

Using Tweets Sentiment Analysis to Predict Stock Market Movement

by

Abdulaziz Sulaiman Almohaimeed

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science in Computer Science and Software Engineering

Auburn, Alabama
August 5, 2017

Key words: Sentiment Analysis, Stock Market, Stock Market Prediction,
Ensemble classifier, majority vote classifiers

Copyright 2017 by Abdulaziz Sulaiman Almohaimeed

Approved by

Richard Chapman, Chair, Associate Professor of Computer Science and Software Engineering
Dean Hendrix, Associate Professor of Computer Science and Software Engineering
Dianne Hall, Professor of Information Systems and Business Analytics

Abstract

Research shows that news affects stock market movement and indicates the possibility of predicting the market by using the news as a signal to a coming movement with an acceptable accuracy percentage. In this research, we introduce an approach that predict the Standard & Poor's 500 index movement by using tweets sentiment analysis classifier ensembles and data-mining Standard & Poor's 500 Index historical data. The data-mining is used to extract the major companies influencing the S&P 500 index, ranking these companies, and finding the market patterns. Sentiment analysis classification is used to determine whether a tweet is positive or negative for a certain company. We show in this thesis that using classifier ensembles such as majority voting classifier formed by Decision Tree, Bernoulli Naive Bayes, leaner SVC classifiers with majority voting selection criteria, and random forest classifier perform better than classic classifiers in classify tweets. Using ensembles classifiers to classifying a number of companies' news rather than the all 500 leads to a predication model with an accuracy rate above 80%.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Prof. Richard Chapman for the continuous support of my masters study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my study. Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Dean Hendrix and Prof. Dianne Hall, for their encouragement, insightful comments, and hard questions.

Table of Contents

Abstract.....	ii
List of Tables	vi
List of Illustrations.....	vii
List of Abbreviations	viii
Chapter 1.....	1
Introduction	1
1. Standard & Poor's 500	3
2. Waikato Environment for Knowledge Analysis.....	4
3. Natural Language Toolkit.....	5
4. Scikit Learn Library.....	5
5. TextBlob Library	6
6. System Design	6
Chapter 2.....	8
Historical data analysis.....	8
1. Historical data collection	8
2. Historical data processing.....	8
3. Factors Extraction	8
4. Rules extraction	10
5. Factor Ranking	13
6. Factors Evaluation.....	14
Chapter 3.....	17
Sentiment Analysis.....	17
1. Tweets Sentiment Analysis	17

2. Feature extraction.....	18
3. Niek Sanders datasets.....	19
4. Tweets classification	20
4.1. Classic classifiers	20
4.2. Ensemble classifier.....	20
5. Tweets Classifiers Evaluation	22
Chapter 4.....	24
Stock Market Prediction	24
1. Factor Prediction.....	24
2. Standard & Poor's 500 prediction.....	25
Chapter 5.....	27
Result and Evaluation	27
1. Data.....	27
2. Result	28
3. Future Work	29
4. Thesis Summary and Conclusions	30
References	31

List of Tables

Table 1 Factors ranking	14
Table 2 Factor evaluation.....	16
Table 3 Feature extraction method	18
Table 4 Niek Sanders dataset.....	19
Table 5 Tweets classifiers evaluation	23
Table 6 Testing dataset	24
Table 7 Result	26

List of Illustrations

Figure 1 News fallow up action	2
Figure 2 System Design	7
Figure 3 The relationship between S&P500 index and factors.....	9
Figure 4 The relationship between Google and Amazon stocks.....	9
Figure 5 The relationship between Exxon stock and TWI oil price	10
Figure 6 Majority voting classifier	21
Figure 7 Majority voting classifier example.....	21
Figure 8 Random forest classifier	22

List of Abbreviations

S&P	Standard & Poor's
LMT	Logistic Model Trees
SVC	Support Vector Machine
NLTK	Natural Language Toolkit
NLP	Natural Language Processing

Chapter 1

Introduction

Research shows that news affects stock market movement and indicates the possibility of predicting the market by using the news as a signal to a coming movement with an acceptable accuracy percentage. [4] [3] Our hypothesis is that if a company has positive news it will lead its stock price to increase in the near future. Also, if a company has a negative news it will lead its stock price to decrease in the near future. In this research, we introduce an approach to predict the Standard & Poor's 500 movement using tweets sentiment analysis with classifier ensembles and datamining. The datamining and data analysis is used to extract the major companies influencing the market, rank these factors, and find some of the Standard & Poor's 500 index patterns. This reduces the number of subject that we need to extract tweets about by 97.6 %. The sentiment analysis classification is used to determined whether a tweet about one of the companies is positive or negative.

We choose Twitter as our news source for many reasons such as:

- News wire APIs cost a lot of money. For example, a Bloomberg account costs about \$2,000 a month, while the Twitter API is free, and we can get Bloomberg news using their account tweets.
- More than 60% of U.S. adults get news on social media. [23]
- More than 60% of Twitter users get their news on the site. [21]
- Using Twitter as our news source allows us to get the most up to date news.
- All major news wire companies have accounts in Twitter.
- Social media and news websites are the most common pathways to online news. [22]
- It is more likely that a fallow up action will be made if the news subject about business and finance or about Government and politics. [22] (See Figure 1). [23]

In this thesis, we reach a result that using classifier ensembles, gives better performance. We use a majority vote classifier formed by Decision Tree, Bernoulli Naive Bayes and learner SVC classifiers focusing on classifying a number of companies' news rather than all 500. This leads us to a predication model with an accuracy rate of 91.42% for single company predictions and a 81.94% accuracy rate to predict the Standard & Poor's 500 index daily movement.

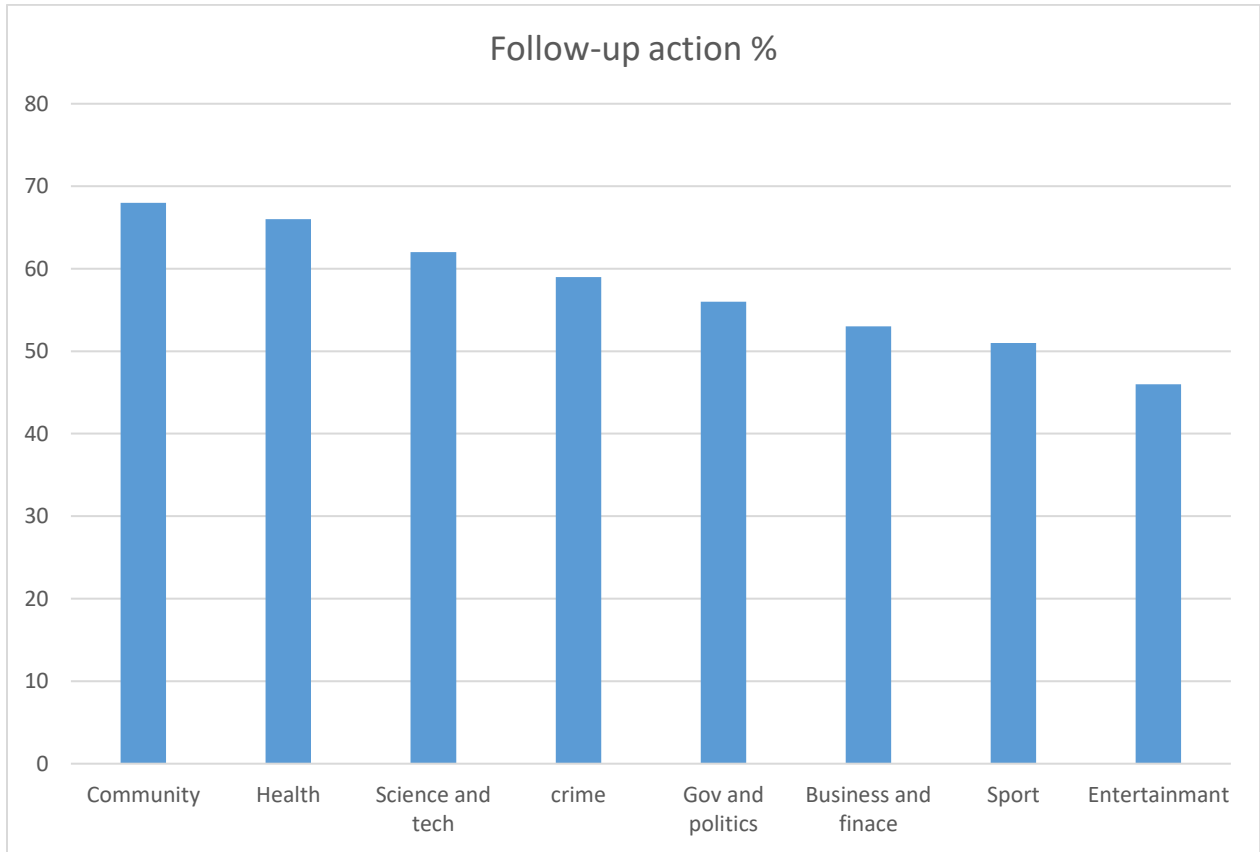


Figure 1

1. Standard & Poor's 500

The S&P 500 is generally regarded as the best single measure of large-cap U.S. equities. There is over USD 7.8 trillion benchmarked to the index, with index assets including almost USD 2.2 trillion of this total. The index includes 500 leading companies and captures around 80% coverage of available market capitalization. [10]

The S&P 500 was called the "Composite Index" when it announced its first stock index in 1923. [11] After three years the Composite Index extended to cover 90 stocks and then in 1957 it expanded to its current 500 companies.[11] Standard & Poor's is a company that was instituted in 1860 by Henry Varnum Poor. In 1941 Poor's Publishing (Henry Varnum Poor's original company) merged with Standard Statistics (founded in 1906 as the Standard Statistics Bureau), and assumed the name Standard and Poor's Corporation. The S&P 500 index in its present form began on March 4, 1957. Technology has allowed the index to be calculated and disseminated in real time. The S&P 500 is commonly used as a measure of the general level of stock prices, as it includes both growth stocks and value stocks.

To calculate the S&P 500 Index value, the sum of the adjusted market capitalization of all 500 stocks is divided by a factor, usually stated as the Divisor.[12][13] For example, if the total adjusted market cap of the 500 component stocks is US\$13 trillion and the Divisor is set at 8.933 billion, then the S&P 500 Index value would be 1,455.28. Although the adjusted market capitalization of the entire index can be accessed from Standard & Poor's website,[14] the Divisor is considered to be proprietary to the firm. However, the Divisor's value is approximately 8.9 billion.[15]

The formula to calculate the S&P 500 Index value is:

$$\text{Index Level} = \frac{\sum (P_i \cdot Q_i)}{\text{Divisor}}$$

where P is the price of each stock in the index and Q is the number of shares publicly available for each stock. In a stock issuance event the divisor is adjusted for spin-offs or similar structural changes, to make sure that such a case does not modify the Index value.[12]

2. Waikato Environment for Knowledge Analysis

The Waikato Environment for Knowledge Analysis, commonly known as Weka, is a workbench that contains a collection of visualization tools and algorithms for data analysis and predictive modeling. [1] Weka is widely used in a number of different application areas, especially for educational purposes and research. Weka Advantages include:

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and it can be run on almost computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Weka graphical user interfaces make it easy to.

Weka supports numerous standard data mining tasks, specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are based on the assumption that the data is accessible as one flat file or relation (database), where each data point is defined by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka offers access to SQL databases using Java Database Connectivity and can process the result obtained by running a database query. [1]

3. **Natural Language Toolkit**

The Natural Language Toolkit, commonly known as NLTK, consists of libraries and programs for symbolic and statistical natural language processing (NLP). NLTK mainly supports research and teaching in NLP and related subjects, as well as empirical linguistics, cognitive science, artificial intelligence, information retrieval, and machine learning [24]. NLTK is used in a number of ways such as a teaching tool, an individual study tool, and a platform for prototyping and building research systems. More than 30 universities in the United States and 25 other countries are using NLTK in their courses. [25] NLTK supports classification, tokenization, stemming, tagging, parsing, and semantic reasoning functionalities. Also it offers an easy-to-use interface to more than 50 corporate and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

4. **Scikit Learn Library**

Scikit Learn is a machine learning library software for the Python programming language.[5] It features numerous classification, regression and clustering algorithms such as support vector machines, random forests, gradient boosting, k-means and DBSCAN. Scikit Learn library is designed to work with the Python numerical and scientific libraries NumPy and SciPy. NumPy is an extension to the Python programming language to support large, multi-dimensional arrays and matrices, along with a great library of high-level mathematical functions to operate on these arrays. [19] SciPy is an open source Python library for scientific computing and technical computing. SciPy includes modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, and additional tasks in science and engineering.

5. TextBlob Library

TextBlob is a Python (2 and 3) library for processing textual data. It offers an easy to use API for diving into common natural language processing (NLP) tasks. TextBlob has many features such as: [9]

- Noun phrase extraction
- Part-of-speech tagging
- Sentiment analysis
- Classification (Naive Bayes, Decision Tree)
- Language translation and detection powered by Google Translate
- Tokenization (splitting text into words and sentences)
- Word and phrase frequencies
- Parsing
- n-grams
- Word inflection (pluralization and singularization) and lemmatization
- Spelling correction
- Add new models or languages through extensions
- WordNet integration

6. System Design

There are three main phases in this thesis, which are historical data analysis, sentiment analysis and stock market prediction. The historical data analysis phase is where we study the S&P 500 history. The result is finding which factors have a major influence on the market that we can use to predict the S&P 500. In the second phase the sentiment analysis phase we extract the news features, which are used to build a number of sentiment analysis classifiers and then test them to find the most accurate classifier to be used in this thesis. Phases one and two are used in the third phase the stock market prediction by classifying tweets to predicting factors, predict the S&P 500, and test our prediction on real data to evaluate the approach. (See Figure 2)

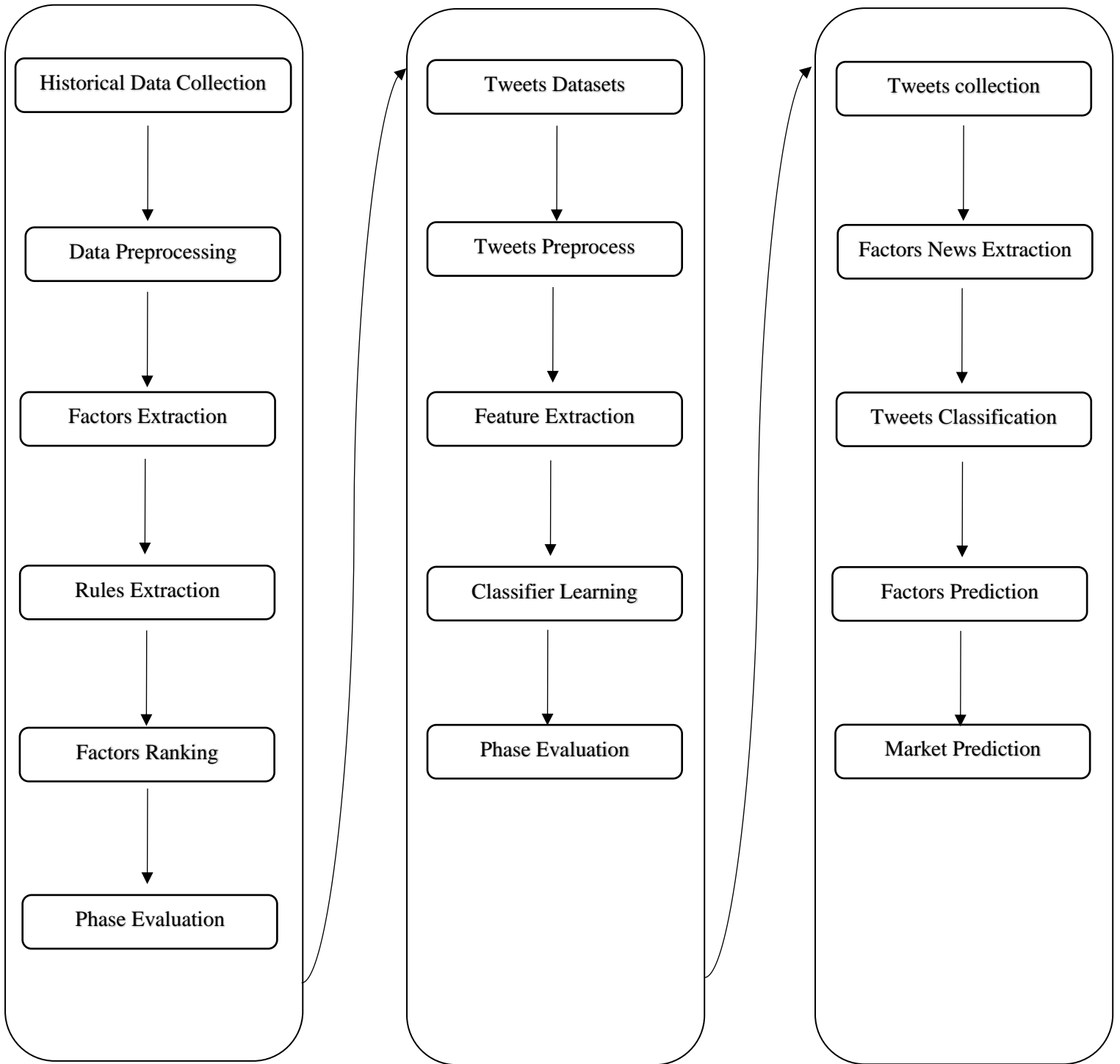


Figure 2

Chapter 2

Historical data analysis

1. Historical data collection

We collected historical data related to the economy and the S&P 500 market such as the Standard & Poor's 500 index values and the companies' stock prices, from <https://www.google.com/finance> and <https://finance.yahoo.com> for the past ten years from 10/16/2006 to 01/17/2017 in a csv format. The data has a different representation between the two sources, (e.g. in Yahoo Finance we get the adjusted closing price as an extra column, which is the stock's closing price on any given day of trading that has been corrected to include any distributions and corporate actions that occurred at any time prior to the next day's opening).

2. Historical data processing

To get more information from the historical data that we collected, we represented the data in binary form (0 and 1), where 0 means that the value of the company or the index in this date is less than the day before, and 1 means that the value of the company or the index in this date is greater than the day before. This data representation allowed us to study the data in a comprehensive way.

3. Factors Extraction

To understand the market and how it behaves, we visualized and analyzed the data using the Weka tool. We found that some companies have more effect on the market than the others by using the data visualization, which used images to represent the data to view it in a better way. There are a large number of conventional ways to represent the data, such as bar graphs, pie charts, histograms, table etc. The main goal is to extract as many details from the data as we can. Using this approach, three patterns appear in the data. The first pattern is that the change percentage in the stock price of Amazon, Microsoft, Apple, Google, Ford, Exxon, Goldman, Morgan, J.P. Morgan and Johnson & Johnson is like the change percentage in the S&P500 index value (see Figure 2), which reveals that there is a relation between these companies and the S&P500 index. The second two additional

patterns were that the prices of (Amazon , Google) and (Exxon , TWI oil price) have analogous movements, which leads to the idea that if there is an effect on one it is more likely to affect the other. See Figure 3 and Figure 4.

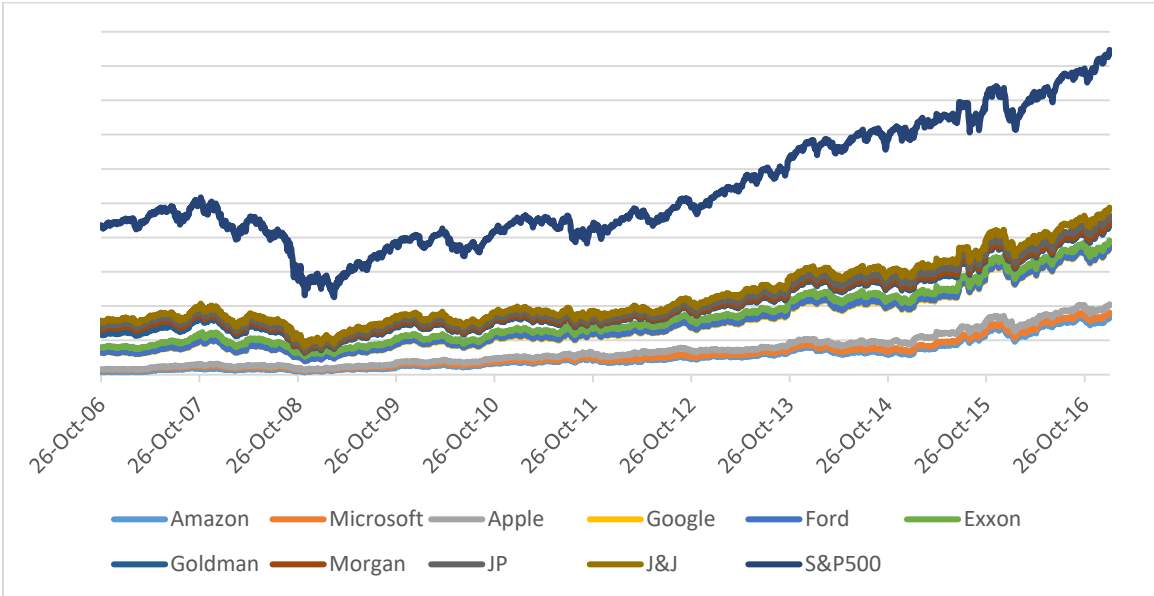


Figure 3

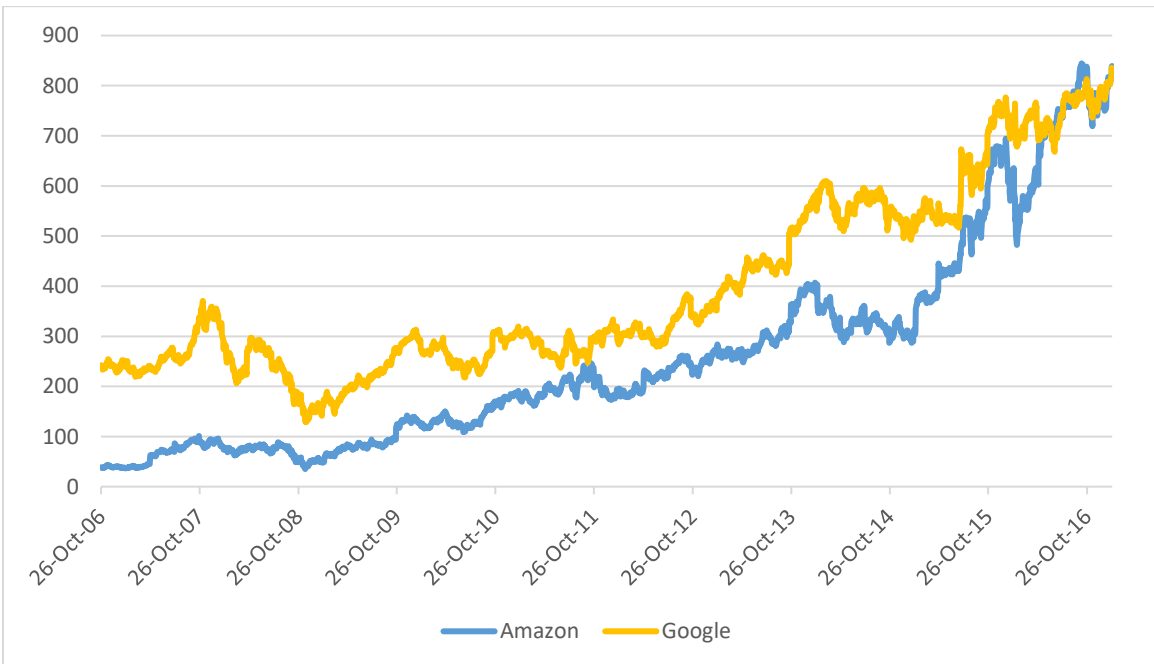


Figure 4

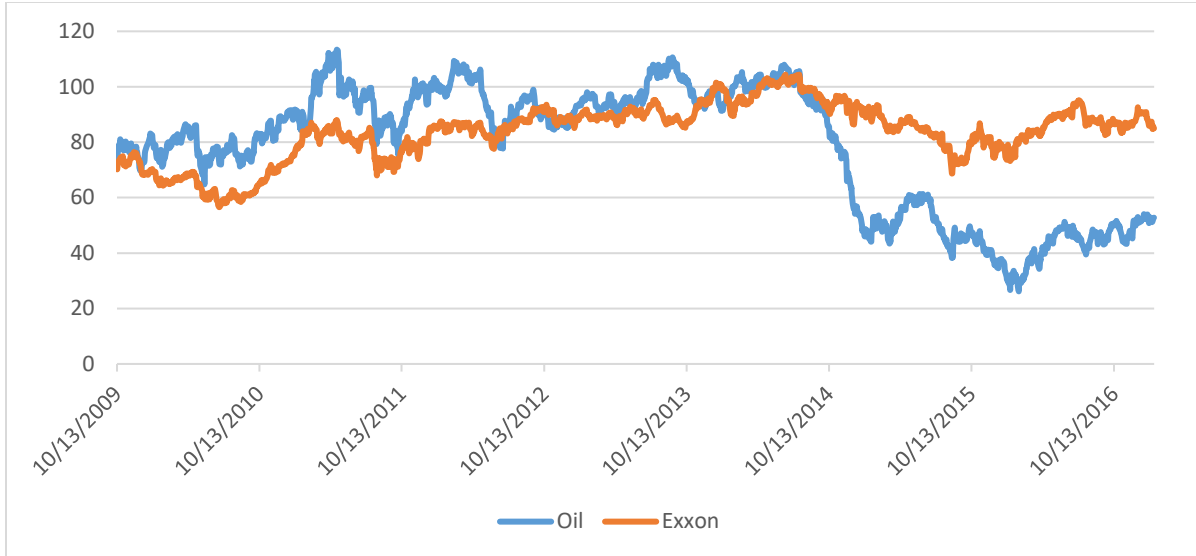


Figure 5

4. Rules extraction

To identify the relationship between S&P 500 and the stock prices, we used an apriori type algorithm, which is an algorithm for frequent item set mining and association rule learning over transactional databases, [8] by iteratively reducing the minimum support until it finds the required number of rules with the given minimum confidence. The algorithm has an option to mine class association rules. The result we collect by using the apriori algorithm can be used to determine association rules which highlight general trends in the dataset. The algorithm results in 23 rules with confidence rate between 97% – 93%:

1. If Ford stock is up, and Exxon stock is up, and Morgan Stanley stock is up, and J.P Morgan. stock is up, then the S&P500 index is up with confidence rate of 97%.
2. If Ford stock is up, and Exxon stock is up, and Goldman stock is up, and J.P Morgan stock is up, then the S&P500 index is up with confidence rate of 97%.

3. If Ford stock is up, and Exxon stock is up, and Goldman stock is up, and Morgan Stanley stock is up, then the S&P500 index is up with confidence rate of 97%.
4. If Exxon stock is up, and Morgan Stanley stock is up, and J.P Morgan stock is up, and Johnson & Johnson stock is up, then the S&P500 index is up with confidence rate of 97%.
5. If Exxon stock is up, and Goldman Sachs stock is up, and Morgan Stanley stock is up, and Johnson & Johnson stock is up, then the S&P500 index is up with confidence rate of 97%.
6. If Ford stock is up, and Exxon stock is up, and Morgan Stanley stock is up, then the S&P500 index is up with confidence rate of 96%.
7. If Ford stock is up, and Exxon stock is up, and Goldman Sachs stock is up, then the S&P500 index is up with confidence rate of 96%.
8. If Ford stock is up, and J.P. Morgan stock is up, and Johnson & Johnson stock is up, then the S&P500 index is up with confidence rate of 96%.
9. If Ford stock is up, and Morgan stock is up, and Johnson & Johnson stock is up, then the S&P500 index is up with confidence rate of 96%.
10. If Exxon stock is up, and Morgan Stanley stock is up, and Johnson & Johnson stock is up, then the S&P500 index is up with confidence rate of 95%.
11. If Exxon stock is up, and Goldman Sachs stock is up, and Morgan Stanley stock is up, and J.P. Morgan stock is up, then the S&P500 index is up with confidence rate of 95%.

12. If Ford stock is up, and Exxon stock is up, and J.P. Morgan stock is up, then the S&P500 index is up with confidence rate of 95%.
13. If Exxon stock is up, and J.P. Morgan stock is up, and Johnson & Johnson stock is up, then the S&P500 index is up with confidence rate of 95%.
14. If Exxon stock is up, and Goldman Sachs stock is up, and Johnson & Johnson stock is up, then the S&P500 index is up with confidence rate of 95%.
15. If Exxon stock is up, and Goldman Sachs stock is up, and J.P. Morgan stock is up, then the S&P500 index is up with confidence rate of 95%.
16. If Ford stock is up, and Goldman Sachs stock is up, and Johnson & Johnson stock is up, then the S&P500 index is up with confidence rate of 95%.
17. If Ford stock is up, and Exxon stock is up, and Johnson & Johnson stock is up then the S&P500 index is up with confidence rate of 95%.
18. Exxon stock is up, and Morgan Stanley stock is up, and J.P. Morgan stock is up, then the S&P500 index is up with confidence rate of 94%.
19. If Ford stock is down, and Exxon stock is down, and Goldman Sachs stock is down, and J.P. Morgan Stanley stock is down, then the S&P500 index is down with confidence rate of 94%.
20. If Exxon stock is down, and J.P. Morgan stock is down, and Johnson & Johnson stock is down, then the S&P500 index is down with confidence rate of 94%.

21. If Goldman Sachs stock is up, and Morgan Stanley stock is up, and J.P. Morgan stock is up and Johnson & Johnson stock is up, then the S&P500 index is up with confidence rate of 94%.
22. If Ford stock is up, Goldman Sachs stock is up, Morgan Stanley stock is up and J.P. Morgan stock is up then the S&P500 index is up with confidence rate of 94%.
23. If Exxon stock is down, Morgan Stanley stock is down and Johnson & Johnson stock is down, then the S&P500 index is down with confidence rate of 94%.

5. Factor Ranking

In this thesis, the companies mentioned in the above rules are called the influencing factors. To know the influence of each factor on the S&P500 index, we ranked these factors using correlation attribute evaluation and Pearson correlation coefficient, which uses a measure of the linear correlation between two variables X and Y. It has a value between 1 and -1 inclusive, where 1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. [27]

It is extensively used in the sciences. In this thesis X is our factor and Y is the S&P 500 index. This algorithm ranks the factors by the following ranking J.P. Morgan, Exxon, Morgan, Goldman, Johnson & Johnson, Ford, Apple, Microsoft, Google, Amazon. See Table 1. Having this information allowed us to understand of S&P 500 and knew what news and information we needed to extract from twitter.

Ranked	Attributes
1	J.P. MORGAN
2	Exxon
3	Morgan Stanly
4	Goldman
5	Johnson & Johnson
6	Ford
7	Apple
8	Microsoft
9	Google
10	Amazon

Table 1

6. Factors Evaluation

Using the processed historical data representation for the factors and the S&P 500 index, we built a number of classifiers and tested them. The test allowed us to evaluate the factors extraction results and to determine whether with only these factors we can predict the S&P 500 index movement or not. Five classifier algorithms were used; J48, LMT, REP Tree, Hoeffding Tree and Decision Table.

We used the ZeroR classifier as our baseline. It is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category. The result of using this classifier was a correctness rate of 57.55%.

Using the historical data collected, we created a dataset that represents these factors and the S&P 500 index in binary representations, where 0 means the factor or index value is lower than the previous day and 1 means that the factor or index value is more than the previous day.

The total number of data instances in the dataset is 2579. By taking the average of the five classifiers test result we showed that if the system knows the movements of these factors it will know the S&P500 index daily movement with a correctness rate above 80%. See Table 2. Having this information reduced the number of subjects that we need to extract tweets about by 98 %. Out of 500 companies, we only focused on these ten factors.

		Testing options				
Classification Algorithms	Correctly Classified	10-Cross Validation	15-Cross Validation	80% Split	90% Split	
	J48	81.7759 %	81.8922 %	81.5891 %	85.2713 %	
	LMT	83.4044 %	83.133 %	84.8837 %	85.2713 %	
	REP Tree	82.4738 %	82.0861 %	82.9457 %	84.8837 %	
	Hoeffding Tree	82.9391 %	82.8616 %	85.0775 %	83.7209 %	
	Decision Table	81.6596 %	82.3187 %	83.1395 %	84.1085 %	
	<hr/>					
	#Correctly Classified	10-Cross Validation	15-Cross Validation	80% Split	90% Split	
	J48	2019 / 2579	2112 / 2579	421 / 516	220 / 258	
	LMT	2151 / 2579	2144 / 2579	438 / 516	220 / 258	
	REP Tree	2127 / 2579	2093 / 2579	428 / 516	219 / 258	
	Hoeffding Tree	2138 / 2579	2137 / 2579	439 / 516	216 / 258	
	Decision Table	2106 / 2579	2123 / 2579	429 / 516	217 / 258	
	<hr/>					
	ROC Area	10-Cross Validation	15-Cross Validation	80% Split	90% Split	
	J48	0.870	0.868	0.867	0.890	
	LMT	0.919	0.919	0.942	0.948	
	REP Tree	0.881	0.880	0.892	0.912	
	Hoeffding Tree	0.915	0.915	0.930	0.929	
	Decision Table	0.906	0.907	0.905	0.940	
	<hr/>					
	Precision	10-Cross Validation	15-Cross Validation	80% Split	90% Split	
	J48	0.820	0.817	0.838	0.823	
	LMT	0.842	0.840	0.884	0.897	
	REP Tree	0.819	0.811	0.830	0.870	
	Hoeffding Tree	0.847	0.846	0.856	0.837	
	Decision Table	0.826	0.834	0.838	0.878	
	<hr/>					
	Recall	10-Cross Validation	15-Cross Validation	80% Split	90% Split	
	J48	0.849	0.856	0.811	0.910	
LMT	0.853	0.849	0.848	0.850		
REP Tree	0.866	0.849	0.886	0.876		
Hoeffding Tree	0.835	0.834	0.865	0.850		
Decision Table	0.837	0.840	0.847	0.850		

Table 2

Chapter 3

Sentiment Analysis

Sentiment analysis is the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically recognize, extract, quantify, and study emotional states and subjective information. [30] Usually, sentiment analysis' purpose is to determine the attitude of a speaker, writer, or other subject with respect to overall contextual emotional reaction to a document or event.

1. Tweets Sentiment Analysis

There are two approaches to using sentiment analysis classifiers. The classic approach is to train one classifier from the training set, and the second approach is train a number of classifiers and combine them to solve the same problem. This is known as the classifier ensembles method [6]. Using the ensemble based approach gives a more effective result for three reasons as Dietterich mentions [20]:

1. **Statistical:** Let's say we use the three classifier algorithms Naive Bayes, SVM and Decision Tree, and each one gives an acceptable accuracy in the training set. if we chose one of them it may not get the best result in unseen data. but by combining all three classifiers we reduce the risk of selecting an insufficient classifier [18].
2. **Computational:** Many classifying algorithms work by applying some sort of local search that may get stuck in local optima which may be far from global optima. For example, decision tree algorithms employ a greedy splitting rule and neural networks algorithms employ gradient descent to minimize an error function over the training set. An ensemble constructed by running the local search from many different starting points may provide a better approximation than any of the individual classifiers [18].

3. **Representational.** In many applications of machine learning the problem is too complicated to be solved by one model. But by combining two or more model it may be possible to solve the problem [18].

Other reasons for combining different classifiers include [26] the following. (1) Training several different classifiers on the same data may show differences in their global and local performances. Each one may have its own area in the feature space where it performs better. (2) Sometimes more than one training set is available, each collected at a different time or in a different environment. These training sets may even use different features.

2. Feature extraction

To extract features from tweets we used a table to represent the tweets, where columns represents words related to finance, economics and influencing factors overall, and the values represent whether these words exist in the tweet or not. Tweets are represented as demonstrated in Table 3. where there are n tweets and m words. A vector represents each tweet as $\mathbf{vt}_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{im})$, in which x_{ij} represent wether the word w_j is in the tweet \mathbf{vt}_i .

	w_1	w_2	w_3	...	w_m
\mathbf{vt}_1	x_{11}	x_{21}	x_{31}	...	x_{m1}
\mathbf{vt}_2	x_{12}	x_{22}	x_{32}	...	x_{m2}
\mathbf{vt}_3	x_{13}	x_{23}	x_{33}	...	x_{m3}
...
\mathbf{vt}_n	x_{n1}	x_{n2}	x_{n3}	...	x_{nm}

Table 3

3. Niek Sanders datasets

Niek Sanders' dataset has 5513 hand-classified tweets. These tweets were classified with respect to one of 4 different topics. Apple, Google, Microsoft and Twitter. Each entry contains tweet id, tweet text, tweet creation date, topic used for sentiment and sentiment label: 'positive', 'neutral', 'negative', or 'irrelevant'. [8] See table 4:

	# Positive	# Negative	# Neutral	# Irrelative	Twitter search term
Apple	191	377	581	164	@apple
Google	218	61	604	498	#google
Microsoft	93	138	671	513	#microsoft
Twiiter	68	78	647	611	#twitter

Table 4

The corpus classifications are:

Positive;

- Positive indicator on topic

Neutral:

- Neither positive nor negative indicators
- Mixed positive and negative indicators
- On topic, but indicator undeterminable
- Simple factual statements
- Questions with no strong emotions indicated

Negative:

- indicator on topic

Irrelevant:

- Not English language

- Not on-topic (e.g. spam)

“Positive” and “Negative” labels were reserved for tweets which clearly express an emotion or where the implications were unambiguous. As a rule of thumb, “neutral” was the preferred label for borderline cases. [8]

4. Tweets classification

In tweets classification, we used two type of classifiers, the classic classifiers and the ensemble classifiers to evaluate which one is the best for sentiment analysis. [28] Tweets are classified as either positive or negative with regards to the related factor. These classifiers are trained using the Niek Sanders datasets. In addition we labeled 200 tweets.

4.1. Classic classifiers

In order to find which algorithms perform well and should be used in the ensemble classifiers model, eight different learning algorithms were used; Naïve Bayes, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Decision Tree, Logistic Regression, Stochastic Gradient Descent, SVM and learner SVM. Naïve Bayes. We created a classifier for each algorithm and used Niek Sanders’ dataset to train and test these classifiers. Decision Tree and learner Support vector machines outperform other classifiers. See Table 5.

4.2. Ensemble classifier

Much research shows that using ensemble classifiers will increase the accuracy rate. [28] Two ensemble classifiers were used. The first ensemble classifier was built using the best three classic classifiers in a majority vote model, which combines output from different learners. This led to a decrease in variance error. [29] See Figure 6 and 7.

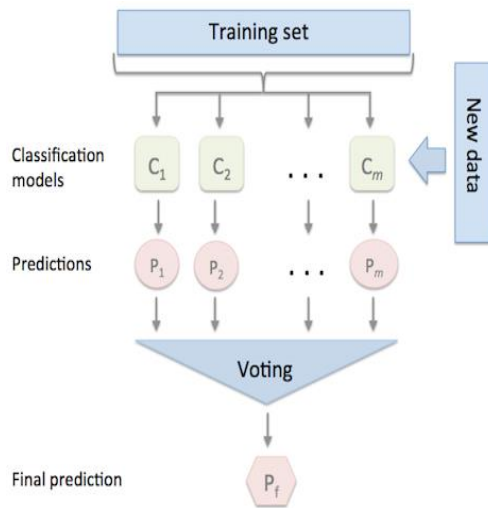


Figure 6

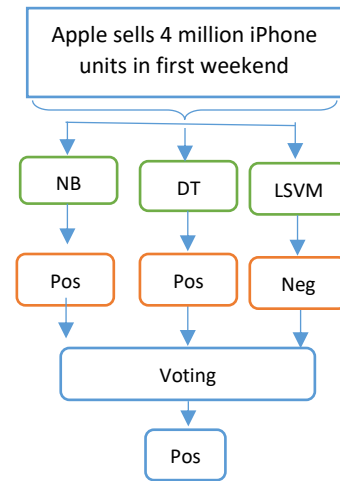


Figure 7

The second ensemble classifier is the random forest classifier which is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size (see Figure 8). Ensemble classifiers give better performance than the other classifiers. The random forest classifier was selected to be the tweets sentiment analysis classifier since it outperforms all other classifiers. See Table 5.

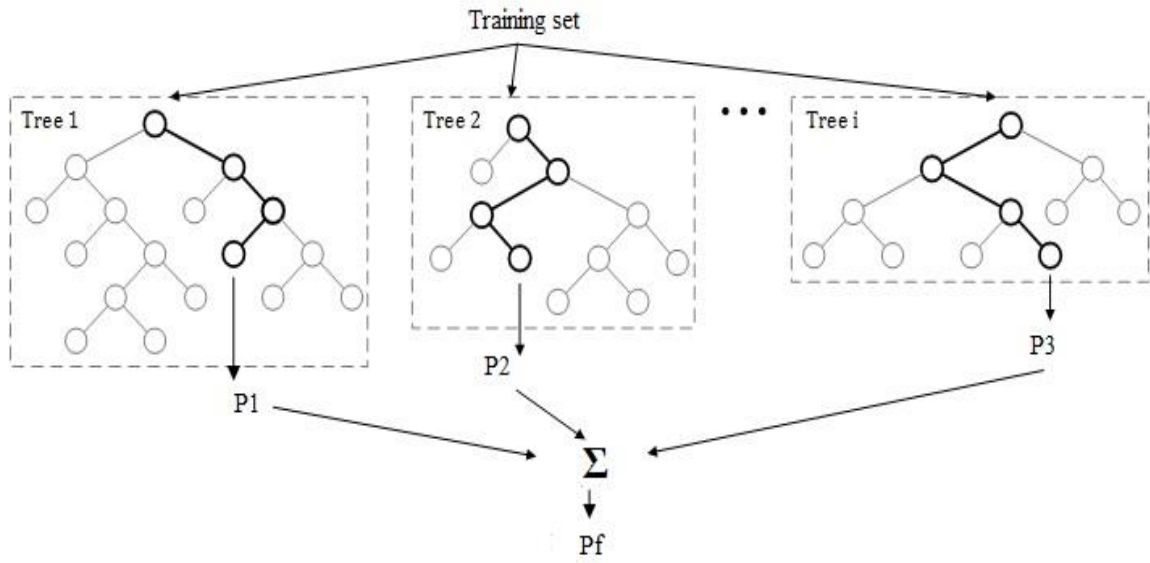


Figure 8

5. Tweets Classifiers Evaluation

To evaluate the classifiers, we divided the datasets into learn and test sets (used to perform three tests for each classifier). The first test performs a test with data split 50% learning set and 50% testing set. The second test was split 80% learning set and 20% testing set. The third was split 90% learning set and 10% testing set. We used the Scikit Learner Dummy Classifier as a baseline, which is a classifier that makes predictions using simple rules. The accuracy baseline is 68.9%. Classic classifiers give a performance with an accuracy rate above 81%, and ensemble classifiers give better performance reaching 88.6% accuracy rate (see Table 5).

Classification Algorithm	50% split	80% split	90% split
Naïve Bayes	86.4%	84.7%	84.4%
Multinomial Naïve Bayes	85.8%	84.4%	84%
Bernoulli Naïve Bayes	86.2%	84.6%	84.3%
Decision Tree	87.4%	85.5%	85.4%
Logistic Regression	85.4%	84.2%	84%
Stochastic Gradient Descent	84.8%	84.5%	84.5%
Linear SVC	86.4%	84.6%	84.2%
SVC	82.4%	81.6%	81.4%
Majority Vote	86.8%	85%	84.7%
Random Forest	88.6%	87%	86.7%

Table 5

Chapter 4

Stock Market Prediction

1. Factor Prediction

Predicting each factor is a key part of the market prediction phase. The system reads all the tweets, text and classifies them using the random forest classifier. The classifier has two classes either 1 or -1, 1 for positive tweets and -1 for negative tweets. Because the factors may have more than one tweets in each day the system gives an average score for each day between 1 and -1. By using the equations below, we predict each factor.

$$p(f) = \frac{\sum_{i=1}^n f_i}{n}$$

In this equation n is the number of tweets related to the factor in the same day, and f is the classifier value of the tweet. If $p(f) > 0$ that means the stock price will increase next day and if $p(f) < 0$ that means the stock price will decrease next day and 0 for no change. To evaluate each prediction, we use the equation below.

$$m(f) = V_{f_{d+1}} - C_{f_d}$$

In this equation V_f is the value of the stock, and C_f is the stock closing price and d is the date of the prediction. To determine whether a prediction is a hit or a miss we use $p(f)$ and $m(f)$. A prediction p is a hit if and only if $p(f)$ and $m(f)$ are both positive, $p(f)$ and $m(f)$ are both negative or $p(f)$ and $m(f)$ are both equal 0. In any other case the prediction p considered as a miss. For example, if we have three tweets related to Amazon stock on 1-3-2017, two of them are positive and one negative, the system predict that the stock price will increase in the next day ((+1, +1,

-1), 0.33). Taking the difference between Amazon stock closing price on 1-3-2017 and the price on the date 1-4-2017 ($757.18 - 753.67 = 3.51$), both the prediction and the stock movement are more than 0, which means the prediction is a hit. Using this approach in the factor prediction phase allows us to reach an accuracy rate of 87%. See Table 7.

2. Standard & Poor's 500 prediction

To predict the market, we need to predict each factor by applying the factor prediction to all the factors, as some factor may have more than one tweet per day and some may not have any. To solve this problem, we use the rules extracted in chapter 1. Applying factor prediction to the factors that have related tweets and extracting the number of rules is applied to the factors. Each rule uses three or four factors to predict the market and gives one of two output either 1 or -1, where 1 means that the S&P 500 index will increase next day and -1 means that the S&P 500 index will decrease next day. These rules are sorted based on the confidence rate and based on the number of rules extracted. The prediction is made using the following equation.

$$mp = \sum_{i=1}^n ri + a(e), \quad n \geq 3$$

$$mp = rc + a(e), \quad n < 3$$

In this equation n is the largest odd number of rules applied to the factors, r is the rule value, rc is the rule with the highest confidence rate and $a(e)$ is the economic policy uncertainty level using the economy tweets which focus on events that effects the economy overall such as the nonfarm payroll employment report, new laws, wars, elections, economy stability and US dollar movements. US economic policy uncertainty reach high level in presidential elections,

Gulf Wars I and II, the 9/11 attacks, the Lehman Brothers bankruptcy, the 2011 debt-ceiling disagreement and other key battles over economic policy. [16] [17] The function $a(e)$ classifies the tweets about these subjects. The result is a number between 1 and -1, where 1 means that the external influence is positive, -1 means that the external influence is negative and 0 is used for neutral influence.

Using a combination of rule and economic policy uncertainty level to predict the S&P 500 reduces the prediction's margin of error. The result for both equations is a number between 1 and -1, where 1 means that the S&P 500 index will increase next day, and -1 means that the S&P 500 index will decrease next day. To evaluate each prediction, we use the equation below to compare the prediction with the S&P 500 index's real change between the index value in the prediction date and the next day.

$$m(x) = V_{d+1} - V_d$$

In this equation V is the S&P 500 index closing value, and d is the date of the prediction. To determine whether a prediction is a hit or a miss we use the mp result and $m(x)$. A prediction mp is a hit only and only if mp result and $m(x)$ are both positive, mp and $m(x)$ are both negative or mp and $m(x)$ are both equals 0. In any other case the prediction mp considered as a miss.

Chapter 5

Result and Evaluation

To estimate the quality of the predictions generated using our model we manually evaluated a sample of 1,102 predictions using tweets that occurred between the date 04-08-2007 and 01-15-2017.

1. Data

For evaluation purposes, we collected tweets from 120 reliable twitter accounts. All the accounts are business or economy related such as Bloomberg, CNN Money, Yahoo Finance, Reuters Business and Financial Times. A total of 387,200 tweets were collected.

To eliminate unrelated and redundant tweets, the 387,200 tweets were scanned and sorted based on which factor it belong to and stored in one of the 12 vectors that represent the factors, if it has been not stored yet. From the 387,200 tweets, 41,420 were extracted and saved. Each factor has its own vector containing tweets text and date. See Table 6.

Factor	Number of Tweet
Apple	1,879
Google	987
Amazon	890
Ford	337
Microsoft	768
Exxon	1,638
Johnson & Johnson	305
Jp	208
Stanly	1,406
Oil	6,584
Gold	4,038
Econ	21,040

Table 6

2. Result

We found that our approach can reach stable results in individual stock predictions and S&P 500 index predictions over a large amount of data (ten years of stock prices and more than 300,00 tweets). The results of the evaluation are displayed in Table 7. The table shows the precision of our model in individual stock predictions and the S&P 500 index predictions. We set our baseline at 70%. Applying our approach for individual stock predictions, we reached a precision of 91.42% in predicting companies stock prices movement for the next day. Also for the S&P 500 we reached a precision of 81.94% in predicting the market index movement for the next day. Out of 1,102 days we correctly predicted 903 days.

Company	Precision
Apple	86.25%
Google	85.07%
Amazon	87%
Ford	85.21%
Microsoft	84.31%
Exxon	78.65%
Johnson & Johnson	91.42%
J.P. Morgan	82.02%
Morgan Stanly	81.78%
Goldman	86.01%
S&P 500	81.94%

Table 7

3. Future Work

While this thesis has demonstrated the potential of efficiently predicting individual stock prices and S&P 500 index movement by using sentiment analysis, many different tests and experiments have been left for the future due to lack of time (i.e. the experiments with real data are usually very time consuming, requiring even days to finish a single run). Future work concerns deeper analysis of mechanisms, new proposals to try different methods, or simply curiosity.

There are some ideas that we would have liked to try during the features extraction, creating sentiment analysis classifiers and predicting the market. The following ideas could be tested:

1. Extending the thesis scope to cover more languages and markets that have major influence on global economy such as Japan Exchange Group, Shanghai Stock Exchange and Deutsche Borse AG.
2. Applying our prediction model over more time periods related to stock markets (5 days, one month, quarter).
3. Applying several different methods in creating sentiment analysis classifiers such as structured events and semantic frames.

4. Thesis Summary and Conclusions

In Chapter 2, we analyze large-scale market historical data. It was clear from Section 3, 4 and 6 that S&P 500 can be predicted by applying a set of rules that define the relationship between ten company's stocks and the S&P 500 index. This information reduced the number of factors we focus on by 98%.

In Chapter 3, we created and trained two sets of sentiment analysis classifiers classical and ensemble, to classify tweets. To increase the classification accuracy rate, we applied different methods in creating the ensemble classifiers and evaluate both sets. Experimental results showed that ensemble classifiers perform better than the classical classifiers in classifying tweets reaching an accuracy rate of 88.6%.

In Chapter 4, we employed Chapter 2 and 3 results in creating the S&P 500 prediction algorithms and equations. We used the rules extracted in chapter 2 and the best classifier in chapter 3 to predict the stocks movements. We also considered the external influences to the market by evaluating the Economic Policy Uncertainty level, which enhanced the prediction model.

In this thesis, we have presented a reliable new approach to predict stock price movement. Combining the trained system on S&P 500 stock historical data and classifying company's news to predict the stock price movement in one system, experimental results showed that this approach outperformed the baseline by 14.51% on average.

References

1. Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016. <http://www.cs.waikato.ac.nz/ml/weka/index.html>
2. P. Reutemann; B. Pfahringer; E. Frank (2004). "Proper: A Toolbox for Learning from Relational Data with Propositional and Multi-Instance Learners". 17th Australian Joint Conference on Artificial Intelligence (AI2004). Springer-Verlag.
3. Alan Ritter, Mausam, Oren Etzioni, Sam Clark. "Open Domain Event Extraction from Twitter". 2012.
4. Xiao Ding, Yue Zhang, Ting Liu, Junwen Duan. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation, Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1415–1425, October 25-29, 2014, Doha, Qatar.
5. Fabian Pedregosa; Gaël Varoquaux; Alexandre Gramfort; Vincent Michel; Bertrand Thirion; Olivier Grisel; Mathieu Blondel; Peter Prettenhofer; Ron Weiss; Vincent Dubourg; Jake Vanderplas; Alexandre Passos; David Cournapeau (2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research. 12: 2825–2830.
6. Z. Zhou, Ensemble Methods: Foundations and Algorithms, Chapman & Hall/CRC Data Mining and Knowledge Discovery Serie, Taylor & Francis, 2012.
7. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
8. Nick Sanders. Sentiment Analysis, <http://sananalytics.com/index.php> , Retrieved October 24, 2011.

9. Steven Loria . TextBlob: Simplified Text Processing. 2017
<https://textblob.readthedocs.io/en/dev/extensions.html#available-extensions>
10. S&P Dow Jones Indices: S&P U.S. Indices Methodology, 2017.
<https://us.spindices.com/indices/equity/sp-500>
11. The Editors of Encyclopædia Britannica, "S&P 500 – stock market". Retrieved 8-11-2005.
<https://www.britannica.com/topic/SandP-500>.
12. S&P Indices Index Mathematics Methodology. The McGraw-Hill Companies, Inc.
January 20, 2013.
13. S&P Dow Jones Index Mathematics Methodology. Retrieved March 2014
<http://www.spindices.com/documents/index-policies/methodology-index-math.pdf>.
14. S&P 500 Details. Standard & Poor's. January 20, 2013.
15. How is the value of the S&P 500 calculated? . Retrieved January 20, 2013.
<http://www.investopedia.com/ask/answers/05/sp500calculation.asp>
16. Scott R. Baker, Nicholas Bloom, Steven J. Davis, Measuring. “Economic Policy Uncertainty”, NATIONAL BUREAU OF ECONOMIC RESEARCH, October 2015.
<http://www.policyuncertainty.com/media/BakerBloomDavis.pdf>
17. Li Liu, Tao Zhang. “Economic policy uncertainty and stock market volatility”, August 2015.
18. Thomas G. Dietterich. Ensemble methods in Machine Learning, International Workshop on Multiple Classifier Systems, MCS 2000: Multiple Classifier Systems pp 1-15.

19. NumPy, <https://en.wikipedia.org/wiki/NumPy>, Retrieved 5 February 2017.
20. T. G. Dietterich, Ensemble methods in machine learning, in: Proceedings of the First International Workshop on Multiple Classifier Systems, MCS '00, Springer-Verlag, London, UK, 2000.
21. Pew Research Center, How do Americans use Twitter for news? <http://www.pewresearch.org/fact-tank/2015/08/19/how-do-americans-use-twitter-for-news/>. August 19, 2015
22. Pew Research Center, How Americans Encounter, Recall and Act Upon Digital News. <http://www.journalism.org/2017/02/09/how-americans-encounter-recall-and-act-upon-digital-news/>. February 9, 2017
23. Pew Research Center, News Use Across Social Media Platforms 2016. <http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/> . May 26, 2016
24. Courses that use NLTK, <http://tinyurl.com/nltk-courses> Retrieved 2016-06-15.
25. Natural Language Toolkit, https://en.wikipedia.org/wiki/Natural_Language_Toolkit, Retrieved 8 January 2017.
26. Zahan Malkani, Evelyn Gillie. Supervised Multi-Class Classification of Tweets, 14 December, 2012.
27. Statistics for the rest of us. Pearson Correlation: Definition and Easy Steps for Use. <http://www.statisticshowto.com/what-is-the-pearson-correlation-coefficient/>. Retrieved 2017.
28. Nadia Felix, Eduardo R Hruschka, Estevam Rafael Hruschka. Tweet Sentiment Analysis with Classifier Ensembles. Article in Decision Support Systems · July 2014.

29. Gareth James. MAJORITY VOTE CLASSIFIERS: THEORY AND APPLICATIONS. May 1998.
30. Sentiment analysis. https://en.wikipedia.org/wiki/Sentiment_analysis. retrieved at 13 March 2017.