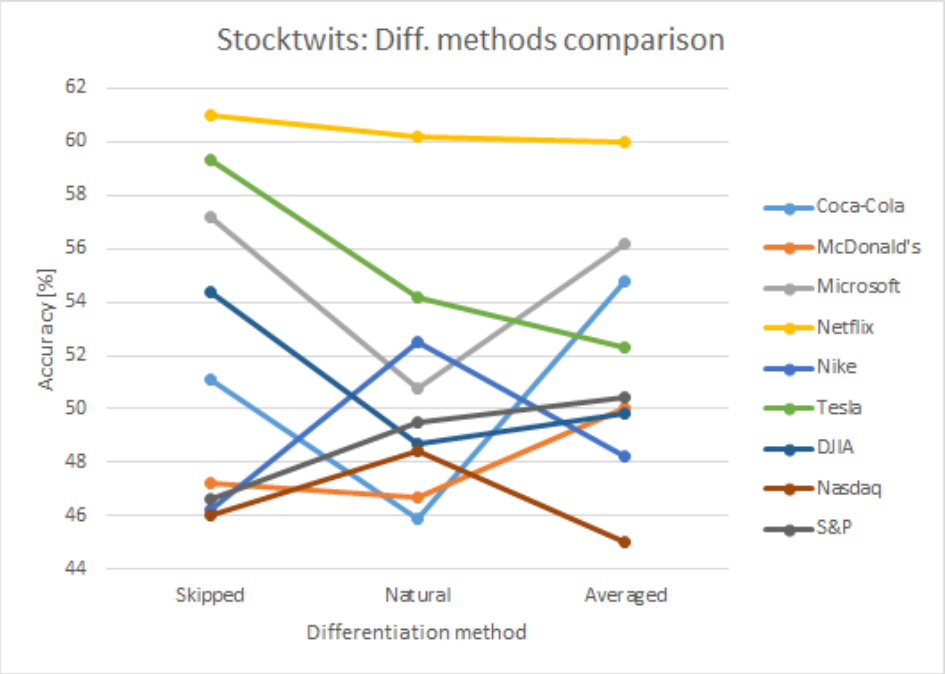


Figure 6.6: Stocktwits: Influence of Differentiation method on classification accuracy

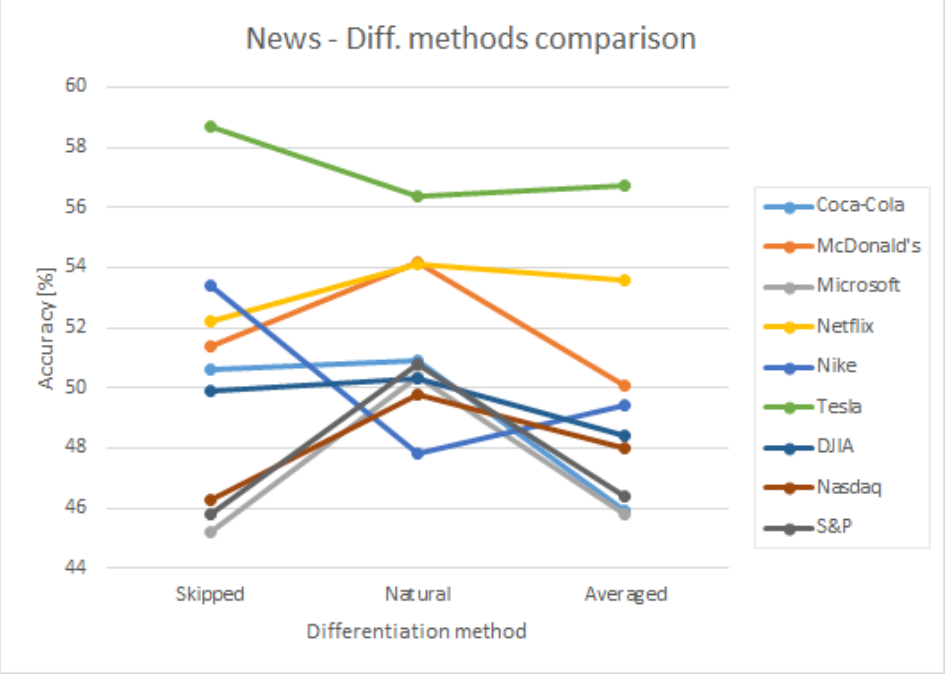


Figure 6.7: News: Influence of Differentiation method on classification accuracy

Figure 6.5 shows the influence of Differentiation methods on the classification accuracy measured for the Twitter data source. As we can see, the Skipped method is clearly the worst performing in almost all the cases. Both Natural and Averaged methods share a similar amount of best performing measurements. Although it is probably impossible to clearly pick the most suitable Differentiation method, the results tell us that the sentiment values measured during weekends or bank holidays are important for the Twitter analysis and they should not be ignored.

A very different tendency can be seen in Figure 6.6 which represents the influence of Differentiation methods on the prediction accuracy measured for the Stocktwits data source. From the chart, it is clearly visible that the Skipped Differentiation method, which ignores weekends, significantly outperforms the Natural method, which includes day-off values in the prediction model. This confirms the assumption which was made in Chapter 3. As it was observed, Stocktwits encounter a critical decrease of user posts during weekends and bank holidays. It was expected that the accuracy of Differentiation methods which use weekend values (natural, averaged) might be decreased as well. This was finally confirmed not only in the Granger causality test but also now in the prediction measurement.

The influence of Differentiation methods on the prediction accuracy measured for the News articles data source is depicted in Figure 6.7. Except for two entities, all the remaining outcomes show the Natural Differentiation method as the best performing one. The conclusion in this case is therefore similar to the one for Twitter: The sentiment values measured during weekends or bank holidays are relevant for the News articles analysis.

### 6.4.3 Best performing results

Now we are going to focus in more detail on the significant results from Figures 6.5, 6.6 and 6.7. In Section 6.3 we have calculated the margin of error being 3%, so we should consider as significant prediction results all the accuracy values higher than 53%. However, as we can see in Figures 6.5, 6.6 and 6.7, the accuracy in some extreme cases drops even to 45%. This tells us that the margin of error is in reality a bit higher than the calculated one, in this case approximately 5%. Given this fact, in the following tables we will consider as significant results only accuracy values of 55% and higher. In order to provide more details about the prediction outcome, we will present also positive/negative precision, positive/negative recall and the most informative feature of classification (the lag of days with the most informative power).

#### Twitter

For the Twitter data source, we obtained three stock entities with significant results which are summarized in Table 6.2. Tesla represents the only company and encounters the accuracy of 57.1% with the most informative feature being the first previous day. This prediction accuracy value is also the very highest one obtained from all the Twitter

prediction measurements. Except for Tesla, there are also DJIA and S&P stock indexes providing some significant outcomes. Both of them consider the sentiment from three days ago as the most informative feature. When looking at the precision and recall values, it is visible that Tesla and S&P classifiers have a slight tendency to prioritize negative outcomes while DJIA does the opposite. However in all the three cases the differences in the precision and recall values are so small that we can consider all the classifiers as well balanced. As we already saw in Figure 6.5, index S&P and Tesla obtained the highest accuracy for the Natural Differentiation method while DJIA shows the best performance using the Averaged method.

When comparing these numbers with the correlation outcomes from Chapter 5, we can see that there is some similarity in the results. Both DJIA and S&P indexes show significant numbers in both correlation and prediction measurements. S&P index has even completely corresponding results including the Natural Differentiation method and 3 days as the most informative lag.

| Stock | Accuracy | Prec-pos | Prec-neg | Rec-pos | Rec-neg | Most inf. Lag |
|---|---|---|---|---|---|---|
| Tesla | 57.1 | 56.9 | 57.4 | 58.9 | 55.2 | 1 |
| DJIA | 56.7 | 57.7 | 56.3 | 52.1 | 61.3 | 3 |
| S&P 500 | 55.6 | 55.5 | 56.1 | 56.4 | 54.6 | 3 |

Table 6.2: Twitter: Details of classification results exceeding 55% prediction accuracy

### Stocktwits

For the Stocktwits data source, we obtained three stock entities with significant prediction results which are shown in Table 6.3. The common aspect of all the results is the Skipped Differentiation method as well as the most informative lag which was identically measured to be the first previous day. Similarly as in the Twitter results, also here the precision and recall values differ only in approximately 1% or 2%. Consequently, we can consider the classifiers as well balanced without any bias tendency. Except for this three stock entities, we also measured the nearly significant accuracy values for Coca-Cola (54.8%) and DJIA (54.4%) but the numbers still fit into the defined margin of error.

The best result of Stocktwits measurement was clearly the prediction accuracy for Netflix which is also the only one from the whole research which crossed the value of 60%. Nonetheless, both Tesla and Microsoft show relatively high accuracy results as well. Moreover, their prediction measurement results accurately match those obtained from the Granger causality test including the most informative lag and the Differentiation method.

### News

In contrast to the results of the Granger causality test where all the values for the News data source were insignificant, the outcome of the prediction measurement actually

| Stock | Accuracy | Prec-pos | Prec-neg | Rec-pos | Rec-neg | Days Lag |
|---|---|---|---|---|---|---|
| Microsoft | 57.2 | 56.9 | 57.8 | 59.6 | 54.8 | 1 |
| Netflix | 61.0 | 61.7 | 61.3 | 59.1 | 63.6 | 1 |
| Tesla | 59.3 | 60.3 | 59.8 | 58.3 | 61.2 | 1 |

Table 6.3: Stocktwits: Details of classification results exceeding 55% prediction accuracy

provides one significant result for Tesla, as it is shown in Table 6.4. The outcome was obtained for the Natural Differentiation method and its most informative feature is a 3-day lag. The small number of significant results combined with the lack of significant outcomes in the Granger causality test slightly undermines the exactness of the result. To the contrary, the correctness of the results is supported by the fact that the highest prediction accuracy is again represented by the same stock entity which had significant outcomes also for the other data sources. Except for Tesla, we measured the nearly significant accuracy values for Netflix (53.6%) and Nike (53.4%) but the numbers fit into the defined margin of error.

| Stock | Accuracy | Prec-pos | Prec-neg | Rec-pos | Rec-neg | Days Lag |
|---|---|---|---|---|---|---|
| Tesla | 58.7 | 58.9 | 59.1 | 59.9 | 57.6 | 3 |

Table 6.4: News: Combination of variables resulting in the highest prediction accuracy

CHAPTER 7

# Summary and future work

## 7.1 Answering the research questions

In this section we will analyze the previously presented results and try to answer the below mentioned research questions which were defined in the beginning of this study.

- What is the best accuracy of the stock price movement predictions obtained from all the observed data and stock variables?
- For which combinations of input data sources and preprocessing techniques do we obtain the best results and why?

### 7.1.1 The highest accuracy performance

As it was defined in Chapter 6, for every combination of input variables we performed 50 classification cycles. The presented prediction accuracy values are thus not the highest outliers obtained from a huge number of independent test runs but they represent the mean values of all the 50 classification cycles performed for one input variable combination. The same values are summarized in Tables 6.2, 6.3 and 6.4.

The highest measured accuracy is 61% which was obtained from the Stocktwits input data for the Netflix company, 2-classes binning and the Skipped Differentiation method.

### 7.1.2 The best input data combinations

#### Data sources

Out of Twitter, Stocktwits and the News articles as our sentiment data sources, Stocktwits is clearly the best performing one for stock prediction purposes. While the best outcomes for Twitter reach only about 57% maximum, Stocktwits exceeds this value for all the

three companies. Moreover, two of the significant results completely correspond to the values obtained in the Granger causality test.

Unfortunately, the obtained results do not tell us the reasons for such a difference in the prediction performance, therefore, finding the proper explanation is only a subject for a discussion. The main benefit of Stocktwits could be its direct relation to the stock market. As all the discussion on this social network is connected to investment, we obtain much higher percentage of relevant information than in the case of News or Twitter where it is hard to filter out all the irrelevant data. When looking at the negatives of these two less performing data sources, another shortcoming of the Twitter analysis might be the unclear target group. The News analysis inaccuracy could be caused by the sentiment analysis approach using every single sentence as an input instead of considering the whole article. In both the Twitter and News data sources there is also a much higher chance of a biased analysis caused by advertisement.

**Stock Entities**

The research analyzed six companies and three stock indexes. Out of all of them, Netflix and Tesla are clearly the best predictable stock entities. Not only because the best overall prediction results were obtained for these two companies but because they both provided positive outputs in various measurements. In the Granger causality test, Neflix had significant values in a correlation with Twitter and Tesla in a correlation with Stocktwits. In the prediction analysis, Netflix occurred as the highest result in the Stocktwits measurement and nearly significant in the News articles. Tesla was significant in all the three data sources.

But what makes the stock movements of Tesla and Netflix so much better predictable in contrast to the other companies and indexes? Although it may seem that a video streaming provider and an electric cars producer do not have much in common, just hitting both their names into Google search together proves such an assumption to be wrong. According to various business journals, Netflix and Tesla are considered to be among the top most overpriced stocks companies. [Col17] [Fis17] If we want to evaluate this statement by using some specific data, we can take a look at the so called PE ratio which is a common measure used in stock market evaluation. The PE ratio calculation is defined in Formula 7.1 where P is a company's stock market price per share and EPS is the earnings per share, which represents the portion of a company's profit allocated to each outstanding share of a common stock.

$$PE = \frac{P}{EPS} \tag{7.1}$$

As it is shown in Table 7.1, Netflix and Tesla have clearly the highest difference between the stock price and the earnings per share (EPS). In the case of Tesla, we can even encounter negative EPS. This makes the resulting PE ratio negative, however, it does not mean that the PE ratio is low, as for example in the case of Nike or McDonald's. On the

contrary, it shows a very big difference between the earning expectations and the reality. Based on these findings, we can conclude that companies with a high (or even negative) PE ratio are more sensitive to the public information sentiment changes and therefore they are more suitable for the sentiment-based prediction of stock price movements.

| Company | Stock Price | Net EPS | PE ratio |
|---|---|---|---|
| Coca-Cola | 44.85 | 0.95 | 47.21 |
| McDonald's | 153.16 | 6.11 | 25.07 |
| Microsoft | 68.93 | 2.71 | 25.44 |
| Netflix | 149.41 | 0.83 | **180.01** |
| Nike | 52.81 | 2.51 | 21.04 |
| Tesla | 361.61 | -4.54 | **-79.65** |

Table 7.1: Price and Earnings as of 6/30/2017

**Differentiation methods**

In the study, there were presented three methods of calculating daily sentiment and stock changes. Although the results do not suggest a single method that would generally outperform the others, some clear source-specific results were measured. For example, in the case of the Stocktwits data source, the Skipped Differentiation method, which ignores all the weekend and bank holiday values, clearly obtained the best results in both the Granger causality test and price movement prediction. As it was already discussed, the main reason is very likely the fact that Stocktwits posts are directly related to the trading days and the bank holidays are intentionally ignored by most users.

In the case of Twitter and News, the advantage of one method over another is not as obvious as in the case of Stocktwits. However, there is a slightly visible opposite tendency to Stocktwits, which means both the Natural and Averaged methods outperforming the Skipped method. In the Granger causality test, all of the four significant results for Twitter belong to these two methods. The results of prediction measurement are providing clearer results, as there are three significant outcomes for the Natural method and one for the Averaged method. This clearly shows that for the Twitter and News data sources, the sentiment values obtained during weekends and bank holidays are as relevant as those obtained during working days so they cannot be ignored in the sentiment and prediction analysis.

**Lag of days**

When examining the results for the lag of days, we can observe a difference between stock companies and stock indexes. When taking both the Granger causality test and the prediction test into consideration, 8 out of 9 significant outcomes for stock companies show a lag of 1 or 2 days while 3 of 4 stock indexes show a lag of 3 days. This result makes sense also from the investment point of view. As stock indexes represent groups of multiple companies which all might have different reaction time to sentiment changes,

the reaction time of the indexes themselves on average increases. On the other hand, the six companies which were analyzed in this research have a lot of daily publicity, which makes it easier for investors to react immediately in case of some new information. This makes their reaction time to the sentiment change smaller.

**Binning**

This preprocessing method was introduced for the prediction test in Chapter 6. As we can see in Figures 6.2, 6.3 and 6.4, for the most of Stocktwits and News articles results the prediction accuracy drops with a higher number of binning classes. In the case of Twitter, the tendency is unclear but definitely not favouring higher binning. Given these results, we can conclude that using a higher number of binning classes does not improve the prediction performance for any of the data sources and there is thus no reason for using other then the basic 2-classes binning.

## 7.2 Summary

This master thesis research successfully performed a comprehensive comparison study of various data sources and processing techniques commonly used for sentiment analysis and stock market price prediction. As a part of the research, there was developed a complex data processing program which was used to download necessary input data, extract their sentiment and predict the stock price movement. Based on the data obtained between November 2016 and August 2017 the correlation and prediction values were calculated, further analyzed and used to answer the research questions.

The main outcomes of the research are the following findings:

- High prediction power of data obtained from Stocktwits social network.
- Direct correlation between PE ratio and the predictability of stock companies.
- Faster sentiment change reaction time of stock companies compared to stock indexes.

Except for the above mentioned conclusions, the research also suggests an optimal setup of the sentiment analysis preprocessing steps and investigates the benefits and shortcoming of various data manipulation methods, as for example sentiment aggregation in time or handling of non-trading days.

## 7.3 Limitations

Although all the described measurements were using various prevention steps to avoid biased results, it is necessary to mention following drawbacks of our analysis:

- The study analyzes only data from the 1st of October 2016 until the 31st of August 2017. This is, however, still a very short period of time to make any solid stock market predictability conclusions.
- The stock market is a dynamic system which encounters various growth trends and changes over time. It is unclear whether the same results would be obtained for a period with a different growth trend.

## 7.4 Future work

The findings presented in the previous sections offer many possibilities of further research and investigation in the same field. One of the biggest options for an additional analysis is the Stocktwits social network which still seems to be rather ignored while Twitter is a very popular subject of various academic studies. They both offer similarly valuable data. Moreover, in comparison with Twitter, Stocktwits introduces a low amount of API limits and it therefore offers a much easier access to a huge amount of historical data for almost all the existing stock market entities. The researched period thus does not need to be bounded by the 10 months, which were used in this study mostly because of the Twitter limitation, but it can be extended much further back in the history in order to increase the accuracy of measurement.

Another space for a new investigation is offered by the findings of an existing correlation between the PE ratio and the price movement predictability. This could serve as an inspiration for another research that could use the existing results and analyze other companies from the top of the PE ratio list, as for example Amazon, Yelp or Salesforce. It would be also interesting to verify the findings by comparing more companies from both the top and bottom of the PE ratio list.

Except for a further analysis of the given findings, there is still a lot of space for improving the existing stock movement prediction accuracy. This could be e.g. achieved by enhancing the existing sentiment analysis approach. While this study was using all the obtained posts and articles without any filtering, it would be possible to increase the relevancy by picking tweets with a certain number of followers or journals with a high number of readers. Another possibility to increase the prediction accuracy is to extend the existing set of input data by combining it with other stock price influencing resources. For example, in Chapter 2 we discussed existing studies which propose to focus on the past movements of the global economy and researched companies, data from annual reports, or maybe even weather conditions.

In general, we can say that there are endless possibilities of further stock price prediction enhancements because, as the Efficient market hypothesis says, the stock market prices reflect the public information made available at any given time. Although it is obviously impossible to gather all the market relevant information, we can try to get as close as possible by further extending the input data sources and improving all the processing and prediction algorithms.

# List of Figures

# List of Tables

# Bibliography


[13] *Twitter: most-used languages*. 2013. URL: https://www.statista.com/statistics/267129/most-used-languages-on-twitter/ (visited on 03/15/2017).

[17a] *About StockTwits*. 2017. URL: https://stocktwits.com/about (visited on 03/15/2017).

[17b] "Granger causality". In: *Wikipedia, The Free Encyclopedia* (June 2017). URL: https://en.wikipedia.org/wiki/Granger_causality (visited on 05/19/2017).

[17c] *McDonald's: Company Profile*. 2017. URL: http://corporate.mcdonalds.com/mcd/investors/company-overview/company-overview-segment-information.html (visited on 11/13/2017).

[17d] *Microsoft Corporation*. 2017. URL: http://www.encyclopedia.com/social-sciences-and-law/economics-business-and-labor/businesses-and-occupations/microsoft-corp (visited on 11/13/2017).

[17e] *Netflix: Most Innovative Company*. 2017. URL: https://www.fastcompany.com/company/netflix (visited on 11/13/2017).

[17f] *NIKE, Inc*. 2017. URL: http://www.encyclopedia.com/social-sciences-and-law/economics-business-and-labor/businesses-and-occupations/nike-inc (visited on 11/13/2017).

[17g] *Top 15 Most Popular Social Networking Sites*. May 2017. URL: http://www.ebizmba.com/articles/social-networking-websites (visited on 06/05/2017).

[17h] *Twitter: number of active users 2010-2017*. Apr. 2017. URL: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ (visited on 03/15/2017).

[BF10] Luciano Barbosa and Junlan Feng. "Robust sentiment detection on twitter from biased and noisy data". In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics. 2010, pp. 36–44.

[BMZ11] Johan Bollen, Huina Mao, and Xiaojun Zeng. “Twitter mood predicts the stock market”. In: *Journal of computational science* 2.1 (2011), pp. 1–8.

[Bro17] Jason Brownlee. *A Gentle Introduction to the Bag-of-Words Model*. Aug. 2017. URL: https://machinelearningmastery.com/gentle-introduction-bag-words-model/ (visited on 11/20/2017).

[CL13] Ray Chen and Marius Lazer. “Sentiment analysis of twitter feeds for the prediction of stock market movement”. In: *Stanford Education, CS229: Machine Learning* (2013). URL: http://cs229.stanford.edu/proj2011/ChenLazer-SentimentAnalysisOfTwitterFeedsForThePredictionOfStockMarketMovement.pdf.

[Col17] Patrick Collinson. “Netflix and Tesla overrated? Yes, in this mad, mad world”. In: *The Guardian* (Mar. 18, 2017). URL: https://www.theguardian.com/money/blog/2017/mar/18/tesla-netflix-share-wall-street-overpriced (visited on 11/13/2017).

[Din+14] Xiao Ding et al. “Using Structured Events to Predict Stock Price Movement: An Empirical Investigation.” In: *EMNLP*. 2014, pp. 1415–1425.

[Fam95] Eugene F Fama. “Random walks in stock market prices”. In: *Financial analysts journal* 51.1 (1995), pp. 75–80.

[Fis17] Adam Fischbaum. *Avoid These 3 Stocks like the Plague: Netflix, Inc. (NFLX), Amazon.com, Inc. (AMZN), Tesla Inc (TSLA)*. Apr. 2017. URL: https://www.smarteranalyst.com/2017/04/17/avoid-3-stocks-like-plague-netflix-inc-nflx-amazon-com-inc-amzn-tesla-inc-tsla/ (visited on 11/29/2017).

[Fol14] Marco Folpmers. *The Twitter Predictor of the Dow: The Rise and Fall*. Apr. 2014. URL: https://folpmers.wordpress.com/2014/04/17/the-twitter-predictor-of-the-dow-the-rise-and-fall/ (visited on 06/12/2017).

[GBH09] Alec Go, Richa Bhayani, and Lei Huang. “Twitter sentiment classification using distant supervision”. In: *CS224N Project Report, Stanford* 1.2009 (2009), p. 12. URL: https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf.

[GK10] Eric Gilbert and Karrie Karahalios. “Widespread Worry and the Stock Market.” In: *ICWSM*. 2010, pp. 59–65.

[Hea17] Reem Heakal. *What Is Market Efficiency?* May 5, 2017. URL: http://www.investopedia.com/articles/02/101502.asp (visited on 01/07/2018).

[Kin15] Harrison Kinsley. *Sentiment Analysis*. 2015. URL: http://sentdex.com/sentiment-analysis/ (visited on 02/18/2017).

[Kol+15] Olga Kolchyna et al. *Twitter sentiment analysis: Lexicon method, machine learning method and their combination*. 2015. URL: http://arxiv.org/abs/1507.00955 (visited on 03/12/2017).

[Lai16] Naomi Lai. *Tesla: Using Social Media to Maximize their Supply Chain Communication*. July 2016. URL: https://smbp.uwaterloo.ca/2016/07/tesla-using-social-media-to-maximize-their-supply-chain-communication/ (visited on 10/28/2017).

[Lam17] Fred Lambert. *Tesla Model S caught fire in Yorkshire, Tesla says cause is due to a crash 2 months before the fire*. Apr. 2017. URL: https://electrek.co/2017/03/31/tesla-model-s-fire-manchester-crash/ (visited on 10/28/2017).

[Lee+14] Heeyoung Lee et al. "On the Importance of Text Analysis for Stock Price Prediction." In: *LREC*. 2014, pp. 1170–1175.

[LHC05] Bing Liu, Minqing Hu, and Junsheng Cheng. "Opinion observer: analyzing and comparing opinions on the web". In: *Proceedings of the 14th international conference on World Wide Web*. ACM. 2005, pp. 342–351.

[Mal03] Burton G Malkiel. "The efficient market hypothesis and its critics". In: *The Journal of Economic Perspectives* 17.1 (2003), pp. 59–82.

[MF70] Burton G Malkiel and Eugene F Fama. "Efficient capital markets: A review of theory and empirical work". In: *The Journal of Finance* 25.2 (1970), pp. 383–417.

[MG12] Anshul Mittal and Arpit Goel. "Stock prediction using twitter sentiment analysis". In: *Stanford Education, CS229: Machine Learning* 15 (2012). URL: http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf.

[MM94] Mark L Mitchell and J Harold Mulherin. "The impact of public information on the stock market". In: *The Journal of Finance* 49.3 (1994), pp. 923–950.

[Moh17] Huda Mohammed. *Netflix Marketing on Social Media*. June 2017. URL: https://smbp.uwaterloo.ca/2017/06/netflix-marketing-on-social-media/ (visited on 11/13/2017).

[Mot13a] David Moth. *How Coca-Cola uses Facebook, Twitter, Pinterest and Google*. Apr. 2013. URL: https://econsultancy.com/blog/62548-how-coca-cola-uses-facebook-twitter-pinterest-and-google (visited on 11/13/2017).

[Mot13b] David Moth. *How McDonald's uses Facebook, Twitter, Pinterest and Google*. Mar. 2013. URL: https://econsultancy.com/blog/62329-how-mcdonald-s-uses-facebook-twitter-pinterest-and-google (visited on 11/13/2017).

[Mot13c] David Moth. "How Microsoft uses Facebook, Twitter, Pinterest and Google". In: *Econsultancy* (Apr. 2013). URL: https://econsultancy.com/blog/62485-how-microsoft-uses-facebook-twitter-pinterest-and-google (visited on 11/13/2017).

[Mot13d] David Moth. *How Nike uses Facebook, Twitter, Pinterest and Google*. Mar. 2013. URL: https://econsultancy.com/blog/62412-how-nike-uses-facebook-twitter-pinterest-and-google (visited on 11/13/2017).

[MRS08] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.

[Per12] Jacob Perkins. *Text Classification for Sentiment Analysis – NLTK Scikit-Learn*. Nov. 2012. URL: https://streamhacker.com/2012/11/22/text-classification-sentiment-analysis-nltk-scikitlearn/ (visited on 05/04/2017).

[PLV02] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques". In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*. Vol. 10. Association for Computational Linguistics. 2002, pp. 79–86.

[RSS13] Michael Rechenthin, W Nick Street, and Padmini Srinivasan. "Stock chatter: Using stock sentiment to predict price direction". In: *Algorithmic Finance* 2.3-4 (2013), pp. 169–196.

[Sch+12] Robert P Schumaker et al. "Evaluating sentiment in financial news articles". In: *Decision Support Systems* 53.3 (2012), pp. 458–464.

[Set] Anil Seth. *Granger causality*. URL: http://www.scholarpedia.org/article/Granger_causality (visited on 06/25/2017).

[She13] Hersh Shefrin. "What Eugene Fama Could Learn From His Fellow Nobel Winner Robert Shiller". In: *Forbes* (Oct. 17, 2013). URL: https://www.forbes.com/sites/hershshefrin/2013/10/17/my-behavioral-take-on-the-2013-economics-nobel/ (visited on 01/07/2018).

[Shi03] Robert J Shiller. "From efficient markets theory to behavioral finance". In: *The Journal of Economic Perspectives* 17.1 (2003), pp. 83–104.

[Sim16] Ioana Sima. *A Netflix Story: The Human Approach to Social Media Marketing*. Aug. 2016. URL: http://www.socialmediatoday.com/marketing/netflix-story-human-approach-social-media-marketing-infographic (visited on 11/13/2017).

[Sto97] Lynn A Stout. *How efficient markets undervalue stocks: CAPM and ECMH under conditions of uncertainty and disagreement*. 1997. URL: https://ssrn.com/abstract=23263 (visited on 02/15/2017).

[Tet07] Paul C Tetlock. "Giving content to investor sentiment: The role of media in the stock market". In: *The Journal of Finance* 62.3 (2007), pp. 1139–1168.

[VW04] Timothy Vick and Barrett Whitener. *How to pick stocks like Warren Buffett*. American Media International, 2004.

[ZS10] Wenbin Zhang and Steven Skiena. “Trading Strategies to Exploit Blog and News Sentiment”. In: *Fourth Int. Conf. on Weblogs and Social Media (ICWSM)*. May 2010.