UNIVERSITETI I EVROPËS JUGLINDORE
SOUTH EAST EUROPEAN UNIVESRSITY
УНИВЕРЗИТЕТ НА ЈУГОИСТОЧНА ЕВРОПА

FAKULTEI I SHKENCAVE DHE
TEKNOLOGJIVE BASHKËKOHORE
ФАКУЛТЕТ ЗА СОВРЕМНИ НАУКИ
И ТЕХНОЛОГИ
FACULTY OF CONTEMPORAY
SCIENCES AND TECHNOLOGIES

POSTGRADUATE STUDIES – SECOND CYCLE

THESIS:

"STOCK PRICE PREDICTION BASED ON FINANCIAL NEWS"

CANDIDATE:
Mubarek Selimi

MENTOR:
Assoc.Prof.Dr. Adrian Besimi

Tetova, 2019

Contents

Declaration of original work

I certify that I am the original author of this thesis. I have not copied from any other students' work or from any other sources apart from reviewed references in accordance with the rules of referencing.

Mubarek Selimi

Abstract

The main intention of this master thesis is to create a model to predict the stock price movements using news articles from relevant sources and past stock prices.

In this thesis it is proposed a new model and needed steps that one should undertake in order to try and predict potential stock price fluctuation solely based on financial news from relevant sources. This thesis starts with providing introduction and background information on the problem and text mining in general, furthermore supporting the idea with relevant research papers needed to focus on the problem that it is investigated in this thesis. The models in this thesis relies on existing text-mining techniques used for sentiment analysis, combined with historical data from relevant news sources as well as stock data.

In this thesis two models are created to predict stock price movements, the stock price movements are predicted whether the prices will go up, down or neutral. The data set is created from collecting news articles from identified relevant sources for targeted companies.

The first model it is created with 4 variables and they are: *source*, *company, 5-day ROC* and *sentimentof_text*, and it achieves an accuracy of 94.29%. While the second model it is created with three variables and they are: *source*, *company* and *sentimentof_text*, and it achieves an accuracy of 49.49%

List of Tables

List of Figures

Chapter 1. Introduction

Internet has become more popular and is a large part of many people's lives. Each day, the dependency on it increases in many areas such as education, technology and the financial markets. The speed at which data is produced has increased to a degree at a rate that is impossible to process, and it has encouraged research in many areas, including data mining and text mining. These two areas have emerged in the last decade mainly due to research in artificial intelligence, machine learning, and inferential statistics (Vale, 2018).

Lots of investors are involved in stock market and they are all interested to know more about the future of market to be able to have more successful investments. Effective market prediction can help investors with trade advices or can be used as a component inside automatic trader agents, ability to predict in a market economy is equal to being generate wealth by avoiding financial losses and making financial gains. Recently some of the researchers have found that news is one of the most influential sources that affect stock market and are necessary in achieving to more accurate predictions. Stock prices are determined by supply and demand of investors, the most important information that investors used to make investment decision is financial news but it is a hard and time consuming task to read and analyze a lot of news published on several sources (Nikfarjam, Emadzadeh, & Muthaiyah, 2010) (Kaya & Karsligil, 2010).

Information published in news articles influence, in a varying degree, the decision of the stock traders, especially if the given information is unexpected. It is important to analyze this information as fast as possible, so it can be used as help for trading decisions by traders before the market has had time to adjust itself to the new information. (Aase, 2011).

In this thesis, there are analyzed relevant financial news articles. Stock price movements are analyzed about the three public listed companies: Apple, Facebook and Tesla. The model is going to leverage the Naïve Bayesian classifier for document classification to make prediction for whether the stock prices goes up, down or neutral, based on the dataset that is generated from the steps conducted in this thesis

Background

The advances in data mining and text mining, allied with the velocity and the way the news articles are published, created opportunities to use text mining applied to financial market prediction (TMFP). Nevertheless, to make possible computers to interpret news articles at the right time and generate profit in financial markets, an interdisciplinary field of research has been created, The Venn diagram in Figure 1 describes the three disciplines involved in this emerging field (Beckmann, 2017).

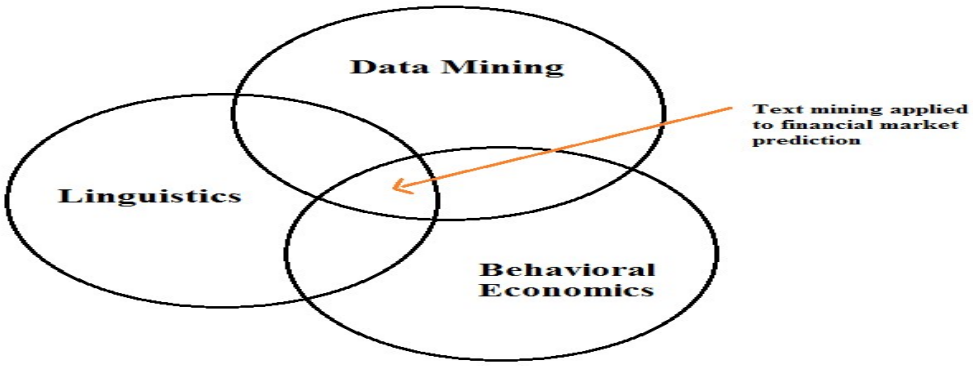


Figure 1 - Venn diagram describing the intersection of discipline involved in this work (Beckmann, 2017)

TMFP in supported by Behavioral Economics (BE) theories, which analyses the psychological, social, cognitive, and emotional aspects of human behavior when taking investment decisions. BE claims that human can make irrational decisions that lead to discrepancies and market inefficiencies. Due to this inefficiency, the stock prices cannot reflect in real time the changes in the world, creating an opportunity for predictive techniques like data mining and text mining (Beckmann, 2017).

One important application of text mining is text sentiment analyses, also referred to as opinion mining, this technique tries to discover the sentiment of a written text. Sentiment analysis is the process of determining people's attitudes, opinions, evaluations, appraisals and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. Sentiment analysis classifies textual data into positive, negative and neutral sentiments so this can be used to categorize text documents into set of predefined sentiment categories (Aase, 2011) (Khedr, Ayman Elsayed; Salama, S E; Yaseen, Nagwa, 2018; Khedr, Ayman Elsayed; Salama, S E; Yaseen, Nagwa, 2018).

Problem Statement

Financial analyst who invest in stock market usually are not aware of the stock market behavior, they are facing the problem of stock trading as they do not know which stocks to buy and which to sell in order to gain more profits. All these users know that the progress of the stock market depends a lot on relevant news and they have to deal daily with vast amount of information. They have to analyze all the news that appears on newspapers, magazines and other textual resources. But analysis of such amount of financial news and articles in order to extract useful knowledge exceeds human capabilities. Text mining techniques can help them automatically extracting the useful knowledge out of textual resources (Falinouss, 2007).

Considering the assumption that news articles might give much better predictions of the stock market than analysis of past price developments, and in contrast to the traditional series analysis, where predictions are made based on solely on the technical and fundamental data, we want to investigate the effects of textual information in predicting the financial markets. We would develop a system which is able to use text mining techniques to model the reaction of the stock market to news articles and predict their reactions. By doing so, the investors are able to foresee the future behavior of their stocks when relevant news are released and act immediately upon them (Falinouss, 2007).

In this thesis is proposed a system predicting stock price fluctuations or movements by analyzing financial news articles, to accomplish this objective, a complete process of data mining and text mining was developed to predict the price movements for the 3 companies listed public.

Research field

In my work there will be analyzed the text news for the purpose of financial market prediction. In general, the design of TMFP (Text Mining applied to Financial Market

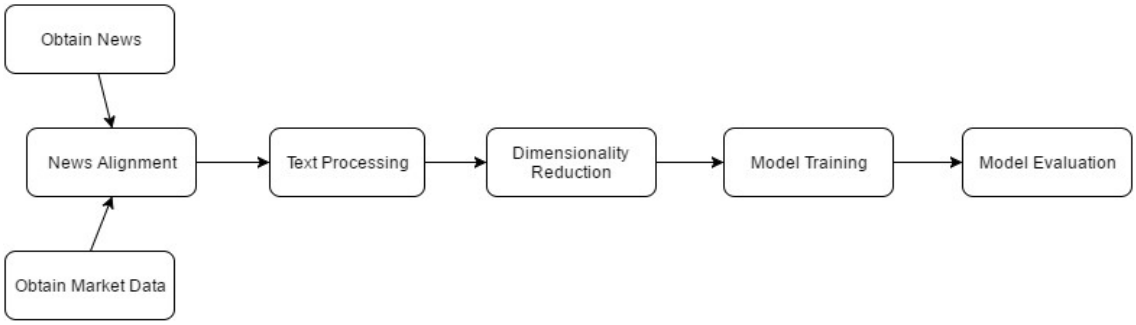


Figure 2 - General design of a TMFP process (Vale, 2018).

Prediction) systems follows a structure like figure 2.

In this thesis there will be applied text mining in financial market, through text mining methods we will try to make a prediction for the direction of stocks about 3 companies. The two fields that are studied are finance and text mining.

A stock is a type of financial asset that denotes part ownership on the assets and profits of a company, it also entitles the owner of the stock to receive dividends if the company chooses to pay some of their profits to the shareholders. Typically, ownership of a stock also gives the investor a right to vote on corporate decisions at shareholder meetings (Aase, 2011).

Stocks are usually traded at one or more stock exchanges. The exchange receives and influx of orders to buy or sell a given volume of a stock at a given price, which are matched together making a trade when the price of a buy order matches the price of a sell order. Historically, stock exchanges have been physical places where stock brokers placed orders to buy or sell stocks in person, but with the technological advances more and more stock exchange have become purely electronical. The world's largest exchange, the New York Stock Exchange (NYSE), still has a physical trading floor (although most orders are now electronically entered). Another popular exchange, NASDAQ, is managed completely electronically (Aase, 2011).

Aims of the research

The trading of stock in public companies is an important part of the economy. Stocks are a type of security that represent ownership interest in a company. Stock trading allows businesses to raise capital to pay off debt, launch new products and expand operations. For investors, stocks offer the chance profit from gains in stock value as well as company dividend payments. Stock prices influence consumer and business confidence, which in turn affect the overall economy. The relationship also works the other way, in that economic conditions often impact stock markets (Basu, 2018).

The financial market is a complex, evolutionary, and non-linear dynamical system. The field of financial forecasting is characterized by data intensity, noise, non-stationary, unstructured nature, high degree of uncertainty, and hidden relationships. Therefore, predicting price movement in financial markets is quite difficult (Falinouss, 2007).

The main aim of this research is to predict the reaction of stock market to news articles, which are rich in valuable information and are more superior to numeric data. To investigate the influence of news articles on stock price movement, different text mining techniques are implemented to make the prediction model. With the application and prediction system would be learned using text classifier. Feeding the system with upcoming news, it forecasts the stock price trend (Falinouss, 2007).

The primary goal in this study is to give prediction of the future stock direction (up, down or neutral) after making a sentiment analysis of the financial news and calculating the Rate of Change (ROC) in average of 5 days. Using the Naïve Bayes classifier to classify as up, down or neutral.

Importance of the thesis

Stock market have been studied repeatedly to extract useful patterns and predict their movements. The reason is that who can beat the market, can gain excess profit. Financial analysts who invest in stock markets usually are not aware of the stock market behavior. They are facing the problem of stock trading as they do not know which stocks to buy and which to sell to gain more profits. If they can predict the future behavior of stock prices, they can act immediately upon it and make profit (Falinouss, 2007).

The more accurate the system predicts the stock price movement, the more profit one can gain from the prediction model. Stock prices trend forecasting based solely on the technical and fundamental analysis enjoys great popularity. But numeric series data only contain the event and not the cause why it happened. Textual data such as news articles have richer information, hence exploiting textual information especially in addition to numeric time series data increases the quality of the input and improved predictions are expected from this kind of input rather than only numerical data (Falinouss, 2007).

Without the doubt, human behaviors are always influenced by their environment. One of the most significant impacts that affect the humans' behavior comes from the mass media or to be more specific, from news articles. On the other hand, the movements of prices in financial markets are the consequences of the actions taken by the investors on how they perceive the events surrounding them as well as the financial markets. As news articles will influence the human' decision will and humans' decision will influence the stock prices, news articles will in turn affect the stock market indirectly (Falinouss, 2007).

An increasing amount of crucial and valuable real-time news articles highly related to the financial markets is widely available on the Internet. Extracting valuable information and figuring out the relationships between the extracted information and the financial markets is a critical issue, as it helps financial analyst predict the stock market behavior and gain excess profit (Falinouss, 2007).

Hypotheses

This thesis will use hypothesis testing to answer the research questions, to conduct the research and verify the findings. Consequently, the following hypotheses were raised, supported by additional hypotheses:

- **H1. The financial news in the media tend to significantly change stock prices for the company on the stock market.**
  - **H1a.** Positive financial news on the media tend to significantly increase stock prices.
- **H2. Even though EMH clearly states that financial stock prices cannot be predicted we argue that based on several attributes from new articles we can reach certain level of prediction and give directions to financial experts.**

In this thesis Rate of Change (ROC) is calculated to define the significant change in stock prices.

The Price Rate of Change (ROC) is a momentum-based technical indicator that measures the percentage change in price between the current price and the price a certain number of periods ago. The ROC indicator is plotted against zero, with the indicator moving upwards into positive territory if price changes are to the upside and moving into negative territory if price changes are to the downside (MITCHELL, 2019). The formula for the Price Rate of Change Indicator is:

$$\text{Rate of Change (ROC)} = \left(\frac{Close\ Price - Clos\quad Price\ n\ Periods\ Ago}{Close\ Price\ n\ Periods\ Ago}\right) * 100$$

Structure of the thesis

Chapter 1: Introduction. This chapter provides an introduction of the research and sets it within a context. It summarizes the problems, importance that have imposed the development of the thesis and gives an overview of the aim of research, research field and the hypotheses raised in this thesis.

Chapter 2: Literature Review. The aim of this chapter is to introduce the knowledge discovery database, data mining and text mining. Explain the approaches to the stock market prediction and the theories available that are discussed when predicting the future stock prices. Giving an overview about the studies that have been conducted in this field.

Chapter 3: Research Methodology. This chapter illustrates the way of research has been conducted by presenting the steps that need to be undertaken to achieve to make the predictions about the stock price movements. The general methods to make sentiment analysis with text mining approach are explained.

Chapter 4: Implementation. The aim of this chapter is to explain in more details the steps that have been undertaken to implement the models that aim to achieve a prediction about the future stock market prices**.** Each step is described, and then summary statistics of the collected data set is given.

Chapter 5: Results and Discussion. This chapter presents the results achieved in this thesis. The results are shown and discussed. The author's findings based on the hypotheses raised are discussed in detail.

Chapter 6: Conclusion. The chapter 6 summarizes the importance of the topic and presents the conclusion of the main findings and the results achieved in this thesis**.** The main results and findings are summarized.

Bibliography

Chapter 2. Literature Review

Knowledge discovery in databases

Information obtained from the data analysis can be used for several applications, ranging from business management, production control and market analysis to engineering and scientific exploration. For (Fayyad & Shapiro, 1996), the value is not in storing the data, but rather in our ability to extract useful reports and to find interesting trends and correlations, using statistical analysis and inference, to support decisions and policies made by scientists and businesses. In this way, researchers motivated by the challenge of transforming information into knowledge, soon come across Knowledge Discovery in Databases (KDD), emphasizing the data mining (DM) application (Vale, 2018).

Knowledge Discovery in Databases includes the entire process of knowledge extraction, including how data is stored and accessed, how to develop efficient and scalable algorithms that can be used to analyze massive data sets, how to interpret and visualize the results, and how to model and support the interaction between human and machine (Vale, 2018).

The KDD consists of a series of defined steps, each destined to the conclusion of a determined task of discovery, and realized by the application of a method discovery:

1. **Data Integration:** data from different sources are collected and put together.
2. **Data Selection:** discard data that might be not useful for your research.
3. **Data Cleaning:** clean data by applying different techniques to deal with data errors, missing values, noisy and inconsistent data.
4. **Data Transformation:** data may need to be transformed into something appropriate for your model to speed up process, associate types, normalize values, etc.
5. **Data Mining:** data discovery process to finding patterns. At this step, a data mining model will be applied into the data.
6. **Pattern Evaluation:** step to visualize the patterns generated.
7. **Decision:** step responsible for heling user to understand the results in order to take better decisions (Vale, 2018).

This Sequence compromises the cycle that the data travels until it becomes

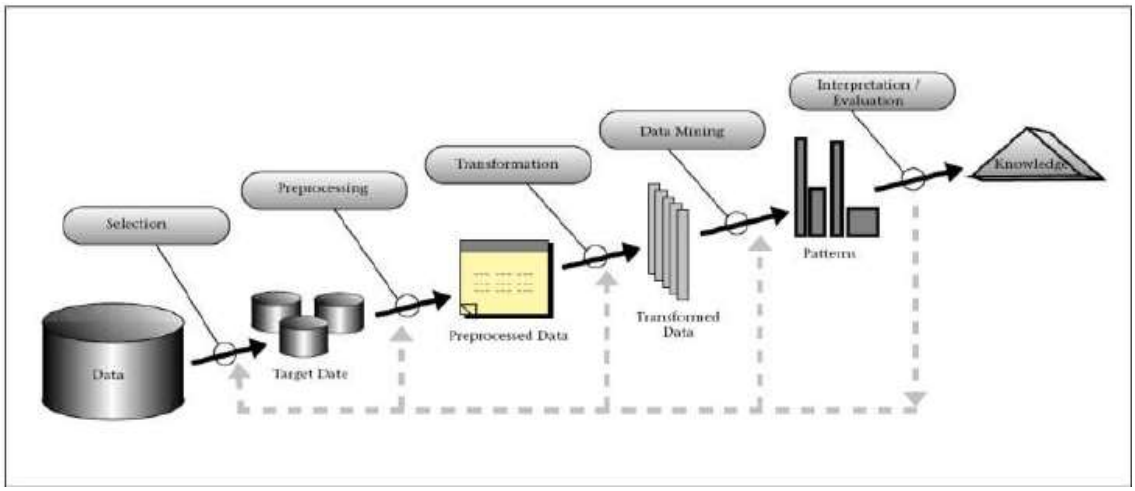


Figure 3- KDD process cycle ((Fayyad & Shapiro, 1996))

useful knowledge, as shown in figure 2:

Although all stages of the KDD process cycle must occur in the best possible way in order to achieve the desired result and thus must be equally important for the transformation of the information into useful knowledge, the data mining phase can be considered the core of the whole process (Vale, 2018).

Data mining and Text mining

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically (Han, Kamber, & Pei, 2011).

Until recently computer scientists and information system specialist concentrated on the discovery of knowledge from structured, numerical databases and data warehouses. However, much, if not the majority, of available business data are captured in text files that are not overtly structured (Falinouss, 2007).

Text Mining is the process of seeking or extracting the useful information from the textual data. It is an exciting research area as it tries to discover knowledge from unstructured texts. It is also known as Text Data Mining (TDM) and knowledge Discovery in

Textual Databases (KDT). KDT plays an increasingly significant role in emerging applications, such as Text Understanding, Text mining process is same as data mining, except, the data mining tools are designed to handle structure data whereas text mining can able to handle unstructured or semi-structured data sets such as emails, HTML files and full text documents etc. (Vijayarani & Ilamathi, 2015).

Approaches to Stock Market Prediction

Mainly there are two methods for forecasting market trends. One is technical analysis and other is Fundamental analysis.

Technical Analysis

Technical analysis considers past price and volume to predict the future trend, those who believe that historic market movements are bound to repeat themselves are known as technical analysts (Nassirtoussi, Aghabozorgi, Wah, & Ling Ngo, 2014).

Technical analysis depends on historical and time-series data. These strategists believe that market timing is critical, and opportunities can be found through the careful averaging of historical price and volume movements and comparing them against current prices (Schumaker & Chen, 2009).

Technical analysis (Pring, 1991) is based on numeric time series data and tries to forecast stock markets using indicators of technical analysis. It is based on the widely accepted hypothesis which says that all reactions of the market to all news are contained in real-time prices of stocks. Because of this, technical analysis ignores news. Its main concern is to identify the existing trends and anticipate the future trends of stock market from charts. But charts or numeric time series data only contain the event and not the cause why it happened (Falinouss, 2007).

Fundamental Analysis

Fundamental analysis of a business involves analyzing financial data to get some insights. In fundamental analysis analysts look at fundamental data that is available to them from different sources and make assumptions based on that. Fundamental data is usually of an unstructured nature and it remains to be a challenge to make the best use of it efficiently

through computing, the research challenge here is to deal with the unstructured data (Nassirtoussi, Aghabozorgi, Wah, & Ling Ngo, 2014) (Joshi & Rao, 2016).

Fundamental analysis (Thomsett, 1998) investigates the factors that affect supply and demand. The goal is to gather and interpret this information and act before the information is incorporated in the stock price. The lag time between an event and its resulting market response presents a trading opportunity. Fundamental analysis is based on economic data of companies and tires to forecast markets using economic data that companies have to publish regularly, i.e. annual and quarterly reports, auditor's reports, balance sheets, income statements, etc. News has an importance for investors using fundamental analysis because news describes factor that may affect supply and demand. (Falinouss, 2007).

This study follows the Fundamental analysis techniques to discover future trends of a stock by considering news articles about a company as prime information and tries to classify news as good (positive) and bad (negative).

The efficacy of both technical and fundamental analysis is disputed by the efficient-market hypothesis which states that the stock market prices are essentially unpredictable (Joshi & Rao, 2016).

Theories of Stock Market Prediction

When predicting the future Stock Market prices, there are several theories available, the most famous and discussed are: Efficient Market Hypothesis and Random Walk Theory.

Efficient-Market Hypothesis

The efficient-market hypothesis (EMH) states that market prices always reflect all available information, or in other words, financial markets are informational efficient. This means that no one can consistently achieve greater returns than that of the average-market returns, not even if they are given all public ally published information that are available at the time of investment (Aase, 2011).

The EMH is divided into three different hypotheses: weak form efficiency, semi-strong form efficiency, and strong form efficiency, each of which has different implications for how the market works (Aase, 2011).

Weak form efficiency states that future prices cannot be predicted from analyzing historical prices. In other words, excess returns, or profits, cannot be gained in the long run by using investment strategies based on historical prices or other historical forms of data. This means that technical analysis will not be able to consistently produce excess returns. This is because one of the main principles vital to technical analysis states that history trends to repeat itself. It states that the stock prices exhibit no serial dependencies, meaning that there exist no "patterns" to asset prices, which is especially important for chartists which is a subfield under technical analysis. Weak form efficiency states that all future price movements follow a random walk, unless there is some change in some fundamental information. It does not state that prices adjust immediately in the advent of new fundamental information, which means that some forms of Fundamental analysis and also news article analysis might provide excess returns. This is because they trade on new information and does not use any historical information to look for patterns (Aase, 2011).

Semi-strong from efficiency implies that share prices adjust in an unbiased fashion to new publicly available information very rapidly such that no excess returns can be earned by trading on that information. This form EMH implies that fundamental analysis, technical analysis nor news trading will be able to reliably produce excess return over time. (Aase, 2011).

In strong-form efficiency, stock prices reflect all information, public and private, and no one can earn excess returns. According to this form of EMH, those traders that are consistently getting profitable return are only lucky since they are among the randomly selected few that are (Aase, 2011).

There has been a lot of criticism against the EMH. Maybe because it assumes that investors always behave rationally, but many behavioral economists argue that the presences of cognitive biases negate the validity of this assumption (Aase, 2011).

In my thesis, for my hypothesis to be true the EMH has to be false. If it is not, only the weak form of EMH can be true. If the semi-strong or the strong form EMH is true, then my hypothesis in this study will be false.

Random Walk Theory

The theory of the random walk hypothesis (Cootner, 1964) claims that stock market prices changes according to a random walk and, consequently, prices cannot be predicted. Therefore, it is impossible to consistently outperform the average market return. This theory is based on the efficient-market hypothesis (EMH), which says that prices fluctuate randomly about intrinsic value. It also holds that the best trading strategy to follow would be a simple "buy and hold" instead of any attempt to "beat the market".

Th primary experiment conducted for this theory was to draw a price graph randomly and have some chartists (technical analysts that analysis price charts) analyze it. The price graph started from an initial value of fifty dollars and all future movements (ups and downs) was chosen by performing a coin flip (fifty-fifty chance for each movement). The chartist, after analyzing the graph, found signs that made him recommend buying. This is then used to argue that the market and stocks could be just as random as flipping a coin (Aase, 2011).

Related work

Many works over the years have continued to prove that the news is closely related to stock prices. With the recent explosive increase in the amount of unstructured text data from the internet, mobile channels, and SNS (Social Network service), there have been attempts to predict stock movements using such text data (Kim, Jeong, & Ghani, 2014).

Different research approaches proved that there is a strong correlation between financial news and stock price changes like (Khedr, Ayman Elsayed; Salama, S E; Yaseen, Nagwa, 2018), in their paper, they proposed and approach that uses sentiment analysis for financial news, along with features extracted from historical stock prices to predict the future behavior of stock market. Three categories of news data were considered: news relevant to market, company news and financial reports that were published by financial experts about stocks. The proposed model consists of two stages, the first stage is to determine the news polarities to be either positive or negative using naïve Bayes algorithm,

and the second stage incorporates the output of the first stage as input along with the processed historical numeric data attributes to predict the future stock trend using K-NN algorithm. The results of their proposed model achieved higher accuracy for sentiment analysis in determining the news polarities by using Naïve Bayes algorithm up to 86.21%. In the second stage of analysis, results proved the importance of considering different values of numeric attributes, and this achieved the highest accuracy compared to other previous approaches, the model for predicting the future behavior of stock market obtained accuracy up to 89.80%. In the proposed model, both Naïve Bayes and K-NN methods lead to the best performance while Support Vector Machines (SVM) has the lowest accuracy. The results of the proposed model are compatible with the researches that state that there is a strong relation between stock news and changes in stock prices (Khedr, Ayman Elsayed; Salama, S E; Yaseen, Nagwa, 2018).

The paper of (Hagenau, Liebmann, & Neumann, 2012), examines whether stock price prediction based on textual information in financial news can be improved as previous approaches only yield prediction accuracies close to guessing probability (50%). Accordingly, they enhance existing text mining methods by using more expressive feature to represent text and by employing market feedback as part of feature selection process. The authors show that a robust feature selection allows selection allows lifting classification accuracies significantly above previous approaches when combined with the complex feature types. This research shows that the combination of advanced feature extraction methods and feedback-based feature selection boosts classification accuracy and allows improved sentiments analytics. According to the authors feature selection significantly improves classification accuracies because their approach allows reducing the number of less-explanatory features, i.e., noise, and thus, may limit negative effects of over-fitting when applying machine learning approaches to classify text messages. When feedback-based feature selection is combined with 2-word combinations, accuracies of up to 76% are achieved. These results were possible as 2-word combinations capture the meaning and context of information pieces in text.

Stock prices prediction is one of the most important issues to be investigated in academic and financial researches. A prediction model, finding and analyzing the correlation between contents off news articles and stock prices and then making predictions for future

prices, was developed (Kaya & Karsligil, 2010). They retrieve financial news articles published in last year and get the stock prices for same period. All articles are labeled positive or negative according to their effects on stock price. So, they use price changes to label the articles. While analyzing textual data, authors use word couples consisting of a noun and a verb as features instead of using single words, they argued that defining word couples as feature is more suitable for classification of financial articles than defining single words as features. Defined word couples composed from a noun and a verb word occurring in a same sentence as features, because it may give an idea to know that, a specific verb occurs in a same sentence, about whether the sentence is positive or not. Authors use chi-square feature selection method which according to them is more suitable for text classification. In this method, dependence between classes and features are calculated and used to weight features. Many classification methods can be used for document classification, in this study is used the support vector machines method to classify financial news articles. In conclusion this system automatically analyzes and classifies news articles and generates recommendations for investors. They acquired 61% accuracy from their study. This accuracy is greater than random prediction, which has 50% accuracy, these results argue that there is a strong relationship between financial news and stock price movements

Many research has been carried in the area of prediction of stocks, (Joshi, Kalyani; Rao, Jyothi; N., Bharathi N., 2016) project is about taking non quantifiable data such as financial news articles about a company and predicting about a company and predicting its future stock trend with news sentiment classification. This research is an attempt to build a model that predicts news polarity which may affect changes in stock trends, authors have taken past three years data from Apple Company as stock price and news articles. They assumed that news articles and stock prices are related to each other. And, news may have capacity to fluctuate stock trend. So, they thoroughly studied this relationship and concluded that stock trend can be predicted using news articles and previous price history. As news articles capture sentiment about the current market, they automate this sentiment detection and based on the words in the news articles, can get an overall news polarity. If the news is positive, then it can be stated that this news impact is good in the market, so more chances of stock price go high. And if the news is negative, then it may impact the

stock price to go down in trend. They used polarity detection algorithm for initially labelling news and making the train set. For this algorithm, dictionary-based approach was used. The dictionaries for positive and negative words are created using general and finance specific sentiment carrying words. Then preprocessing of text data was also a challenging task, they created own dictionary for stop words removal which also includes finance specific stop words. Based on this data, they implemented three classification models and tested under different test scenarios. Then after comparing their results, Random Forest worked very well for all test cases ranging from 88% to 92% accuracy. Accuracy followed by SVM is also considerable around 86%. Naïve Bayes algorithm performance is around 83%. Given any news article, it would be possible for the model to arrive on a polarity which would further predict the stock trend.

News articles serve the purpose of spreading company's information to the investors either consciously or unconsciously in their trading strategies on the stock market. Recently, the number of online news have rocketed which make it hard for the investors to cover all the latest information. As a result, many kinds of automated systems have been implemented to support the investors. Take stock trends as example, if the direction of the selected stock was predicted to be "up" in the next 24 hours, investors would definitely hold that stock shares to earn more profit. The research paper conducted by (Dang, Minh; Doung, Duc;, 2016), they propose an approach of using time series analysis and improved text mining techniques to predict daily stock market directions. Initially, they scrawl the articles from the online websites then they extract the file's content in plain text format. After pre-processing all the news in the collection, they created an optimized dataset. Next, they label each news into a specific class of positive, negative or neutral by using the stock prices. Several terms weighting methods have been proposed by researchers in the text mining field and they used delta TF-IDF method instead of the normal TF-IDF, the goal of this new algorithm was to increase the importance of the word which unevenly distributed between positive and negative class, reduce the importance of the word which evenly distributed between positive and negative class. They used support vector machine as their machine learning method to classify the financial news articles. They gathered the stock news automatically from popular financial websites between May 1st, 2014 and April 30th, 2015 by using a web crawler tool. They collected 1884 different articles. In their work, they

selected only the news relating to companies in the VN30 Index. Authors have proved the correlation between the financial news and the stock prices. To achieve that, the financial news and the stock prices were gathered for careful experiment. They have achieved quite a high accuracy at 73%. Moreover, they removed the weak stock tickers in VN30 index by the technical indicators and the results proved that the performance of the system has greatly been improved. However, according to the authors the success ratio of their proposed system would be increased if they analyzed the news from more reliable sources (Dang, Minh; Doung, Duc;, 2016).

The research study from (Shynkevich, Yauheniya; Coleman, Sonya; Belatreche, Ammar;, 2015) explores whether simultaneous usage of different financial news categories can provide an advantage in financial prediction system based on news. Five categories of news articles were considered: news relevant to a target stock and news relevant to its sub industry, industry, group industry and sector. Each category of news articles was pre-processed independently, and five different subsets of data were constructed. The Multiple Kernel Learning (MKL) approach was used for learning from different news categories, independent kernels were employed to learn from each subset. A number of different types of kernels and kernel combinations were used. The findings have shown that the highest prediction accuracy and return per trade were achieved for MKL when all five categories of news were utilized with wo separate kernels of the polynomial and Gaussian types used for each news category. The highest kernel weights were assigned to the polynomial kernels indicating that this kernel type contributes the most to the final decision. The SVM and kNN methods based on single category of news, either stock-specific (SS), sub-industry-specific (SIS), industry-specific (IS), group-industry-specific (GIS) or sector-specific (SeS), demonstrated worse performance than MKL. These results indicate that dividing news items into different categories based on their relevance to the target stock and using separate kernels for learning from these categories allows the system to learn and utilize more information about the future price behavior which gives an advantage for more accurate predictions. The achieved results are promising and can be enhanced further. The introduction of additional data sources such as historical prices can potentially improve the performance.

Stock market data analysis needs the help of artificial intelligence and data mining techniques. The volatility of stock prices depends on gains or losses of certain companies. News articles are one of the most important factors which influence the stock market. The study conducted by (Dnyaneshwar , Kirange K ; Ratnadeep , Deshmukh;, 2016), basically shows the effect of emotion classification of financial news to the prediction of stock market prices. In order to find correlation between sentiment predicted from news and original stock price, they plot the sentiments of two companies (Infosys and Wipro) over a period of 10 years. For emotion classification, various classifiers such as Naïve Bayes, Knn and SVM are evaluated. They have identified the various sentiment categories as positive, negative, neutral and conflict. The comparison between positive sentiment curve and stock price trends reveals co-relation between them. The accuracy level of this predictive model is satisfactory for the authors, the achieved accuracy with SVM, Knn and Naïve Bayes are 75.45%, 47.98% and 72.64 respectively. Authors believe that it can be further developed by incorporating more complex classifiers, and analysis techniques of machine learning or data mining (Dnyaneshwar , Kirange K ; Ratnadeep , Deshmukh;, 2016).

Predicting the behaviors of the stock markets are always an interesting topic for not only financial investors but also scholars and professionals from different fields, because successful prediction can help investors yield significant profits. In their work (Thanh, Hoang T. P.; Meesad, Phayung;, 2013), they proposed an approach of using time series analysis and text mining techniques to predict daily stock market trends. The research is conducted with the utilization of a database containing stock index prices and news articles collected from Vietnam websites over 3 years from 2010 to 2012. A robust feature selection and a strong machine learning algorithm are able to lift the forecasting accuracy. By combining Linear Support Vector Machine Weight and Support Vector Machine algorithm. The results showed that data set represented by 42-features achieves highest accuracy by using one-against-one Support Vector Machines up to 75% and one-against-one method outperforms one-against-all method which achieved about 68%. In conclusion, this work will support investors to make their decisions (Thanh, Hoang T. P.; Meesad, Phayung;, 2013).

Stock market prediction has always had a certain appeal for researchers, (Shcumaker, Robert P.; Chen, Hsinchun;, 2009) conducted a research that examines a predictive machine learning approach for financial news articles analysis using several

different textual representation: bag of words, noun phrases, and named entities. Through this approach, they investigated 9,211 financial news articles and 10,259,042 stock quotes covering the S&P 500 stocks during a five-week period. Stock quotes are gathered on a per minute basis for each stock. When a news article is released, they estimate what the stock price would be 20 minutes after the article was released. To do this they perform linear regression on the quotation data using an arbitrary 60 minutes prior to article release and extrapolate what the stock price should be 20 minutes in the future. To test the types of information that need to be included, authors developed four different models and varied the data given to them. The first model, regress, was a simple linear regression estimate of the +20-minute stock price. Next the three models use the supervised learning of SVM regression to compute their +20-minute predictions. Model M1 uses only extracted article terms for its prediction. While no baseline stock price exists within this model. Model M2 uses extracted article terms and the stock price at the time the article was released and Model M3 uses extracted terms and a regress estimate of the +20-minute stock price. And their first conclusion was that model M2, using both article terms and the stock price at the time of article release, had a dominating performance in all three metrics: measures closeness at 0.04261, directional accuracy at 57.1% and simulated trading at a 2.06% return. These results were the direct consequence of this model's ability to capitalize on the article terms and stock price for machine learning. The second conclusion was that proper nouns had the better textual representation performance. While it performed best in 2 of the 3 metrics, directional accuracy at 58.2% and simulated trading at 2.84%, it pulled up short in measures of closeness, 0.04433, as compared to named entities with 0.03407, all p-values < 0.05. However, this subset representation performed better than its parent, noun phrases, in all three metrics. Authors believe that proper nouns can attribute its success to being freer of the term noise used by noun phrases and free of the constraining categories used by named entities (Shcumaker, Robert P.; Chen, Hsinchun;, 2009).

In the research of (Lee, Heeyoung; Mihai, Surdeanu; MacCartney, Bill; Jurafsky, Dan;, 2014), they introduce a system that forecasts companies' stock price changes (UP, DOWN, STAY) in response to financial events reported in 8-K documents. In this wok authors built a corpus that can be used to investigate the importance of text analytics for stock price movement. Our corpus aligns descriptions of financial events reported in 8-K documents

with the corresponding stock prices, which facilitates the development of stock price forecasting systems that combine financial and textual information. Using this corpus, they showed that incorporating textual information is indeed important, especially in the short term (the two days immediately following the event), the experiments indicated that predicting next day's price movement is improved by 10% (relative) if text is considered, the research made several simplifying assumptions and, because of this, should not be considered as a complete trading strategy. However, it does indicate that text carries predictive power for stock price movement.

To date, "traditional" risk management tools have largely neglected one of the largest sources of information, in other words unstructured qualitative words. Therefore, existing risk management approaches are not able to sufficiently capture/predict extreme intraday market movements (at event time) triggered by new information released to the market. However, as the mitigations of intraday market risk is important to many market participants, so (Groth, Sven S.; Jan, Muntermann;, 2010) propose an intraday risk management approach that also makes use of qualitative, unstructured data. This study explores the risk implications of information being newly available to market participants. Four different learners – Naïve Bayes, k- Nearest Neighbor, Neural Network, and Support Vector Machine – have been applied in order to detect patterns in the textual data that could explain increased risk exposure. Two evaluations are presented in order to assess the learning capabilities of the approach in the context of risk management. First "classic" data mining evaluation metrics are applied and, second, a newly developed simulation-based evaluation method is presented. Evaluation results provide strong evidence that unstructured (textual) data represents a valuable source of information also for financial risk management. They show that today's technology is capable of extracting valuable information from corporate disclosures for risk management purposes. Both a "classic" and a newly developed domain-specific simulation-based evaluation confirm the suitability of our approach to identify most critical, in the other words' volatility-enhancing, market events. Therefore, they conclude that intraday market risk exposures can be discovered utilizing text mining techniques. Moreover, they show that more sophisticated classifications methods such as kNN, NNet, or SVM perform better than NB. Taking into

account both classification results and computational efficiency, SVM turn out to be the method of choice for this particular learning task (Groth, Sven S.; Jan, Muntermann;, 2010).

Many previous studies revealed that the news influences stock prices, and a number of studies on stock price prediction have been made using actual news articles. However, (Kim, Yoosin; Jeong, Seung Ryul; Ghani, Imran;, 2014) have conducted a novel attempt to compile a stock domain specific sentimental word dictionary from the news as unstructured big data, to analyze 'sentiment' using that dictionary, and to mine opinions in order to predict stock price fluctuations (up/down movements). Authors introduce a method of mining text opinions to analyze Korean language news in order to predict rises and falls on KOSPI (Korea Composite Stock Price Index). They have built a stock domain specific dictionary via the Natural Language Processing (NLP) of 78,216 news articles take from two different style media and showed the sentimental words, news sentiments and opinions calculated using the dictionary. The aim of their experiment was to determine whether there is a correlation between the news sentiment and the rise of fall of stock prices, and whether the results of prediction differ according to each media. For the experiment, they separated three groups, media H, M and H+M, from the data set. Media M news' opinions showed the best prediction accuracy of 65.2% at the critical value of 0.22. H news' opinions, meanwhile, showed60.3% prediction accuracy at the critical value of 0.19. The difference in accuracy between two media is 5%, which means that M's news, compared to H's news, could predict with greater accuracy any rises and falls in stock prices. On the other hand, M+H news' opinions showed 60.1% prediction accuracy at the critical value of 0.11. It is understood that the mixes analysis of news articles of the two somewhat contrasting media made the classification of opinions unclear, thereby lowering accuracy. By conducting a stock prediction experiment, they found that the opinions extracted from the news could be useful in accurately predicting stock market movements and F1 features. Furthermore, they recognized that the media have their own characteristics, so the accuracy pf predicting stock market price movements could differ depending on the media (Kim, Yoosin; Jeong, Seung Ryul; Ghani, Imran;, 2014).

Machine learning paradigms are increasingly being used along with text categorization for knowledge discovery in unstructured data. (Vakeel, Khadija; Dey, Shubhamoy;, 2014), in their paper used machine learning techniques to textually analyze

online news articles. Their main hypothesis was whether in addition to technical approaches, the information in news articles can influence stock prices. Indian elections, an exceptionally dynamic period for the Indian economy us chosen to form two corpuses: pre and post-election. These corpuses comprised online news articles collected over a period of 4 months, they have collected data from 2 online newspapers of India. They have collected the data from web crawler. The rise and fall of the national index during the corresponding period have been recorded every hour then the news has been classified on the basis of 'rise' or 'fall' of the index into 2 classes "Positive" and "Negative" class. They had hypothesized that the information content of news articles is absorbed by the market and reflected in the price of stocks. Authors, through their analysis of 3253 articles using Support Vector Machine, found that there is a significant impact of the information content of news articles on the stock prices. In both cases, pre-election and post-election, the classification accuracy w 62.04% and 63.98%, which is significantly more than a chance prediction. Form this they conclude that apart from technical analysis, textual analysis of news can also be used to predict the movements of the Stock Market.

With the advent of electronic and online news sources, analysts must deal with enormous amounts of real-time, unstructured streaming data. In the research paper of (Anurag, Nagar; Hahsler, Michael;, 2012), they present an automated text-mining based approach to aggregate news stories from diverse sources and create a News Corpus. The corpus is filtered down to relevant sentences and analyzed using Natural Language Processing (NLP) techniques. A sentiment metric, called NewsSentiment, utilizing the count of positive and negative polarity words is proposed as a measure of the sentiment of the overall news corpus. In this research paper, authors use the open source software R to build a news engine for gathering and aggregating news items from various financial sources. The software has the capability to gather news in real time and analyze it for key financial terms and phrases. The phrases are analyzed for positive/negative affect and polarity by comparing them with publicly available lexicons and dictionaries. The proposed algorithm filters out irrelevant sentences and noise from news stories and created a text corpus from only those sentences which are relevant to the stock in question. For breaking a story down to sentences, they again used R and various Natural Language Processing (NLP) tools available therein, such as the package NLP. The method can work at different granularity

levels – such as sentences, headlines, paragraphs, or even the entire news article. After extracting the relevant instances, they identify the key words contained in them and match them against available sources of positive or negative sentiment terms. They have used the Multi-Perspective Questions Answering (MPQA) Subjectivity Lexicon and list of sentiment words from the R package to create a database of sentiment carrying words. An instance is classified as positive if the count of positive words is greater than or equal to the count of negative words. Similarly, an instance is negative if the count of negative words is greater than the count of positive words. Score of a corpus is defined as the ratio of positive instances to the total number of instances. The method can thus clearly detect the direction of the market sentiment, which bears a close association with the direction of the actual stock price movement. NewsSentiment scores had a very strong correlation with the actual stock price variations and can pick up the direction of the market sentiment (Anurag, Nagar; Hahsler, Michael;, 2012).

Technological advancements that cultivate vibrant creation, sharing, and collaboration among Web users, investors can rapidly obtain more valuable and timely information. Meanwhile, the adaption of user engagement in media effectively magnifies the information in the news. With such rapid information influx, investor decisions tend to be influenced by peer and public emotions. (Li, Qing; Wang, Tiejun; Li, Ping; Gong, Qixu; Chen, Yuanzhu;, 2014) propose a quantitative media-aware trading strategy to investigate the media impact on stock markets. To identify the essential reasons how Web media influences the stock market, authors studied the internal functions of sentiment, firm characteristics, and news content regarding the relation of Web media to stock markets. They found that sentiment influence originates from two sources: the sentiment in a news article captured by the financial-specific sentiment words and revealed public feelings via posting and comments on financial discussion boards. Such sentiment in Web media cause investors emotions to fluctuate and intervene in their decision making. This result corroborates the basic assertion in behavioral finance that investors are sentimental. Other enlightened findings include that Web media influence on stock varies by news content and firm characteristics. Specifically, stocks are sensitive to news articles on restricting and earnings issues. The firms strongly involved with public interests or daily life, especially with utility supply, real estate, social service, and wholesale and retail trade, are more

predictable in stock markets. However, the article origin, whether official, leaked, or rumored, may have different influences on investors (Li, Qing; Wang, Tiejun; Li, Ping; Gong, Qixu; Chen, Yuanzhu;, 2014).

People tend to buy a company with good reputation. One way to know company's reputation is by seeing relationship between company and customer. The explosion of social media usage forces many companies to create their official account in social media in order to keep in touch with their customer. This make customer can express their opinion about products easily. One of the social media that commonly used by company is Twitter. (Cakra, Yahya Eru; Trisedya, Bayu Distiawan;, 2015) purpose on their research paper is to predict the Indonesian stock market using simple sentiment analysis. In the research are used to kinds of data. Stock prices of several companies in Indonesia and data contained opinions about certain products produced by the companies mentioned. Opinions was shared via Twitter. Companies that were being chosen were companies with fluctuating stock prices and its products already popular in Indonesia. There are many algorithms that can be used in supervised classification. In this research the used algorithms are Support Vector Machine (SVM), Naïve Bayes, Decision Tree, Random Forest and Neural Network (with single layer perception). There are three things that were predicted in this research paper. They are price fluctuation prediction, margin percentage prediction and price prediction. Price fluctuation prediction used classification by supervised learning method. Margin Percentage and price prediction used linear regression, since the value that was being predicted was not categorized. The gathered Tweets were being classified into three classes (positive, negative and neutral). Each tweet was examined, whether it contains positive lexicon, negative lexicon or none of both. From the results of this research authors conclude that created sentiment analysis model using Random Forest algorithm can classify tweet data with 60.39% accuracy, and the one with Naïve Bayes algorithm can classify tweet data with 56.50% accuracy. In price fluctuation prediction, created models can predict whether the upcoming price will go up or down with the highest accuracy of 67.37% for tweets data classified by Naïve Bayes and 66.34% for tweets data classified by Random Forest. In margin percentage prediction, created models have R^2 value which close to 0, it means that created models fitted only few data. In price prediction, created model have R^2 value which close to 1, it means that created models fitted lots of data. The highest R^2 value retrieved

from model with previous price as the feature (Cakra, Yahya Eru; Trisedya, Bayu Distiawan;, 2015).

A research paper conducted by (Bollen, Johan; Mao, Huina; Zeng, Xiao Jun;, 2010) analyze the text content of daily Twitter feeds, they use two tools to measure variations in the public mood from tweets submitted to the twitter service from February 28, 2008 to December 19, 2008. The first tool, OpinionFinder, analyses the text content of tweets submitted on a given day to provide a positive vs. negative daily time series of public mood. The second tool, Google-Profile of Mood States (GPOMS), similarly analyses the text content of tweets to generate a six-dimensional daily time series of public mood to provide a more detailed view of changes in public along a variety of different mood dimensions, the 6 dimensions mood are: calm, alert, vital, kind and happy. The resulting public mood time series are correlated to the Dow Jones Industrial Average (DIJA) to assess their ability to predict changes in the DIJA over time. Their results indicate that the prediction accuracy of standard stock market prediction model is significantly improved when certain mood dimensions are included, but not others. They find an accuracy of 87.6% in predicting the daily up and down changes in the closing value of DIJA and a reduction of the Mean Average Percentage Error by more than 6%. In particular variations along the public mood dimensions of Calm and Happiness as measured by GPOMS seemed to have a predictive effect, but not general happiness as measured by the OpinionFinder tool.

According to investment theories, investors' behaviors will influence the stock market, and the way people invest their money is based on the history trend and information they hold. Due to the serious feature sparse problem in tweets and unreliability of using average sentiment score to indicate one day's sentiment, a hybrid feature selection method is proposed in the work done by (Meesad , Phayung; Li, Jiajia;, 2014), to increase the quality of selected features. Instead of applying sentiment analysis to add the sentiment related features, this paper used SentiWordNet to give the weight to the selected features and then used weighted example sets to train a support vector machine (SVM) classification model. History tweets were collected from Topsy.com which is real-time search engine for social comments. They were "$AAPL" keyword focused and related with apple Inc. news. There are 4,622 tweets gathered from July 1st, 2013 to May 30th, 2014. The stock data were collected from yahoo finance. Cross-validation is an effective method to get the estimate

performance by iterating train and test process together. The mostly used are 10-fold cross validation and Leave-one-out cross validation. For 10-fold cross validation, it separates the data sets into 10 folds, left one-fold for test and others for training, then iterate 10 times. For Leave-one-out cross validation, one observation is taken as the validation set and the remaining observations are taken as the training set. The experiments were conducted using both of them. In conclusion SVM linear algorithm based on leave-one-out cross validation yields the best performance of 90,34% (Matsubara, Takashi; Akita, Ryo; Uehera, Kuniaki, 2018).

Economic analysis indicates a relationship between consumer sentiment and stock price movements. In a study done by (Vu, Tien Thanh; Chang, Shu; Ha, Quang Thuy; Collier, Nigel;, 2012), they harness features from Twitter messages to capture public mood related to four Tech companies for predicting the daily up and down price movements of these companies' NASDAQ stocks. In the study is proposed a novel model combining features namely positive and negative sentiment, consumer confidence in the product with respect to 'bullish' or 'bearish' lexicon and three pervious stock market movement days. The features are employed in a Decision Tree classifier using cross-fold validation to yield accuracies of 82.93%, 80.49%, 75.61% and 75.00% in predicting the daily up and down changes of Apple, Google, Microsoft and Amazon stocks respectively. Data containing 5,001,460 daily Tweets was crawled by using Twitter online streaming API from 1st April 2011 to 31st May 2011.

Predicting the future has always been an adventurous and attractive task for the probing individuals. This kind of prediction becomes more fascinating when it involves money and risk like predicting the Stock Market. The main objective of the research paper examined by (Usmani, Mehak; Adil, Syed Hasan; Raza, Kamran; Azhar Ali, Syed Saad;, 2016), was to predict the market performance of Karachi Stock Exchange (KSE) on day closing using different machine learning techniques. The prediction model used different attributes as an input and predicts market as Positive and Negative. The attributes used in the model includes: Oil rates, Gold and Silver rates, Interest rate, Foreign Exchange (FEX) rate, News and social media feed. The old statistical techniques including Simple Moving Average (SMA) and Autoregressive Integrated Moving Average (ARIMA) are also used as input. The machine learning techniques including Single Layer Perceptron (SLP), Multi-Layer Perceptron (MLP),

Radial Basis Functions (RBF) and Support Vector Machine (SVM) are compared. The data used in this study spread over three months, from September 2015 to January 2016. News and twitter data were available in the form of feed which was processed using text mining techniques. Single Layer Perceptron model trained by training set and tested on the different data set. The model gave about 60% accurate results, the model was unable to predict the Positive class well. When the SLP algorithm was tested on the same data set that was used for training, it gave 83% accuracy. Similarly, MLP algorithm was applied on data set first on training and then on testing, the model gave 77% correct results when verified on test set and 67% on training set. RBF algorithm gave 63% results when demonstrated on test set and 61% when demonstrated on training set. Applying SVM algorithm on the training and test set gives different results, the algorithm produced 100% accuracy on training set but 60% on the test set. The algorithm MLP performed best as compared to other techniques, and the oil rate attribute was found to be most relevant to market performance. The results suggest that the performance of KSE-100 index can be predicted with machine learning techniques (Usmani, Mehak; Adil, Syed Hasan; Raza, Kamran; Azhar Ali, Syed Saad;, 2016).

Business and financial news bring us the latest information about the stock market. Studies have shown that business and financial news have strong correlation with future stock performance. Therefore, extracting sentiments and opinions from business and financial news is useful as it may assist in the stock price prediction. In the research paper conducted by (Im, Tan Li; San, Phang Wai; On, Chin Kim; Alfred, Rayner; Anthony, Patricia;, 2013), they present a sentiment analyzer for financial news articles using lexicon-based approach. In this study, the lexicon is manually created. They find the common words used in financial news, classify each word into one of the two sentiment classes (positive or negative), and the final step is to add it to the lexicon. The targeted news articles were obtained from Malaysia's local online newspapers, since their focus is on business and financial domain only business and financial news articles were collected. They conducted two experiments for the lexicon-based sentiment analysis algorithm, one with the stemming algorithm and the other one without using stemming algorithm then they compared the result obtained from these two experiments. The positive predictive value for non-stemming and stemming lexicon-based approach recorded 76.7% and 82.4% respectively. In

addition, the negative predictive value for non-stemming lexicon-based approach gives 68.3% whereas stemming lexicon-based approach recorded 73.2%. The lexicon-based approach sentiment analysis algorithm performs better in assigning positive news articles compared to negative news articles. The true positive rate for non-stemming lexicon-based approach recorded 83.61% whereas the true negative rate recorded only 58.11%, which is much lower than the true positive rate. According to the authors this is because of the nature of the news articles itself where reporters tend to write something positive in news even though the news is negative. It appears that there are more positive words that negative words in negative news articles. Thus, the algorithm can be considered as positive bias. The overall accuracy of the stemming lexicon-based approach is 79.1% which is higher approach that recorded a 74% accuracy rate. Based on the results, authors conclude that utilizing stemming algorithm produced higher accuracy for all the measurements. Stemming also increases the chances for each token in the news article to be matched with the words in the lexicon. As a direct consequence, it increases both the accuracy of predicting positive and negative news articles (Im, Tan Li; San, Phang Wai; On, Chin Kim; Alfred, Rayner; Anthony, Patricia;, 2013).

Investors make decisions according to various factors, including consumer price index, price earnings ratio, and miscellaneous events reported in newspapers. In order to assist their decisions in a timely manner, many automatic ways to analyze those information have been proposed in the last decade. However, many of them used either numerical or textual information, but not both for a single company. In the study done by (Akita, Yoshihara, Takashi, & Kuniaki, 2016), they propose an approach that converts newspaper articles into their distributed representation via Paragraph Vector and models the temporal effects of past events on opening prices about multiple companies with Long Short-Term Memory (LSTM). Their approach predicts 10 company's closing stock prices by using LSTM, which can memorize the previous timesteps due to its architecture. They use multiple companies to learn the correlations between companies. For example, an event like "Nisan recalls…" might make Nissan's stock price decrease while making the stock price of Toyota (another company in the same industry) to increase at the same time. They decided that the number of predicting companies to be 10 due to computation time constraint. To evaluate the effectiveness of the approach, they conducted experiments on market simulation.

Experimental results showed that distributed representation of textual information are better that the numerical-data-only methods and Bag-of-Words based methods, LSTM was capable of capturing time series influence of input data than other models, and considering the companies in the same industry was effective for stock price prediction (Akita, Yoshihara, Takashi, & Kuniaki, 2016).

Investors read all publicly available information, such as news articles, to learn about events both in the past and the future, then react via the corresponding financial commodity. The price of the financial commodity then responds to these events rather quickly. Based on this relationship, machine learning approaches have been directed to predict price movements by employing classifiers. In other words, these approaches build a model describing the indirect influence that news articles have on the prices, however, a large number of news articles are published daily, and each article contains rich information regardless of whether or not it is related to the financial market. More specifically, machine learning approaches are given numerous news articles as explanatory variables that explain the limited number of price movements, such approaches are prone to overfitting to price movements used for parameter adjustment and have not been generalized to accurate prediction of future price movements. To counter this problem, a research paper is conducted by (Matsubara, Takashi; Akita, Ryo; Uehera, Kuniaki, 2018), they propose a deep neural generative model (DGM) of news articles to predict the price movements, according to the authors knowledge, this is the first time a DGM has been used to tackle such a problem. The DGM is an implementation of generative model on deep neural networks. Their proposed model generates news articles embedded to vectors, given the assumption of future price movement as a condition. Thanks to the nature of generative modeling, their proposed model is expected to have a lower risk of overfitting to past price movements for training. Authors evaluate the proposed model using historical datasets of Nikkei 225 (Nikkei Stock Average) and Standard & Poor's 500 Stock Index, as well as related news articles. In the Nikkei with DGM there is achieved a prediction performance accuracy of 56.4% while with SVM and MLP the achieved prediction performance accuracy is 49.2% and 50.4% respectively. In the S&P with the DGM there is achieved a prediction accuracy of 61.1% while with SVM and MLP the achieved prediction performance accuracy is 59.0% and 55.3% respectively. The experimental results demonstrate that their proposed model better

predicts the movements of these stock indices versus two keys conventionally baseline methods, in other words, support vector machines (SVMs) and multilayer perceptron's (MLPs). Results of a simplified market simulation also demonstrate that the proposed model by the authors is more capable of making profit versus baseline methods (Matsubara, Takashi; Akita, Ryo; Uehera, Kuniaki, 2018).

With the development of information and communication technologies, financial text-mining techniques recently have drawn many attentions from both professional and individual investors. Financial text-mining is computerized methods to extract valuable and useful information for investments automatically from vast textual data such as news articles, SNS, and tweets. The use of an appropriate dictionary, especially a polarity dictionary of positive and negative expressions in financial contexts is central to this area. The purpose of (Ito, Tomoki; Izumi, Kiyoshi; Tsubouchi, Kota; Yamashita, Tatsuo;, 2016) , in their research paper is to give the financial term whose polarity is unknown positive-negative score, and to make the feature vector of a document useful for predicting stock price trends. They propose a new technology for giving a word positive-negative polarity score using existing polarity dictionary. First, they assigned a numerical vector to a word appeared in financial news documents using word2vec algorithm and defined the feature vector of the documents. The certain percentage of those words have polarity scores defined by financial professionals. Then, in order to acquire polarity scores of news words, they analyzed the relationships between the feature values of the words and the stock price trends, and sentiment score which can be evaluated from the textual data of Yahoo! Finance board. In the machine learning process, the polarity scores of words in the dictionary propagated to the news words, which are not in the dictionary. Authors compared the prediction power of the feature value made from their method against those of the feature value made from conventional methods. They tested it using economic news documents that were distributed via Thomson-Reuters news service, the individual stock price data, and the sentiment score data of the Yahoo! Finance board. As a result, their II algorithm performed well in 10-fold cross validation. It also had better predictability than other traditional classification methods. These results indicate that the method proposed by the authors for giving a polarity score is valid (Ito, Tomoki; Izumi, Kiyoshi; Tsubouchi, Kota; Yamashita, Tatsuo;, 2016).

Stock market prediction is always a challenging task because it is highly dynamic. Several methods have been deployed to forecast the future direction of the stock market, in their work (Huyhnh, Huy D.; Dang, Minh L.; Duong, Duc;, 2017), they introduce a new prediction model that depend on Bidirectional Gated Recurrent Unit (BGRU). Their predictive model relies on both online financial news and historical stock prices data to predict the stock movements in the future. The experiment used financial news from Reuters and Bloomberg between October 2006 and November 2013. Reuters and Bloomberg dataset contained approximately 106,521 news and 447,145 news respectively. They also used the public price data from Yahoo Finance from 2006 to 2013 which match the time period of the financial news to conduct the experiments on forecasting (S&P500) index and its individual stocks. Authors introduce the BGRU networks. The idea behind BGRU is the idea of bidirectional recurrent neural network. The BGRU presents each training sequence forwards and backwards to two separate recurrent nets, both of which are connected to the same output layer. They used three companies selected in different sectors for evaluating the effectiveness of the approach on the aspect of individual stock prediction. They chose Google Inc. in Information Technology, Wal-Mart Stores in Consumer Staples and Boeing Company in Industrial. They extracted all the news, regard to the three mentioned companies in Reuters News. Authors compared BGRU model with the standard Long-short-term-memory (LSTM) and Gated Recurrent Unit (GRU). Experimental results have shown that authors proposed method was simple but very effective, which could significantly improve the stock prediction accuracy on a standard financial database over the other systems which only used the historical price movements. Their model accuracy achieves nearly 60% in S&P index prediction whereas the individual stock prediction is over 65% (Huyhnh, Huy D.; Dang, Minh L.; Duong, Duc;, 2017).

As news events affect human decisions and the volatility of stock prices is influenced by human trading, it is reasonable to say that events can influence the stock market. Accurate extraction of events from financial news may play an important role in stock market prediction. However, previous work represents news documents mainly using simple features, such as bag-of-words, noun phrases, and named entities. With these unstructured features, it is difficult to capture key events embedded in financial news, and even more difficult to model the impact of events on stock market prediction. For example,

representing the event “Apple has sued Samsung Electronics for copying ‘the look and feel’ of its iPad tablet and iPhone smartphone.” using term-level features {“Apple”, “sued”, “Samsung”, “Electronics”, “copying”, …} alone, it can be difficult to accurately predict the stock price movements of Apple Inc. and Samsung Inc., respectively, as the unstructured terms cannot indicate the actor and object of the event. In a study done by (Ding, Xiao; Zhang, Yue; Liu, Ting; Duan, Junwen;, 2014), they propose using structured information to represent events, and develop a prediction model to analyze the relationship between events and the stock market. The problem is important because it provides insights into understanding the underlying mechanisms of the influence of events on the stock market. There are two main challenges to this method. On the one hand, how to obtain structured event information from large-scale news streams is a challenging problem. Authors after identifying the aforementioned challenges and problems they propose to apply Open Information Extraction techniques (Open IE), which do not require predefined event types or manually labeled corpora. Subsequently, two ontologies (in other words VerbNet and WordNet) are used to generalize structured event features in order to reduce their sparseness. On the other hand, the problem of accurately predicting stock price movement using structured events is challenging, since events and the stock market can have complex relations, which can be influenced by hidden factors. In addition to the commonly used linear models, they build a deep neural network model, which takes structured events as input and learn the potential relationships between events and the stock market. Experiments on large-scale financial news datasets from Reuters (106,521 documents) and Bloomberg (447,145 documents) show that events are better features for stock market prediction than bag-of-words. In addition, deep neural networks achieved better performance than linear models. The accuracy of S&P 500 index prediction by their approach outperforms previous systems that they compared in their research paper with 58.94% accuracy, and the accuracy of individual stock prediction can be over 70% on the large-scale data. Authors claim that their system can be regarded as one step towards building an expert system that exploits rich knowledge for stock market prediction. Their results are helpful for automatically mining stock price related news events, and for improving the accuracy of algorithm trading systems (Ding, Xiao; Zhang, Yue; Liu, Ting; Duan, Junwen;, 2014).

Recent advances in computing power and Natural Language Processing (NLP) technology enables more accurate models of events with structures. Using open information extraction (Open IE) to obtain structured events representations, (Ding, Xiao; Zhang, Yue; Liu, Ting; Duan, Junwen;, 2014) find that the actor and object of events can be better captured. One disadvantage of structured representation of events is that they lead to increased sparsity, which potentially limits the predictive power. Another work done by (Ding, Xiao; Zhang, Yue; Duan, Junwen;, 2015), they propose to address this issue by representing structured events using event embeddings, which are dense vectors. For the predictive model, they propose to use deep learning to capture the influence of news events over a history that is a longer than a day. They model long-term events as events over the past month, mid-term events as events over the past week, and short-term events as events on the past day of the stock price change, the prediction model learns the effect of these three different time spans on stock prices based on the framework of a CNN. Authors use financial news from Reuters and Bloomberg over the period from October 2006 to November 2013, released by (Ding, Xiao; Zhang, Yue; Liu, Ting; Duan, Junwen;, 2014). In this paper, authors extract events only from news titles, they conduct the experiments on predicting the Standard & Poor's 500 stock (S&P 500) index and its individual stocks, obtaining indices and prices from Yahoo Finance. Experiments on large-scale financial news datasets from Reuters and Bloomberg show that event embeddings can effectively address the problem of event sparsity. In addition, the CNN model gives significant improvement by using longer-term event history. The accuracies of both S&P 500 index prediction and individual stock prediction by their approach outperform state-of-the-art baseline methods by nearly 6%. The achieved accuracy for S&P 500 index is 64.21% and the accuracy for the Individual Stock Prediction is 65.48% (Ding, Xiao; Zhang, Yue; Duan, Junwen;, 2015).

Numerous studies have attempted to examine whether the stock price forecasting through text mining technology and machine learning could lead to abnormal returns. However, few of them involved the discussion on whether using different media could affect forecasting results. Financial sentiment analysis is an important research area of financial technology (FinTech). This research paper by (Day, Min Yuh; Lee, Chia Chou;, 2016), focuses on investigating the influence of using different financial resources to investment and how to improve the accuracy of forecasting through deep learning. They designed a

web crawler with web mining method to capture information from 4 electronic news providers. The grabbed financial information is semi-structured or unstructured information. The study adopted deep learning as the machine learning method to perform classification forecasting of the experimental data. Deep learning is a branch of the machine learning. They wrote a web crawler in Python language, grabbing respectively from NowNews, AppleDaily, LTN and MoneyDJ, the four electronic news providers, choosing 18 public companies, news released for 1 year, with a total of 8,472 articles used in this experiment. The output items of the prediction model are rise, fall and remain. The experimental results showed various financial resources have significantly different effects to investors and their investments, while the accuracy of news categorization could be improved through deep learning (Day, Min Yuh; Lee, Chia Chou;, 2016).

In the work of (Abdullah, Sheikh Shaugat; Rahaman, Mohammad Saiedur; Rahman, Mohammad Saidur;, 2013), they proposed and new data processing framework that takes text from different sources as input where the source may be authentic or unauthentic. For authentic sources like company web source or stock exchange database, the information is retrieved using a text analyzer algorithm and stored in database. However, information from unauthentic sources are passed to a natural language processing tool to extract as much information as possible and therefore set a rank or weight on that information by comparing with historical data. Though they used the data of Dhaka Stock Exchange (DSE) in their analysis, according to them this technique is applicable on any other stock exchange as well. Any news can lead to decision by comparing it with chronological data which may or may not be true depending on other factors like price trends.

The ability to predict stock trend is crucial for stock investors. Using daily time series data, one is able to predict the trend with the help of simple moving average technique. Daily news, particularly financial news has a great role in deciding stock trend. Each news has a sentiment value classified into positive, negative, and neutral sentiment that directly affects whether the trend goes up or down. In the research paper worked by (Lauren, Stefan; Harlili, Dra;, 2014), artificial neural network is used as a model to combine simple moving average and financial news' sentiment to predict stock trend more responsively. To achieve this they retrieved, processed and classified daily financial news to get the sentiment value. They used both stock and news data from May 2013 to May 2014. The

stock in this case is JSKE (Jakarta Composite Exchange) and the news are retrieved only on market and corporate category. Authors claimed that artificial neural network does the combination of both components very well. However, only several parameters on stock trend prediction and simple moving average combination give the best result and thus improve stock trend prediction to be more responsive (Lauren, Stefan; Harlili, Dra;, 2014).

Stock market collapses can cause significant loss of investor wealth, and deeply affect listed corporations and the whole economy. The objectives of the study examined by (Wang, Wanbin Walter; Ho, Kin Yip; Liu, Wai-Man Raymond; Wang, Kun Tracy;, 2013), were to investigate the ability of news events to predict stock price jumps and the occurrence of stock market collapses. To achieve these objectives, they first used causality tests to find factors affecting stock price movements, and then they built an artificial network to predict stock price jumps using these factors. The Granger causality test has been used to find factors causing stock price changes. They used daily closing prices (2004-2012) of the Dow Jones Price Index. The news data used in this study are provided by RavenPack Inc., a leading provider of news analytic data. The data in RavenPack database are derived from all news articles and press release that appeared in the Dow Jones newswire, and there are 20,354,107 news articles in RavenPack database. In this study, the PNN (probabilistic neural network) is implemented using the MATLAB Neural Network Toolbox, where the network structures were set to the default. The Granger causality tests showed that news sentiment is the Granger cause of stock price change, and stock price change is the Granger cause of news volume and news sentiment. The results of their PNN forecast model is promising. They evaluated the model through sensitivity, specificity and predication rate, and they had these results 37.04%, 89.25% and 84.66% respectively. Authors claimed that there is much work need to be done to improve the prediction accuracy of their model, especially the sensitivity of the predictions (Wang, Wanbin Walter; Ho, Kin Yip; Liu, Wai-Man Raymond; Wang, Kun Tracy;, 2013).

Financial news articles are believed to have impacts on stock price return. In the research paper worked by (Li, Xiadong; Xie, Haoran; Chen, Li; Wang, Jianping; Deng, Xiaotie;, 2014), they analyzed the news impact from sentiment dimensions. They first implement a generic stock price prediction framework. Secondly, they used Harvard psychological dictionary and Loughran-McDonald financial sentiment dictionary to construct the

sentiment dimensions. News articles are then quantitatively projected onto sentiment space. They evaluate the models' prediction accuracy and empirically compare their performance at different market levels. To make it a fair comparison, instance labelling method is rigorously discussed and tested, and the threshold *th* is carefully chosen. Experiments, are conducted on five years historical Hong Kong Stock Exchange prices and news articles, showed that sentiment analysis does help improve the prediction accuracy. At stock, sector and index levels, the models with sentiment analysis outperform the bag-of-words model in both validation and independent testing data sets. Simply focusing on positive and negative dimensions could not bring useful predictions. The models which used sentiment polarity do not perform well in all the tests. There is a minor difference between the models using two different sentiment dictionaries.

Now days lots of research ongoing to predict the price. This study (Patel, Hiral R.; Parikh, Satyen;, 2016), considers news impact as semantic analysis and as a technical view stock prices and index is measured. This paper showed that short-term stock price movements can be predicted using financial news article. Given a stock price time series, for each interval classify the price movements as "up", "down" or "unchanged" relative to the volatility of the stock and the change in a relevant index. Each article in a training set of news articles is then labelled Negative, Positive or Neutral according to the contextual semantic analysis of published news and movement of the associated stock in a time interval surrounding publication of the article. The proposed model based on dataset fetched from selective news article and classifies the news using combine approach of Bag of Words, Noun Phrase and TF-IDF. The text mining is used for to convert textual information in to machine readable approach. This paper collects news from pre-selected financial news. The Bag of Words approach is basically process of to maintain tokenization which represents tokens of documents also implies stop word removal and stemming. Noun phrases are also applied for to remove unused words and noun from the text. Another approach is Term Frequency – Inverse of the Document Frequency (TF-IDF). Semantic analysis is used to identify the news impact to predict the stock price prediction. This fusion approach achieved 72% accuracy.

Chapter 3. Research Methodology

In this thesis it is worked on analyzing data, concretely news articles and stock prices to make future prediction about stock price direction, to achieve this data is crucial in this work. Many steps are conducted to achieve the aim of this research starting from identifying the news sources and the targeted companies, in this chapter the steps that are undertaken in this thesis are presented, but before that there is an explanation in details about text mining and about how is the sentiment analysis done using text mining approach.

Text Mining

In this study I will analyze financial news so text content, and text mining is considered a set of methodologies to extract useful information from text content. For this purpose, it is necessary to transform unstructured text content into a structured format readable by algorithms (Beckmann, 2017).

Text Mining may be defined as the process of examining data to gather valuable information. Text mining, also known as text data mining involves algorithms of data mining, machine learning, statistics, and natural language processing, attempts to extract high quality, useful information from unstructured formats (Bose, 2018).

The main activities of a text mining are: entity extraction, taxonomy extraction, sentiment analysis, document summarization, text categorization, text clustering, entity relationship, and visualization. Most part of these activities relies on data mining algorithms, but these algorithms are not able to deal directly with unstructured data, as they need a structured format, normally in a matrix shape. A text mining system normally has the architecture depicted in Figure below (see 4 Figure).

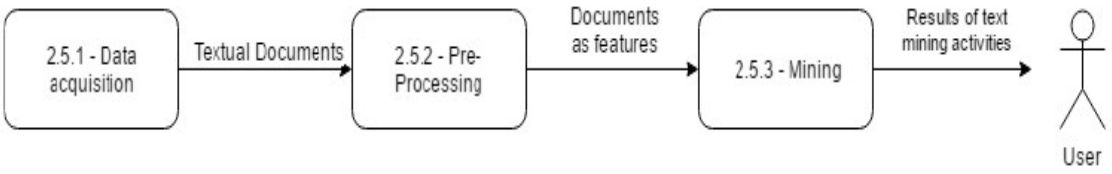


Figure 4 Text mining system architecture (Beckmann, 2017)

Data Acquisition

A good and reliable source of data is the key for building good text mining models. According to DHAKA (2013), a set of textual documents, known as corpus, can be collected from databases, web crawler systems, textual files from file systems (e.g. manuscripts, journalists, digitalized documents, etc.), or any other automated system designed to collect unstructured data from resources like news reports, social media (platform/sites), emails, etc. (Vale, 2018).

Pre- Processing

To transform unstructured data in features, the textual documents must be parsed into simple words, with the blank spaces and punctuation used to distinguish and separate the words. This process is also known as tokenization. A list will all existing words and the respective number of occurrences in the corpus can also be generated during this phase. After this, the words or terms are selected to form features. In this context, a feature can be understood as a value, and the feature name is the meaning of this value. Features can represent a word, as sequence of words or *n*-grams, which consists in a series of consecutive *n* words, types of entities (e.g. company names, stock symbols), quantitative value (e.g. stock prices, date, time), syntactical structures like noun-phrases and part-of-speech, etc. (Beckmann, 2017).

Not all the words carry information in the textual content. The stop words are terms with low importance for information retrieval (normally prepositions), and its removal is recommended. Terms with occurrence per document lower or above a specific threshold are also recommended for removal, because a few numbers of words have no representation, and do not carry significant information in the document. The same applies to repeated and abundant words. The min/max thresholds must be adjusted according to

the problem under study, but normally values lower than ~5%, or greater than ~90% are reported in the literature (Beckmann, 2017).

The use of stemming reduces the number of words, by replacing a word to its base or stem (Lovins, 1968), (Porter, 1980), e.g., fruit = fructify, fruity, fruitful. The use of stemming requires caution must be adjusted according to the problem under study, as it may remove important information existing in the original words (Beckmann, 2017).

The most common type of features representation is the Bag of words (BOW), first mentioned by (Harris, 1954) and still predominant technique nowadays. A BOW is basically a matrix, where each document is represented as a vector row, and the features (normally words) as the columns of this matrix. The columns of this matrix must contain not only the existing terms in the document, but also all existing terms in the corpus. Not all the document share the same terms, then the missing terms in a document are filled with zero or null, which can result in a sparse matrix (Beckmann, 2017).

The feature values can be represented as categorical, binary (i.e., existence, nonexistence of a feature in a document), and numerical values. The numerical values can contain any integer or continuous value extracted from the textual content (e.g., prices, counting, etc.), or some measurement or weighting regarding that feature. For example, the Term Occurrence (TO), is the number of times a term occurs in a document, Term Frequency (TF) is the TO divided by the total number of terms in the document, since every document has a different length, it is possible that a term would appear many more times in a ling documents than shorter ones, then the division is a way of normalization. (Beckmann, 2017)

Mining

Once the data is prepared to be processed, two approaches can be taken to analyze the data: unsupervised learning and supervised learning. In unsupervised learning techniques can be used to group similar documents, identify common rules, taxonomies and sentiments. These groups are then labeled and used in supervised learning (classification and regression) and recommendation system (Vale, 2018).

Sentiment Analysis of News Articles using text mining approaches

Sentiment analysis is crucial when working with text news and there are 5 general steps to analyze sentiment data and here's the graphical representation of the methodology to do the same (see 5 Figure).

Figure 5. 5 Steps to analyze sentiment data (Shankhdhar, 2019)

- **Data Collection**

Data collection is a process of collecting information from all the relevant sources to find answers to the research problem, test the hypothesis and evaluate the outcomes. Data collection methods can be divided into two categories: secondary methods of data collection and primary methods of data collection. Secondary data is a type of data that has already been published in books, newspapers, magazines, journals, online portals etc. There is an abundance of data available in these sources about your research area in business studies, almost regardless of the nature of the research area. Therefore, application of appropriate set of criteria to select secondary data to be used in the study plays an important role in terms of increasing the levels of research validity and reliability. Primary data collection methods can be divided into two groups: quantitative and qualitative (Dudovskiy, 2018).

Quantitative data collection methods are based in mathematical calculations in various formats. Methods of quantitative data collection and analysis include questionnaires with closed-ended questions, methods of correlation and regression, mean, mode and median and others (Dudovskiy, 2018).

Qualitative research methods, on the contrary, do not involve numbers or mathematical calculations. Qualitative research is closely associated with words, sounds, feeling, emotions, colors and other elements that are non-quantifiable (Dudovskiy, 2018).

Data for sentiment analysis generally is collected from various news articles or social network sites like Facebook and Twitter. Manual analysis of sentiment data is virtually

impossible. Therefore, special programming languages like Python or 'R' are used to process and analyze the data.

- **Text Preparation**

Text preparation is nothing but filtering the extracted data before analysis. It includes identifying and eliminating non-textual content and content that is irrelevant to the area of study from the data (Shankhdhar, 2019).

- **Sentiment Detection**

Nowadays the interesting in investigating social media data has been increasing because it offers a platform that allows users to express their opinions towards things. Sentiment detection is a significant process of text analysis which the primary target is detected sentiment within a particular text. The notion of sentiment detection was introduced first by [9] that date back to 1998. Sentiment detection has applied to a variety of social media, especially to reviews of products or services, tweets, etc. This section is focusing on existing research on sentiment detection overall (Salma Al-Asmari; Mohammed Dahab, 2017).

- **Sentiment Classification**

Sentiment classification is a special task of text classification whose objective is to classify a text according to the sentimental polarities of opinions it contains (Shoushan Li; Sophia Yat Mei Lee; Ying Chen, Chu-Ren Huang; Guodong Zhou, 2002).

Sentiments can be broadly classified into two groups, positive and negative.

- **Presentation of Output**

The main idea of sentiment analysis is to convert unstructured text into meaningful information.

Types of classification algorithms in Machine Learning

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation (Sidana, 2017). In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data (Shetty, 2018).

Those are the types of classification algorithms in Machine Learning mostly used for text mining applied for financial market prediction:

1. Linear Classifiers: Naive Bayes Classifier, Logistic Regression
2. Support Vector Machines
3. Decision Trees
4. Random Forest
5. Neural Networks
6. Nearest Neighbor

Naive Bayes Classifier (Generative Learning Model):

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. In my study, Naïve Bayes classifier is going to be used to predict the stock movements as up, down or neutral (Sidana, 2017).

Logistic Regression (Predictive Learning Model):

It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables (Sidana, 2017).

Decision Trees:

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data (Sidana, 2017).

Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set (Sidana, 2017).

Neural Network:

A neural network consists of units (neurons), arranged in layers, which convert an input vector into some output. Each unit takes an input, applies a (often nonlinear) function to it and then passes the output on to the next layer. Generally, the networks are defined to be feed-forward: a unit feeds its output to all the units on the next layer, but there is no feedback to the previous layer. Weightings are applied to the signals passing from one unit to

another, and it is these weightings which are tuned in the training phase to adapt a neural network to the particular problem at hand (Sidana, 2017).

Nearest Neighbor:

The k-nearest-neighbors algorithm is a classification algorithm, and it is supervised: it takes a bunch of labelled points and uses them to learn how to label other points. To label a new point, it looks at the labelled points closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever label most of the neighbors have is the label for the new point (the "k" is the number of neighbors it checks) (Sidana, 2017).

Applied Text-Mining approach for stock price prediction based on financial news

In this thesis the main aim is to make a prediction for the future stock price movements, to achieve to build the models that gives the direction of the stock price. There are proposed 9 steps in this thesis to be conducted and they are:

1. Identifying the news sources and targeted companies
2. Data collection and data cleaning of news articles
3. Sentiment Analysis of news articles
4. Data collection of stock prices
5. Calculating Rate of Change (ROC)
6. Categorizing the data
7. Applying Naive Bayesian classifier
8. Training and Test
9. Evaluation

1. Identifying the news sources and targeted companies

In this thesis stock price prediction is based on news articles, 8 news sources are identified from which the news are collected, and they are: The Washington Post, Cnn, Market Watch, BGR, The Street, Fox Business, The Verge and Breitbart. All the 8 sources were analyzed and decided to gather the news articles from them, because those news sources have the necessary news articles needed for this thesis and are the most reliable in the market. On the other side 3 companies are targeted: Apple, Facebook and Tesla. These

3 companies are targeted because they are worldwide known and are in worlds eye as the most attractive companies.

2. Data Collection and data cleaning of news articles

For conducting this study, the main type of data needed to collect are the news articles. The process of this research starts from collecting relevant financial news articles about the companies that will be analyzed to predict the future directions of the stock prices, the financial news articles are collected for a period of 1 year, which are from the last year. The news articles are collected from the identified news sources that were discussed in the previous step.

News article's links are collected through using Web Scraper Google Chrome extension, after that to collect the features of news articles such as: title, published date, author and content there is built a python script based on Scrapy[1]. It is decided to use Scrapy because:

- It is easier to build and scale large crawling projects.
- It has a built-in mechanism called Selectors, for extracting the data from websites.
- It handles the requests asynchronously and it is fast.
- It automatically adjusts crawling speed using Auto-throttling mechanism.
- Ensures developer accessibility (Scrapy, 2019).

Scrapy also has many features such as:

- Scrapy is an open source and free to use web crawling framework.
- Scrapy generates feed exports in formats such as JSON, CSV, and XML.
- Scrapy has built-in support for selecting and extracting data from sources either by XPath or CSS expressions.

[1] Scrapy - An open source and collaborative framework for extracting the data you need from websites. www.scrapy.org

- Scrapy based on crawler, allows extracting data from the web pages automatically (Scrapy, 2019).

It also provides many advantages and some of the most important advantages are:

- Scrapy is easily extensible, fast, and powerful.
- It is a cross-platform application framework (Windows, Linux, Mac OS and BSD).
- Scrapy requests are scheduled and processed asynchronously (Scrapy, 2019).

After collecting the articles, text cleaning is a necessary before the data is ready for analysis because data obtained from web is highly unstructured.

3. Sentiment Analysis of news articles

In the sentiment analysis step every news article gets a polarity. All the financial news is classified as positive and negative. After getting the positive and negative score for each article, a column named *Sentimentof_text* its added and if the positive score is greater than negative score it gets positive classified and if negative score is greater than the positive score it gets negative classified and the neutral score it is not taken in consideration. To make the sentiment analysis in this thesis is used the NLTK package. The Natural Language Toolkit (NLTK) is a Python package for natural language processing.

4. Data collection of stock prices

After collection of the news articles, the second type of data needed for this research thesis is stock prices from the companies that are going to be analyzed, the past stock quotes are collected for a period of one year. Stock prices were collected form Yahoo! Finance.

5. Calculating the Rate of Change (ROC)

In the data set it is added one other column that is 5-day ROC that is the rate of change in average of 5 days, that is calculated to define the significance change, and one more column added as Future ROC that is the 5-day ROC after five days to see the effect the news has after 5 days. After created the two data sets, one created from relevant financial articles

and the other with the stock prices. The two data sets are merged in one data set to build the models.

6. Categorizing the data

Categorizing the data is done to apply Naive Bayesian classifier. In the data set all news are collected with their features such as: title, author, published date, content, and link. Then the sentiment analysis added new columns such as: pos (positive score), neg (negative score) and compound, two new columns are added: *Sentimentof_text* that could be "positive" if the sentiment score is greater than zero and "negative" if the sentiment score is less than zero. Neutral score is not taken in consideration of the text content because that could result in majority of neutral results. The second column is the *ROC_sentiment* that can be "positive" if the *Future ROC* is greater than zero and "negative" if the future ROC is less than zero.

7. Applying the Naïve Bayesian classifier

This thesis applies the Naïve Bayes classifier. Naive Bayes classifier has been widely used for text categorization due to its simplicity and efficiency. The Naïve Bayes applies the Bayes' theorem with the "naive" assumption that any pair of features are independent for a given class (Tang, Bo; Kay, Steven; He, Haibo;, 2016).

This thesis applies the Naïve Bayesian classifier because it provides many advantages such as:

- It is easy and fast to predict the class of the test data set. It also performs well in multi-class prediction. When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
- It performs well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption) (Chauhan, 2018).

8. Training

In this thesis the data set is split into training and test set, ten months starting from March 2018 to December 2018 it is used to train the model. Ten months Is used for the model to be well trained. January and February 2019, it is for the test set. 2 months testing to see how the model performs.

9. Evaluation

Evaluation is the last step of the process, and it assess the quality of the created predictive model.

Chapter 4. Implementation

Steps that are conducted were defined in the third chapter, while in this chapter there is a detail description regarding how the steps are conducted during the implementation phase.

The data collected is for a period of one year, starting from 1st of March 2018 until 1st of March 2019. To make the prediction there are used different variable like the polarity of the news, positive or negative, rate of change in stocks prices average of 5 days, source of news and the company name.

The following are the steps needed to undertake to perform stock price prediction of financial news, the steps are depicted in the figure below (see 6 Figure).

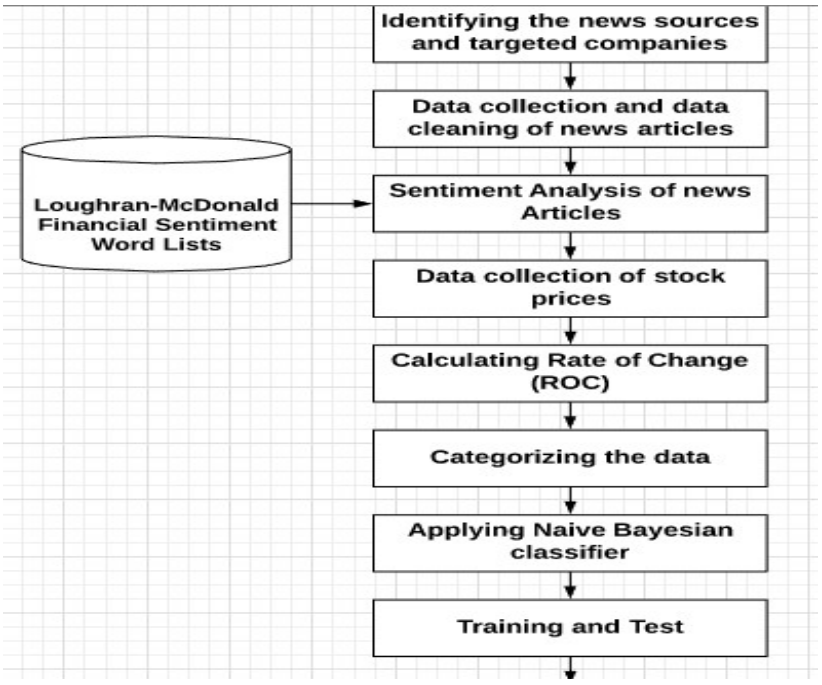


Figure 6. 9 steps to be conducted for implementation

1. Identifying the news sources and targeted companies

Is crucial to understand the data that is analyzed in the study. The information collected must be relevant and trustworthy. As such, the relevant data from financial news articles from top reliable sources have been identified as: The Washington Post, Cnn, MarketWatch, BGR, Fox Business, The Street, The Verge and Breitbart, and the targeted companies for this study are: Tesla, Facebook and Apple. News sources are proven to be reliable in the market as the most unbiased.

2. Data collection and data cleaning of news articles

Links of the news from the sources mentioned in step 1 are collected using Web Scraper extension of Google Chrome browser. After collecting all the links, a python script was built based on Scrapy framework that is extracting data from the links and organizing them in the following structure: article's title, date, author and the text content. The scrapy python script used modules, the code reads the links from the csv file created from web scraper Google chrome extension (see 7 Figure). To gather title, published date, author and content was done specifying xPath selectors for each element (see 8 Figure). All the news article features were appended together (see 10 Figure). It was succeeded to collect the news articles from 8 different news sources, totaling 20226 news articles, split into Table 2 for Training Set and Table 3 for Test Set. Appropriate data cleaning has been applied to remove unnecessary text as well as to format the data from different sources to one standard (see 9 Figure). News sources was blocking them from being scraped, in order to prevent the blocking in this thesis *scrapy-user-agents* was used, the code was written in the project settings.py (see 11 Figure).

Figure 8. Scrapy script, imported modules and connections

```
def parseNews(self, response):
    title = response.selector.xpath("//h1[@data-pb-field='custom.topperDisplayName']/text()").extract_first()
    if not title:
        title = response.selector.xpath("//h1[@class='entry-title']/span/text()").extract_first()
    author = response.selector.xpath("//span[@class='author-name']/text()").extract_first()

    if not author:
        author = response.selector.xpath("//a[@class='author-name']/text()").extract_first()

    if not author:
        author = response.selector.xpath("//div[@class='author-byline']/text()").extract_first()


    published_at = response.selector.xpath("//span[@class='author-timestamp']/text()").extract_first()
    if not published_at:
        published_at = response.selector.xpath("//p[@id='published-timestamp']/span/text()").extract_first()
    content = ""
    for p in response.selector.xpath("//div[@id='article-body']//p/text()"):
        content += str(p.extract().encode("utf-8"))

    link = response.url

    if len(content) < 3:
        for p in response.selector.xpath("//p/text()"):
            content += str(p.extract().encode("utf-8"))

    try:
        date = str(parse(published_at, fuzzy=True))
    except:
        date = str(datetime.now())
```

Figure 7. Scrapy script, parse news

```
import scrapy
from datetime import datetime
from dateutil.parser import parse
import json, csv
import time
class washingtonpost(scrapy.Spider):
    name = "washingtonpost"
    start_urls = ['https://www.washingtonpost.com/19b85bb0-8c1a-11e8-9d59-dccc2c0cabcf_story.html']
    base_urls = ['https://www.washingtonpost.com/']
    def __init__(self):
        self.data = []
        self.files = ["the-washingtonpost-facebook-url.csv"]
        self.headers = {}
    def start_requests(self):
        url = self.start_urls[0]
        request = scrapy.Request(url, self.parse, headers=self.headers)
        yield request
    def close(self, spider):
        with open('washingtonpost-facebook-news.json', 'w') as outfile:
            json.dump(self.data, outfile)
    def parse(self, response):
        for file in self.files:
            with open(file) as csvfile:
                reader = csv.DictReader(csvfile)
                for row in reader:
                    request = scrapy.Request(row["URL-href"], callback=self.parseNews, headers=self.headers,
                                             dont_filter=True)
                    time.sleep(1)
                    yield request
```

```
    def cleanString(incomingString):
        newstring = incomingString
        newstring = newstring.replace("b\'", "")
        newstring = newstring.replace("xe2", "")
        newstring = newstring.replace("x80", "")
        newstring = newstring.replace("x99s", "")
        newstring = newstring.replace("r", "")
        newstring = newstring.replace("'\n", "")
        newstring = newstring.replace("'b'", "")
        newstring = newstring.replace("xc2", "")
        newstring = newstring.replace("xa0 ", "")
        newstring = newstring.replace("xc2", "")
        newstring = newstring.replace("x94 ", "")
        newstring = newstring.replace("xa0", "")
        newstring = newstring.replace("('b') ", "")
        newstring = newstring.replace("x9d", "")
        newstring = newstring.replace("'b'", "")
        newstring = newstring.replace("b\"", "")
        newstring = newstring.replace("x9c'", "")
        newstring = newstring.replace("\'", "")
        newstring = newstring.replace("\"", "")
        newstring = newstring.replace("\\", "")
        newstring = newstring.replace("/", "")
        newstring = newstring.replace("x9c", "")
        newstring = newstring.replace("x99", "")
        newstring = newstring.replace("  |  | ", "")
        newstring = newstring.replace("  n  ", "")
        newstring = newstring.replace("x98", "")
        newstring = newstring.replace("  ", " ")
        newstring = newstring.replace("nn", "")
        return newstring
```

Figure 9. Scrapy script, content cleaning

```
self.data.append({
    "title": title,
    "author": author,
    "published_at": date,
    "content": cleanString(content),
    "link": link,
    "source": "The Washington post",
    "Company": "Facebook"
})
```

Figure 10. Data append scrapy script

```
DOWNLOADER_MIDDLEWARES = {
    'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware': None,
    'scrapy_user_agents.middlewares.RandomUserAgentMiddleware': 400,
}
```

Figure 11. Scrapy-user-agents

3. Sentiment Analysis of news articles

Sentiment analysis was applied to every news record based on the news content by using Vader Sentiment Analysis. VADER (Valence Aware Dictionary for sEntiment Reasoning) is a pre-built sentiment analysis model included in the NLTK package. It can give both positive/negative (polarity) as well as the strength of the emotion (intensity) of a text. VADER however is focused on social media and short texts unlike Financial News which are almost the opposite. It was updated the VADER lexicon with words plus sentiments from other sources/lexicons such as the Loughran-McDonald Financial Sentiment Word Lists, to be appropriate for our collected financial news (Yip, 2018). At the end of this step we had the polarity of the news content recorded in our dataset. The python script that conducted the sentiment analysis with nltk package is shown in the below figure (see 12 Figure).

```
import nltk
import csv
from nltk.sentiment.vader import SentimentIntensityAnalyzer
nltk.downloader.download('vader_lexicon')
sia = SentimentIntensityAnalyzer()

result = []
with open('thewashingtonpost-tesla-news.csv') as readfile:
    reader = csv.DictReader(readfile)
    for row in reader:
        sentimentResult = sia.polarity_scores(row["Column1.content"])
        row["neg"] = sentimentResult["neg"]
        row["pos"] = sentimentResult["pos"]
        row["neu"] = sentimentResult["neu"]
        row["compound"] = sentimentResult["compound"]
        result.append(row)

with open('thewashingtonpost-tesla-withsentiment.csv', 'w', newline='') as writeFile:
    writer = csv.DictWriter(writeFile, result[0].keys())
    writer.writeheader()
    for row in result:
        writer.writerow(row)

readfile.close()
writeFile.close()
```

Figure 12.Sentiment analysis script

4. Data collection of stock prices

For each of targeted companies stock prices were collected from Yahoo Finance portal for a period of 1 year starting from 1st March 2018 until 1st March 2019 where the following information were collected: date, open price, high, low, close price, volume and Adj close. These data are important for correlation with the appropriate news from first data set.

5. Calculating Rate of Change (ROC)

The 5-day ROC and Future ROC are the two variables that are calculated from the data set from Step 4. The rate of change (ROC) in stocks in average of 5 days is an existing formula that refers to the last 5 days of stock fluctuation. In our case we also added a column with the Future ROC (the ROC after 5 days), having in mind that the effect of this positive or negative news will be reflected in the future and not the past. Since we are dealing with historical data, the Future ROC is easy to calculate.

6. Categorizing the data

Categorizing the data must be done in order to apply Naive Bayesian classifier. In the data set all news are collected with their features, it is added two new columns:

Sentimentof_text that could be "positive" if the sentiment score is greater than zero and "negative" if the sentiment score is less than zero. Neutral score of the text content is not taken in consideration because that could result in majority of neutral results. The second column is the *ROC_Sentiment* that can be "positive" if the *Future ROC* is greater than zero and "negative" if the future ROC is less than zero.

7. Applying Naive Bayesian classifier

Naive Bayesian classifier is applied to make the prediction of the future stock movements, the Naive Bayes applies the Bayes' theorem with the "naive" assumption that any pair of features are independent for a given class (Tang, Bo; Kay, Steven; He, Haibo;, 2016). To prepare the data set to make prediction with the Naïve Bayesian, a new column was added with the name *class* that is "UP" if the *Sentimentof_text* is "positive" and the *ROC_Sentiment* is "positive", if the *Sentimentof_text* is "negative" and the *ROC_Sentiment* is "negative" then the class is "DOWN", otherwise is "NEUTRAL" classification. The implementation of Naïve Bayes classifier was done with the data analysis add-in solution for Microsoft Excel, the XLSTAT.

XLSTAT is a suite of statistical add-ins for Microsoft Excel that has been developed since 1993 by Addinsoft to enhance the analytical capabilities of Microsoft Excel. Since 2003, Addinsoft is a Microsoft partner and all the XLSTAT analytical add-ins are registered on the Office Marketplace. The XLSTAT software relies on Microsoft Excel for the input of data and the display of results. This makes the software very convenient to share data and results. Computations are done using autonomous software components that are optimized for speed and efficiency, Naive Bayes classifier is built in the XLSTAT (XLSTAT, 2019).

The training dataset results are summarized in Table 4, where for each company in our target list the classification results are shown. As general finding is that the algorithm applied as explained in this step, classifies 15.71% of the articles in the training set as "DOWN" (meaning the stock will go down in the following days), 50.71% is classified as "NEUTRAL" (there is no clear picture on what the prediction will be) and 33.59% of the data as "UP" (meaning the stock will go up). The "UP" classification is relevant to our study and can be used for simulating investments on our test data from the test set.

Different research approaches proved that there is a strong correlation between financial news and stock price changes like (Khedr, Ayman Elsayed; Salama, S E; Yaseen, Nagwa, 2018), in their paper, they proposed and approach that uses sentiment analysis for financial news, along with features extracted from historical stock prices to predict the future behavior of stock market. Three categories of news data were considered: news relevant to market, company news and financial reports that were published by financial experts about stocks. The proposed model consists of two stages, the first stage is to determine the news polarities to be either positive or negative using naïve Bayes algorithm. The results of their proposed model achieved higher accuracy for sentiment analysis in determining the news polarities by using Naïve Bayes algorithm up to 86.21%. The results of the proposed model are compatible with the researches that state that there is a strong relation between stock news and changes in stock prices (Khedr, Ayman Elsayed; Salama, S E; Yaseen, Nagwa, 2018).

Many research has been carried in the area of prediction of stocks, (Joshi, Kalyani; Rao, Jyothi; N., Bharathi N., 2016) project is about taking non quantifiable data such as financial news articles about a company and predicting about a company and predicting its future stock trend with news sentiment classification. This research is an attempt to build a model that predicts news polarity which may affect changes in stock trends, authors have taken past three years data from Apple Company as stock price and news articles. They assumed that news articles and stock prices are related to each other. And, news may have capacity to fluctuate stock trend. In this work they also applied the Naïve Bayes algorithm performance is around 83%.

The volatility of stock prices depends on gains or losses of certain companies. News articles are one of the most important factors which influence the stock market. The study conducted by (Dnyaneshwar , Kirange K ; Ratnadeep , Deshmukh;, 2016), basically shows the effect of emotion classification of financial news to the prediction of stock market prices. In order to find correlation between sentiment predicted from news and original stock price, they plot the sentiments of two companies (Infosys and Wipro) over a period of 10 years. Naïve Bayes classifier was applied in this study and got an accuracy of 72.64%.

8. Training and Test

The data that is collected (see Table 2 and Table 3) contains records for 12 months from which 10 months will be used to train the model and the last 2 months will be used for the test set, to evaluate how it performs. In total 18236 records will be used as training dataset and the remaining 1990 records (roughly 10%) out of 20226 will be used as test set.

In this thesis are created 2 models to make stock price prediction. The following variables are used to train and test the first model: *Source, Company, Sentimentof_text* and the *5-day ROC* while in the second model is trained only with these variables: *Source, Company and Sentimentof_text*.

After conducting all the steps for collection of data, giving polarity to the news for each article and merging with stock data the data set look like the table below (see Table 1).

Table 1. Sample training dataset after the processing. Data from single day, limited to 30 records.

Author	Date	Content	link	Source	Company	neg	pos	neu	compound	5-day ROC	Future RC	Sentiment WithoutNeutraClass	ROC_Sentiment	CLASS
Tauhid C	4/5/2018			The Washington post	Apple	0.031	0.08	0.889	0.9954	3.79626	0.77546	POSITIVE	POSITIVE	UP
Steve He	4/5/2018			The Washington post	Apple	0.064	0.036	0.9	-0.9818	3.79626	0.77546	NEGATIVI	POSITIVE	NEUTRAL
Juli Brisk	4/5/2018			The Washington post	Facebook	0.029	0.041	0.93	0.708	4.12337	2.84298	POSITIVE	POSITIVE	UP
Geoffrey	4/5/2018			The Washington post	Facebook	0.03	0.076	0.894	0.9966	4.12337	2.84298	POSITIVE	POSITIVE	UP
Amy Dick	4/5/2018			The Washington post	Facebook	0.028	0.021	0.952	-0.4404	4.12337	2.84298	NEGATIVI	POSITIVE	NEUTRAL
Eugene I	4/5/2018			The Washington post	Facebook	0.048	0.063	0.89	0.7114	4.12337	2.84298	POSITIVE	POSITIVE	UP
Joshua F	4/5/2018			The Washington post	Facebook	0.057	0.054	0.889	-0.7778	4.12337	2.84298	NEGATIVI	POSITIVE	NEUTRAL
David Ig	4/5/2018			The Washington post	Facebook	0.062	0.05	0.888	-0.8442	4.12337	2.84298	NEGATIVI	POSITIVE	NEUTRAL
Lenny Be	4/5/2018			The Washington post	Facebook	0.055	0.03	0.915	-0.9832	4.12337	2.84298	NEGATIVI	POSITIVE	NEUTRAL
	4/5/2018			market-watch	Apple	0.122	0.013	0.865	-0.8834	3.79626	0.77546	NEGATIVI	POSITIVE	NEUTRAL
	4/5/2018			market-watch	Apple	0.051	0.14	0.809	0.998	3.79626	0.77546	POSITIVE	POSITIVE	UP
Brett Are	4/5/2018			market-watch	Facebook	0.044	0.033	0.924	-0.7651	4.12337	2.84298	NEGATIVI	POSITIVE	NEUTRAL
	4/5/2018			market-watch	Facebook	0.015	0	0.985	-0.2382	4.12337	2.84298	NEGATIVI	POSITIVE	NEUTRAL
	4/5/2018			market-watch	Tesla	0.056	0.025	0.919	-0.6249	18.5973	-3.80741	NEGATIVI	NEGATIVI	DOWN
	4/5/2018			market-watch	Tesla	0	0.081	0.919	0.296	18.5973	-3.80741	POSITIVE	NEGATIVI	NEUTRAL
Paul R. La	4/5/2018			CNN	Apple	0.042	0.113	0.845	0.9944	3.79626	0.77546	POSITIVE	POSITIVE	UP
Richard C	4/5/2018			CNN	Facebook	0.049	0.089	0.862	0.9816	4.12337	2.84298	POSITIVE	POSITIVE	UP
Ivana Ko	4/5/2018			CNN	Facebook	0.033	0.073	0.894	0.9231	4.12337	2.84298	POSITIVE	POSITIVE	UP
Zach Eps	4/5/2018			BGR	Apple	0.021	0.069	0.911	0.985	3.79626	0.77546	POSITIVE	POSITIVE	UP
Chris Sm	4/5/2018			BGR	Facebook	0.085	0.028	0.888	-0.9849	4.12337	2.84298	NEGATIVI	POSITIVE	NEUTRAL
Motley Fo	4/5/2018			Fox Business	Apple	0.009	0.046	0.945	0.9606	3.79626	0.77546	POSITIVE	POSITIVE	UP
Megan H	4/5/2018			Fox Business	Facebook	0.054	0.021	0.925	-0.791	4.12337	2.84298	NEGATIVI	POSITIVE	NEUTRAL
Motley Fo	4/5/2018			Fox Business	Tesla	0.034	0.103	0.863	0.9844	18.5973	-3.80741	POSITIVE	NEGATIVI	NEUTRAL
Kinsey Gi	4/5/2018			The Street	Apple	0.03	0.076	0.894	0.9568	3.79626	0.77546	POSITIVE	POSITIVE	UP
Eric Jhon	4/5/2018			The Street	Facebook	0.026	0.037	0.937	0.8157	4.12337	2.84298	POSITIVE	POSITIVE	UP
Tedd Coh	4/5/2018			The Street	Facebook	0.03	0.062	0.907	0.8504	4.12337	2.84298	POSITIVE	POSITIVE	UP
Eric Jhon	4/5/2018			The Street	Tesla	0.017	0.077	0.906	0.9957	18.5973	-3.80741	POSITIVE	NEGATIVI	NEUTRAL
Chaim Ga	4/5/2018			The Verge	Apple	0.074	0.071	0.855	-0.2996	3.79626	0.77546	NEGATIVI	POSITIVE	NEUTRAL
Adi Robe	4/5/2018			The Verge	Facebook	0.017	0.052	0.93	0.9806	4.12337	2.84298	POSITIVE	POSITIVE	UP
Loren Gru	4/5/2018			The Verge	Tesla	0.015	0.034	0.951	0.9771	18.5973	-3.80741	POSITIVE	NEGATIVI	NEUTRAL

Table 2. Total news articles obtained for Apple, Tesla and Facebook organized by Source for Training Set. Period March 2018-December 2018

Variable	Categories	Frequencies	%
Source	BGR	1073	5.884

	Breitbart	435	2.385
	CNN	687	3.767
	Fox Business	813	4.458
	The Street	3810	20.893
	The Verge	2847	15.612
	The Washington post	6051	33.182
	market-watch	2520	13.819
Company	**Apple**	7591	41.626
	Facebook	7513	41.199
	Tesla	3132	17.175

Table 3. Total news articles obtained for Apple, Tesla and Facebook organized by Source for Test Set. Period January 2019-March 2019

Variable	**Categories**	**Frequencies**	**%**
Source	BGR	185	9.30
	Breitbart	167	8.39
	CNN	211	10.60
	Fox Business	147	7.39
	The Street	590	29.65
	The Verge	603	30.30
	The Washington post	87	4.37
	market-watch	0	0
Company	**Apple**	1144	57.49
	Facebook	416	20.90
	Tesla	430	21.61

Table 4. Training set classification data organized by Company and frequency

CLASS

Company	DOWN	NEUTRAL	UP	Total
Apple	1,006	3,930	2,655	7,591
%	*5.52*	*21.55*	*14.56*	*41.63*
Facebook	1,390	3,683	2,440	7,513
%	*7.62*	*20.20*	*13.38*	*41.20*
Tesla	468	1,634	1,030	3,132
%	*2.57*	*8.96*	*5.65*	*17.17*
Total	2,864	9,247	6,125	18,236
%	*15.71*	*50.71*	*33.59*	*100.00*

9. Evaluation

On the two models created, the XLSTAT have calculated the global accuracy of the model. XLSTAT display the model accuracy, which is the proportion of correct predictions. For the first model created there is an accuracy of 94.29% while for the second model there is an accuracy of 49.49%.

Chapter 5. Results and discussion

With the completion of the implementation phase (collecting the necessary data and applied sentiment analysis), it is applied Naïve Bayes classifier to train the model and to predict the classes up, down or neutral for the test set. In this thesis two models are created. For the two models the data set is split into training and test set and the models are trained and then tested.

First stock price prediction model

The first model was first given the training set records, in the training set there are 18236 records that are used to train the model, in the training set are used these columns: *Source*, *Company*, *Sentimentof_text*, 5-day ROC and the class. To test the first model 4 variables are used as input are given: Source, Company, *Sentimentof_text* and 5-day ROC to

predict the class up, down or neutral. The results achieved in the first model are shown in the figure below (see 13 Figure).

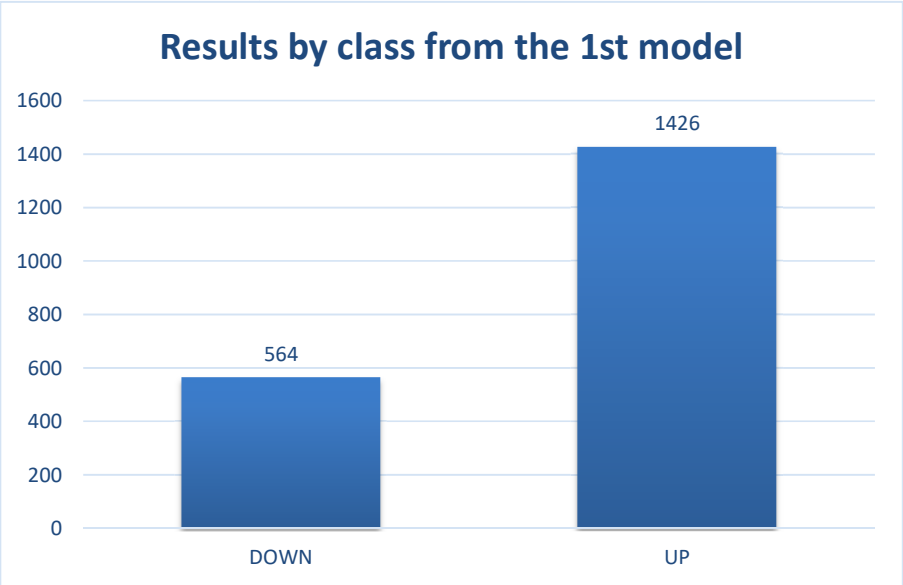


Figure 13.Results by class from 1st model

In the first model 564 down and 1426 up stock price direction were predicted from 1900 records being tested. The global accuracy of the first model is 94.29%, which is calculated from XLSTAT, which is the proportion of correct predictions.

Second stock price prediction model

The second model 18236 records are used to train the model, but there are only these columns: *Source*, *Company*, *Sentimentof_text* and the class. To train the second model 3 variables as input are given: Source, Company and *Sentimentof_text* to predict the class as up, down or neutral. The results achieved in the second model are shown in the figure below (see 14 Figure). The difference from the first model is that in the second model 5-day ROC variable is not used taken in consideration.

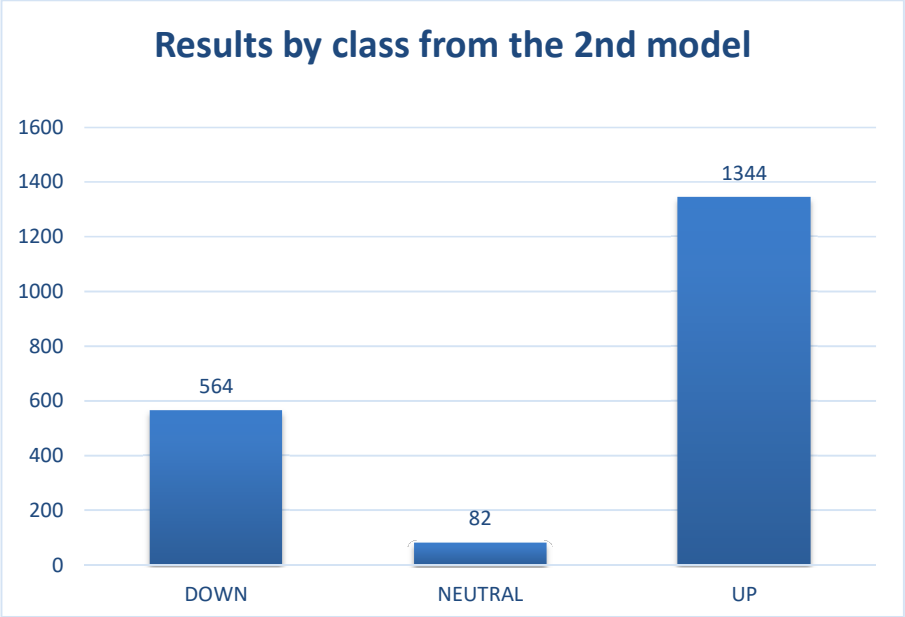


Figure 14. Results by class the 2nd model

In comparison with the first model that has an accuracy of 94.29% the second model has 49.49% which is significantly lower accuracy than the first model that has just one more variable the 5-day ROC. It can be stated that aside from sentiment of text, stock data as in this case 5-day ROC plays a vital role in prediction of the future stock price movements.

Research Findings

In this thesis two hypotheses have been raised, that are the main goal of this thesis findings. The hypotheses are discussed in detail, and the author's findings are shown.

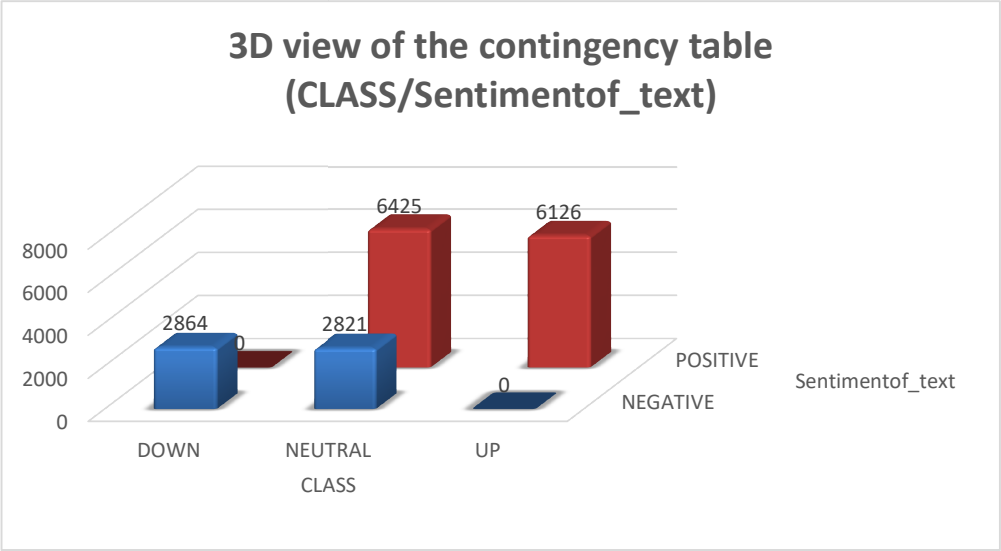
The first hypothesis raised in this thesis claimed that: *The financial news in the media tend to significantly change stock prices for the company on the stock market.* In order to find whether the financial news in the media tend to significantly change stock prices for the company on the stock market, it has to be analyzed the impact on the future ROC after a news has been posted. After a news classified as positive has been posted how has impacted the future ROC and after a news classified as positive has been posted how has impacted the future ROC. To see the relationship between the sentiment of the text and the sentiment of the future ROC, is created a contingency table between two variables: *ROC_sentiment* and *Sentimentof_text* (see 5 Table).

Table 5.Contingency table (ROC_Sentiment/Sentimentof_text)

	SENTIMENT OF TEXT	
ROC_Sentiment	**NEGATIVE**	**POSITIVE**
NEGATIVE	2864	6425
POSITIVE	2821	6126
Total	**5685**	**12551**

In the table above created a contingency table to see the relationship between the future ROC and sentiment of text, it is shown that in 5685 negative news there are also 2864 negative effected future ROC, or in 50,3% of the cases negative news affect negatively the future rate of change for a given stock. In the case of 12551 positive news articles, it is shown that 6126 cases, or 48.8% of the cases, the positive news affect positively the future rate of change of the given stock. Based on the literature review and the research that have been conducted in this thesis it is proved that financial news in the media has evident impact on the stock price movements.

The first hypothesis is supported by additional hypothesis: *Positive financial news on the media tend to significantly increase stock prices.* This hypothesis aim is to find whether a positive classified news increase the stock prices, above it was created a contingency table between two variables: *ROC_sentiment* and *Sentimentof_text* (see 5 Table) and gave a description of the table, according to the data set created and the results obtained in this thesis there is a weak correlation that the positive news increase the stock prices while when a negative news has more impact in decreasing the stock prices. To see the relationship between positive news and the direction of the stock prices it was created a 3D view of the contingency table (see 15 Figure).



In the above graph it is clearly that in all cases that there is a positive news the stock price direction goes up or neutral, 6126 positive classified news have led to up stock price direction and on the other side when there is a negative news the stock price direction goes down or neutral, 2864 negative classified news have led to down stock price direction.

All in all, from analyzing the contingency table that is created between two variables: *ROC_sentiment* and *Sentimentof_text (see 5 Table)* and the 3D view of the contingency table created between the class and Sentimentof_text (see 15 Figure), it can be stated that the financial news tends to change significantly the stock price for the company on the stock

Figure 15. 3D view of the contingency table (CLASS/ Sentimentof_text)

market, and positive financial news on the media tend to significantly increase stock prices. Positive news articles make a moderate positive impact on stock prices for the company while the negative news has more impact on the stock prices, negative impact by making stock prices for the company move to down direction, decrease the price.

The second hypothesis raised in this thesis is*: Even though Efficient Market Hypothesis (EMH) clearly states that financial stock prices cannot be predicted we argue that based on several attributes from new articles we can reach certain level of prediction and*

give directions to financial experts, in the first model trained with four variables: source, company, Sentimentof_text and 5-day ROC got an accuracy of 94.29%, the accuracy achieved shows that there is a very high chance to predict the stock price movements. By this is valid the hypothesis raised that with several attributes there can be reached a certain level of prediction but EMH still remains because there is no 100% prediction. The accuracy rate of the first model is a high accuracy and it can be seen that there is a strong relationship between financial news and stock price movements.

While in the second model without 5-day ROC, without the stock data the model trained with only three variables: source, company and Sentimentof_text got an accuracy of 49.49% that is less than the guessing probability (50%), in this case EMH cannot be rejected. The probability to make a correct prediction about the future stock price movements is low.

In this thesis it can be stated that the weak-form of EMH is true. Weak form efficiency states that all future price movements follow a random walk, unless there is some change in some fundamental information. It does not state that prices adjust immediately in the advent of new fundamental information, which means that some forms of fundamental analysis and news article analysis might provide excess returns. This is because they trade on new information and does not use any historical information to look for patterns.

Simulation

With 10.000$ investments per company per day

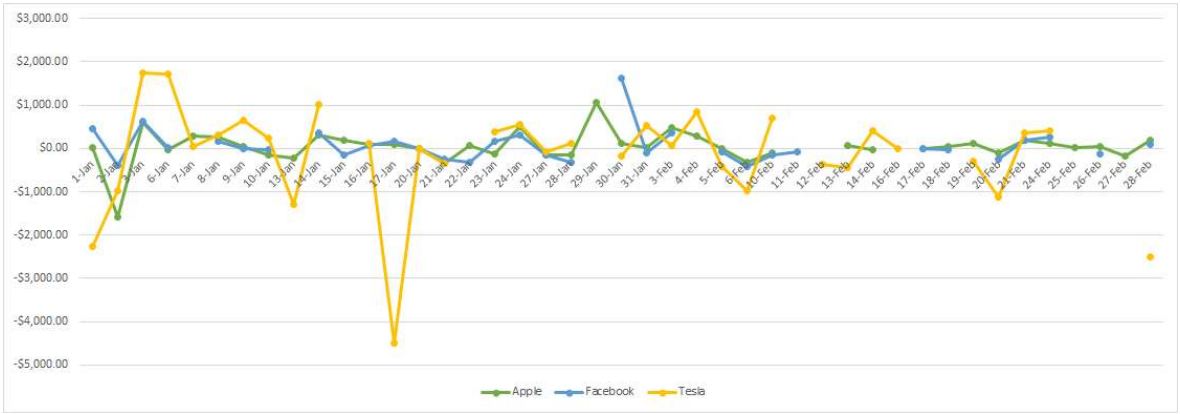


Figure 16. Profit/Loss simulation on Test set data based on classification model (Naive Bayes) used in this study

This chart clearly shows the fluctuation on the investment simulation based on real data from stock closing prices, for the three companies in combination with the model as explained in previous table. TESLA data shows more fluctuation and as such it was excluded in the next chart to see if the model prediction can be used for investment.

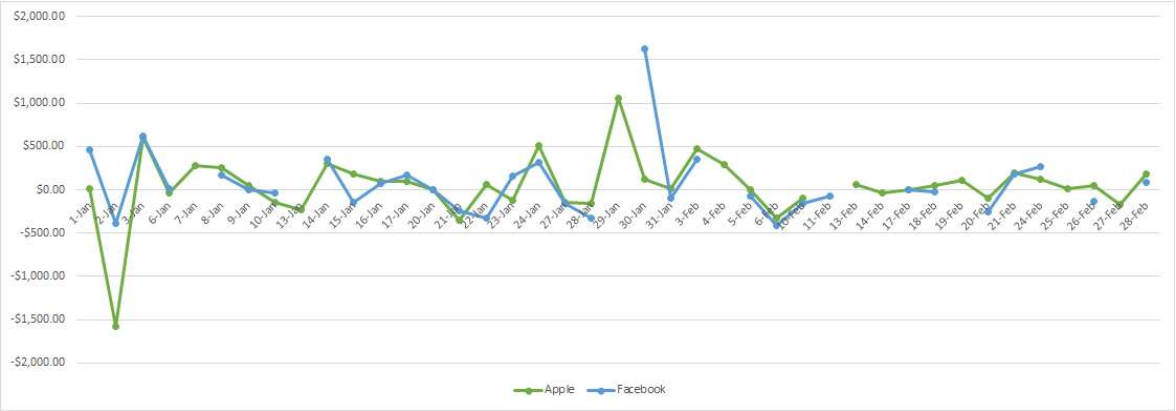


Figure 17. Profit/Loss simulation on Test set data based on classification model (Naive Bayes) used in this study, excluding TESLA company

Table 6. Profit and Loss table for Apple and Facebook based on the Test Set simulation.

Date (2019)	Apple	Facebook	Grand Total
1-Jan	$18.00	$459.00	$477.00
2-Jan	$(1,573.00)	$(394.00)	$(1,967.00)
3-Jan	$607.00	$621.00	$1,228.00
6-Jan	$(33.00)	$10.00	$(23.00)
7-Jan	$282.00		$282.00
8-Jan	$256.00	$170.00	$426.00
9-Jan	$49.00	$(3.00)	$46.00
10-Jan	$(151.00)	$(40.00)	$(191.00)

13-Jan	$(229.00)		$(229.00)
14-Jan	$307.00	$356.00	$663.00
15-Jan	$187.00	$(141.00)	$46.00
16-Jan	$92.00	$76.00	$168.00
17-Jan	$96.00	$174.00	$270.00
20-Jan	$-	$-	$-
21-Jan	$(352.00)	$(247.00)	$(599.00)
22-Jan	$62.00	$(327.00)	$(265.00)
23-Jan	$(122.00)	$153.00	$31.00
24-Jan	$506.00	$318.00	$824.00
27-Jan	$(146.00)	$(154.00)	$(300.00)
28-Jan	$(162.00)	$(328.00)	$(490.00)
29-Jan	$1,057.00		$1,057.00
30-Jan	$119.00	$1,627.00	$1,746.00
31-Jan	$8.00	$(98.00)	$(90.00)
3-Feb	$473.00	$354.00	$827.00
4-Feb	$293.00		$293.00
5-Feb	$6.00	$(67.00)	$(61.00)
6-Feb	$(330.00)	$(411.00)	$(741.00)

10-Feb	$(98.00)	$(154.00)	$(252.00)
11-Feb		$(75.00)	$(75.00)
13-Feb	$62.00		$62.00
14-Feb	$(38.00)		$(38.00)
17-Feb	$-	$-	$-
18-Feb	$51.00	$(21.00)	$30.00
19-Feb	$110.00		$110.00
20-Feb	$(97.00)	$(252.00)	$(349.00)
21-Feb	$191.00	$185.00	$376.00
24-Feb	$126.00	$273.00	$399.00
25-Feb	$10.00		$10.00
26-Feb	$54.00	$(132.00)	$(78.00)
27-Feb	$(172.00)		$(172.00)
28-Feb	$182.00	$83.00	$265.00
Grand Total	$1,701.00	$2,015.00	$3,716.00

In the table above (see 6 Table) Profit and Loss table for Apple and Facebook based on the Test Set simulation with 10.000 investments was conducted. The simulation conducted does not show 100%-win case for the classification of stock prediction and as such it does not apply to all companies. The difference where there are better results relies on the targeted companies, such as Apple and Facebook, which are more stable ones rather

than Tesla, which as a case had different fluctuations that in long term did not bring good results in our simulation.

Chapter 6. Conclusion

The role and impact of the Internet by the end of the second decade is by taking unimaginable dimensions and changing people's lives.

The trading of stock in public companies is an important part of the economy, so in this thesis stocks are analyzed through using data mining and text mining techniques to make a prediction for stock price directions of the stocks for 3 companies listed public, which are Apple, Facebook and Tesla.

The contribution in this thesis are the 9 steps undertaken to complete this research thesis and the conducted steps are: Identifying the news sources and targeted companies, Data collection and data cleaning of news articles, Sentiment Analysis of news articles, Data collection of stock prices, Calculating Rate of Change (ROC), Categorizing the data, Applying Naive Bayesian classifier, Training the model then testing and Evaluation of the model.

The financial news which is the most crucial data in this thesis were collected from relevant sources that are: The Washington Post, Cnn, Market Watch, BGR, Fox Business, The Street, The Verge and Breitbart. The collection of the links of all news articles has been done through Web Scraper Google chrome extension, and for collection of news features such as: title, author, published date and content a scrapy python script has been created. All the past stock quotes are collected from the portal Yahoo! Finance.

In this thesis created two models to make prediction about the stock price movements, the first model created with 4 variables and trained with 18236 records and tested with 1900 records achieved an accuracy about 94.29%. The second model created with 3 variables and trained with 18236 records and tested with 1900 records achieved an accuracy about 49.49%. In the second model the 5-day ROC is not used and there is very low accuracy compared the first model, it can be concluded that together with the sentiment of

text and the past stock prices has an important role on prediction of the stock price movements.

Previous models for sentiment analysis of financial news articles are limited in news articles from relevant sources and as such, based only on sentiment of the news do not provide enough information for future movements. The models in this this adds more variables to the dataset in order to give more accuracy to the prediction.

In this thesis two hypotheses were raised, the first hypothesis raised in this thesis claimed that: *The financial news in the media tend to significantly change stock prices for the company on the stock market,* and the findings in this thesis showed that financial news in the media has a moderate impact on the fluctuation of the stock prices. The first hypothesis is supported by additional hypothesis: *Positive financial news on the media tend to significantly increase stock prices.* The analysis done showed that positive news has a moderate impact on increasing the stock but in the analysis, it was found that the negative news has a higher impact on the stock price movements compared to the positive news, negative news impacts negatively the future ROC and all those findings are based on the data set created in this thesis.

The second hypothesis raised in this thesis claimed that*: Even though Efficient Market Hypothesis (EMH) clearly states that financial stock prices cannot be predicted we argue that based on several attributes from news articles we can reach certain level of prediction and give directions to financial experts*, in the first model trained with four variables: source, *company*, *Sentimentof_text* and 5-day ROC got an accuracy of 94.29%, the accuracy achieved shows that there is a very high chance to predict the stock price movements. By this is valid the hypothesis raised that with several attributes can be reached a certain level of prediction but EMH still remains because there is no 100% prediction.

Unfortunately, there is no 100% prediction for the future of stock prices, and the main reason is that there are too many variables included that can change and that are unpredictable.

As the results are probabilistic weights (predictions) and thus classification, the simulation conducted does not show 100%-win case for the classification of stock prediction

and as such it does not apply to all companies. The difference where there are better results relies on the targeted companies, such as Apple and Facebook, which are more stable ones rather than Tesla, which as a case had different fluctuations that in long term did not bring good results in our simulation.

Bibliography

Aase, K. G. (2011). *Text Mining of News Articles for stock Price Predictions.* Trondheim.

Abdullah, Sheikh Shaugat; Rahaman, Mohammad Saiedur; Rahman, Mohammad Saidur;. (2013). *Analysis of Stock Market using Text Mining and Natural Language Processing.* Dhaka: IEEE.

Akita, R., Yoshihara, A., Takashi, M., & Kuniaki, U. (2016). *Deep Learining for Stock Prediction Using Numerical and Textual Infomration.* Okayama: ICIS.

Anurag, Nagar; Hahsler, Michael;. (2012). *Using Text and Data Mining Techniques to extract Stock Market Sentiment from Live News Streams.* Singapore: International Conference on Computer Technology and Scienc.

Basu, C. (2018, May 29). *ZACKS*. Retrieved from What Is a Stock and How Do Stocks Affect the Economy?: https://finance.zacks.com/stock-stocks-affect-economy-2233.html

Beckmann, M. (2017). *Srock Price Change Prediction Using News Text Mining.* Rio de Janeiro.

Bollen, Johan; Mao, Huina; Zeng, Xiao Jun;. (2010). *Twitter mood predicts the stock market.* arXiv.

Bose, B. (2018, 7 10). *Digital Vidya*. Retrieved from Techniques and Applications of Text Mining: https://www.digitalvidya.com/blog/techniques-applications-text-mining/

Cakra, Yahya Eru; Trisedya, Bayu Distiawan;. (2015). *Stock Price Prediction using Linear Regression based on Sentiment Analysis.* Depok.

Catanzarite, J. (2018, 12 14). *Towards Data Science*. Retrieved from The Naïve Bayes Classifier: https://towardsdatascience.com/the-naive-bayes-classifier-e92ea9f47523

Chauhan, G. (2018, 10 8). *Towards Data Science*. Retrieved from All about the Naive Bayes: https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf

Dang, Minh; Doung, Duc;. (2016). *Improvements Methods for Stock Market Prediction using Financial News Articles.* Danang: IEEE.

Day, Min Yuh; Lee, Chia Chou;. (2016). *Deep Learning for Financial Sentiment Analysis on Finance News Providers.* San Francisco: IEEE/ACM.

Ding, Xiao; Zhang, Yue; Duan, Junwen;. (2015). *Deep Learning for Event-Driven Stock Prediction.* IJCAI.

Ding, Xiao; Zhang, Yue; Liu, Ting; Duan, Junwen;. (2014). *Using Strucutred Events to Predict Stock Price Movement: An Empirical Investigation.* Doha: Association for Computational Linguistics.

Dnyaneshwar , Kirange K ; Ratnadeep , Deshmukh;. (2016). *Sentiment Analysis of News Headlines for Stock Price Prediction.* COMPUSOFT.

Dudovskiy, J. (2018, 1 1). The Ultimate Guide to Writing a Dissertation in Business Studies: A Step-by-Step Assistance. Jonathan, Pittsburgh, USA. Retrieved from Data Collection Methods.

Falinouss, P. (2007). *Stock Trend Prediction Using News Articles.* Luleå.

Fayyad, U., & Shapiro, G.-S. P. (1996). *From Data Mining to.*

Góralewicz, B. (2018, March 6). *Elephate*. Retrieved from The TF-IDF Algorithm Explained: https://www.elephate.com/blog/what-is-tf-idf/

Groth, Sven S.; Jan, Muntermann;. (2010). *An intraday market risk managemnt approach based on textual analysis.* Frankfurt.

Hagenau, M., Liebmann, M., & Neumann, D. (2012). *Automated news reading: Stock prices prediction based on financial news using context-capturing features.* Freiburg.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining : concepts and techniques.*

Huyhnh, Huy D.; Dang, Minh L.; Duong, Duc;. (2017). *A New Model for Stock Price Movements Prediction Using Deep Neural Network.* Nha Trang : ACM New York, NY, USA.

Im, Tan Li; San, Phang Wai; On, Chin Kim; Alfred, Rayner; Anthony, Patricia;. (2013). *Analysing Market Sentiment in Financial News using Lexical Approach.* Sarawak.

Ito, Tomoki; Izumi, Kiyoshi; Tsubouchi, Kota; Yamashita, Tatsuo;. (2016). *Polarity Propagation of financial terms for market trend analyses using news articles.* Tokyo.

Joshi, K., & Rao, J. (2016). *Stock Trend Prediction Using News Sentiment Analysis.*

Joshi, Kalyani; Rao, Jyothi; N., Bharathi N. (2016). *Stock Trend Prediction Using News Sentiment Analysis.* Mumbai.

Kaya, Y., & Karsligil, E. (2010). *Stock price prediction using financial news articles.* Istanbul.

Khedr, Ayman Elsayed; Salama, S E; Yaseen, Nagwa. (2018). *Predicitng Stock Market Behaviour using Data Mining Technique and News Sentiment Analyis.* New cairo.

Kim, Y., Jeong, S. R., & Ghani, I. (2014). *Text Opimion Mining to Analyze News for Stock Market Prediction.* Seoul.

Kim, Yoosin; Jeong, Seung Ryul; Ghani, Imran;. (2014). *Text Opinion Mining to Analyze News for Stcok Market Prediction.* Seoul: SRCG Publication.

Lauren, Stefan; Harlili, Dra;. (2014). *Stock Trend Prediction Using Simple Moving Avarage Supported by News Classififcation.* Bandung: IEEE.

Lee, Heeyoung; Mihai, Surdeanu; MacCartney, Bill; Jurafsky, Dan;. (2014). *On the Importance of Text Analysis for Stock Price Prediction.* lrec-conf.org.

Li, Qing; Wang, Tiejun; Li, Ping; Gong, Qixu; Chen, Yuanzhu;. (2014). *The effect of news and public mood on stock movemtns.*

Li, Xiadong; Xie, Haoran; Chen, Li; Wang, Jianping; Deng, Xiaotie;. (2014). *News impact on stock price return vi sentiment analysis .* kowloon: Elsevier B.V.

Matsubara, Takashi; Akita, Ryo; Uehera, Kuniaki. (2018). *Stock Price Prediction by Deep Neural Generative Model of News Articles.* Osaka.

Meesad , Phayung; Li, Jiajia;. (2014). *Stock Trend Relying on Text Mining and Sentiment Analysis with Tweets.* Bangkok.

MITCHELL, C. (2019, 4 16). *Investopedia*. Retrieved from Price Rate Of Change Indicator - ROC Definition and Uses: https://www.investopedia.com/terms/p/pricerateofchange.asp

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ling Ngo, D. C. (2014). *Text mining for market prediction: A systematic review.*

Nikfarjam, A., Emadzadeh, E., & Muthaiyah, S. (2010). *Text mining approaches for stock market predicition.* Cyberjaya.

Patel, Hiral R.; Parikh, Satyen;. (2016). *Comparative Analytical Study for News Text Classification Techniques Applied for Stock Market Price Extrapolation.* Singapore: Springer.

Ramya, M.; Pinakas, J. Alwin;. (2014). *Different Type of Feature Selection for Text Classification.* IJCTT.

Salma Al-Asmari; Mohammed Dahab. (2017). *Sentiment Detection, Recognition and Aspect.* Abha: International Journal of Computer Applications .

Schumaker, R., & Chen, H. (2009). *Textual Analysis of Stock Market Prediction Usineg Financail News Articles.*

Scrapy. (2019, 5 25). *tutorialspoint*. Retrieved from Scrapy - Overview: https://www.tutorialspoint.com/scrapy/scrapy_overview.htm

Shankhdhar, G. (2019, 5 22). *edureka!* Retrieved from Sentiment Analysis Methodology: https://www.edureka.co/blog/sentiment-analysis-methodology/

Shcumaker, Robert P.; Chen, Hsinchun;. (2009). *Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System.* New York.

Shetty, B. (2018, 12 12). *Towards Data Science*. Retrieved from Supervised Machine Learning: Classification: https://towardsdatascience.com/supervised-machine-learning-classification-5e685fe18a6d

Shoushan Li; Sophia Yat Mei Lee; Ying Chen, Chu-Ren Huang; Guodong Zhou. (2002). *Sentiment Classification and Polarity Shifting.* Hong Kong: The Hong Kong Polytechnic University.

Shynkevich, Yauheniya; Coleman, Sonya; Belatreche, Ammar;. (2015). *Predicting Stock Price Movements Based on Different Categories of News Articles.* Nottingham.

Sidana, M. (2017, 2 28). *Medium*. Retrieved from Types of classification algorithms in Machine Learning: https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14

Tang, Bo; Kay, Steven; He, Haibo;. (2016). *Toward Optimal Feature Selection in Naive Bayes for Text Categorization.* IEEE.

Thanh, Hoang T. P.; Meesad, Phayung;. (2013). *Stock Market Trend Prediction Based on Text Mining of Corporate Web and Time Series Data.* Bangkok: Journal of Advanced Computational Intelligence and Intelligent Informatics.

Usmani, Mehak; Adil, Syed Hasan; Raza, Kamran; Azhar Ali, Syed Saad;. (2016). *Stock Market Prediction using Machine Learning Techniques.* Karachi.

Vakeel, Khadija; Dey, Shubhamoy;. (2014). *Impact of News Articles on Stock Prices: An Analysis using Machine Learning.* Bangalore: Permision@acm.org.

Vale, M. N. (2018). *Dow Jones Index Change Prediction Using Text Mining.* Rio de Janeiro: Marcos Neves do Vale.

Vijayarani, S., & Ilamathi, J. (2015). *Preprocessing Techniques for Text Mining - An Overview.* Tamilnadu.

Vu, Tien Thanh; Chang, Shu; Ha, Quang Thuy; Collier, Nigel;. (2012). *An Experiment in Integrating Sentiment Features for Tech Stock Prediction in Twitter.* Milton Keynes: The Open University .

Wang, Wanbin Walter; Ho, Kin Yip; Liu, Wai-Man Raymond; Wang, Kun Tracy;. (2013). *The relation between news evens and stock price jump: an analysis based on neural network.* Canberra.

XLSTAT. (2019, 5 23). *XLSTAT*. Retrieved from The data analysis add-in solution for Microsoft® Excel®: https://www.xlstat.com/en/company/microsoft-partner

Yip, J. (2018, 11 26). *Towards Data Science*. Retrieved from Algorithmic Trading using Sentiment Analysis on News Articles: https://towardsdatascience.com/https-towardsdatascience-com-algorithmic-trading-using-sentiment-analysis-on-news-articles-83db77966704