2014-10-16

Portfolio Optimizer Tool (POT) Based on Trend Market Analysis and News Sentiment Analysis

Al-jomai, Riad

UNIVERSITY OF CALGARY

Portfolio Optimizer Tool (POT)
Based on Trend Market Analysis and News Sentiment Analysis

by

Riad Yahya Qaid Al-Jomai

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER IN SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

CALGARY, ALBERTA
October, 2014

Abstract

Stock market and portfolio analysis on financial data conducted by professional investors, who seek informative decision making, is found to be complicated and hard to understand by non-professional investors. In this thesis, we suggest to develop a portfolio optimizer framework which is intended to be beneficial to investors ranging from professionals to non-professionals. The proposed framework integrates data mining techniques such as frequent pattern mining, clustering, classification, and sentiment analysis. To achieve the target, we will apply sentimental analysis to investigate stock market trend in order to predict the actual stock price index movement. This has been realized as an important task for the last decades following the development in technology, though investments in the stock market are active for over a century. According to the accuracy of prediction results, the investors can be guided to investments that might help them make more money. On the other hand, loss of investment is the risk when the prediction reported to investors suffers from lower accuracy or when the prediction is missing some of the factors or objectives that directly or indirectly affect the trend of the market. The integrated framework will help to find the correlation between various stocks in the market to determine groups of stocks predicted to behave similarly, i.e., maintain same trend whether up or down. We will concentrate on the classification of stocks by sector and by industry. This will allow investors to diversify their investments and hence reduce the risk if they prefer. In conclusion, this thesis provides useful guidelines for investors to select appropriate market areas and formulate efficiently diversified investment portfolios. The reported test results demonstrate the applicability and effectiveness of the proposed framework.

Acknowledgements

My first word of thanks goes to Allah the Almighty for His uncountable blessings and for easing the journey of my Master's program.

I would like to thank deeply my supervisor Dr. Reda Alhajj for his invaluable help and support throughout my Master's program. Specifically I want to thank him for suggesting the interesting thesis topic; for his continuous meetings with me to exchange ideas and thoughts, track my progress and keep me on track; for his financial, logistic, and moral support; for his patience with my shortcomings and for mentoring me to quickly catch up with the high caliber of the research group; and for his calm and kind character that inspired me greatly. I want also to thank Dr. Jon Rokne for his feedback and support during my thesis work. The feedback from my examination committee member Dr. Maen Husein is also appreciated. I would like also to thank Alper Aksac for his help and support when I felt the need for more motivation.

Dedication

I dedicate my work to my mother Sadah; my father Yahya; my uncles Derhim, and Ibrahim, my aunt Horia, my brothers and sisters, and my fiance Sarah, for their indescribable love, care, and support.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	**Definition**
DJIA	Dow Jones Industrial Average
IR	Information Retrieval
HC	Hierarchical Clustering
SN	Social Network
DM	Data Mining
KDD	Knowledge Discovery in Databases

Chapter 1

Introduction

For long time the stock market has established itself as one of the attractions for investment despite the several crisis it suffered from almost periodically and the fluctuation in stock prices affected by various factors. Investors are mostly optimistic and ready to take the risk because they expect high return. However, investors range from professionals to new comers who are mostly unexperienced and more subject to mistakes leading to losses. It is generally assumed that professionals have developed over time reasonable prediction skills and hence have the tendency to make more profit than loss unless the market is affected by unexpected and unpredicted factors they are not aware of. The tendency to make profit attracts new comers who may think they will catch the same trend as professionals.

Unfortunately, the stock market is a very volatile and complicated domain that depends on the optimization of various factors. The influence of each factor and the duration of its effect varies over time and based on specific stocks. Thus, researchers and practitioners felt the need to develop some automated techniques that could guide investors and draw their attention to factors which they might have missed. This way, investors will be able to take more or less risk based on their financial state and expectations.

Indeed, the world of the stock market has become attractive to people because of the quick profit making potential. Discovering a suitable set of shares for financial investment to eventually get more return and deal with less risk, in comparison to many other alternatives, draws numerous individuals, whether domain specialists or not. In the stock domain, organizations from different domains (including Energy, Products, Business, Customer Discretionary, Customer Staples, Wellness Care, Financial, Ideas Technology, Telecommunication providers and Utilities, etc.) sell some part of their organizations stocks. People can have a comprehensive aggregate review of all of the available shares to purchase, and therefore, can attain a dependable choice on just how much cash to spend and on which shares. The possible advantage of the stock marketplace is bringing in more and more people to spend. Every investor works like in isolation as if he/she does not have the complete picture about the other investors who are mostly spread over the world. Each investor has his own skills and some investors may depend on specific brokers thinking that will lead to more profit. When to sell and when to buy are critical decisions that once taken right and on the right time will lead to more profit. Unfortunately the stock market is a time sensitive environment and there is almost no time to think because while thinking another investor might take initiative and buy stocks which have the potential to make profit. For investors portfolio optimization is an essential process for more guided investment.

A portfolio is a collection of possessions used by an individual. Portfolio optimization is the method of generating expense choices by keeping a collection of monetary possessions to satisfy different requirements. Simply put, portfolio optimization is the choice procedure of asset weighting and choice, such that the collection of possessions fulfills an investors goals. The stock portfolio optimization procedure requires forecasting the overall performance and also the volatility of shares. This will help designers who may use these forecasts in an effort to acquire a portfolio that suits the investors preference profile.

In economic assets, it is essential for buyers to handle the amount of danger to which they are subject to on their own while looking for portfolio return. As a whole, investment opportunities that provide greater return also require greater risk. Thus, there is constantly a trade-off between risk and return within the financial investment choice procedure. Monetary ideas define the expense threat in a manner that information technology can be used to assess and after that link the quantifiable risk to the level of return which can be anticipated through the expense [22]. The trade-off between risk and return, released in modern portfolio theory, is a vital topic in creating financial investment decisions and therefore ought to be considered by any automatic evaluation.

To cope aided by the portfolio optimization issue, a number of scientists focus on the growth of designs which, in addition to the two fundamental requirements of return and risk, also give consideration to other similarly crucial requirements produced by fundamental evaluation as the investors choices and guidelines [36]. Consequently, the portfolio optimization issue has already been understood as the process of locating the maximum portfolio of possessions according to multiple goals such as return and risk along with other monetary criteria.

The risk/return tradeoff underlies the variation idea, such that portfolio danger/risk can be decreased by incorporating possessions whoever returns are weakly correlated. The concept of diversification is to decrease risk by investing in a variety of stocks selected from various groups which mostly witness opposite trends. In the event that the chosen stocks have no or little similarity in accordance for their developments, a portfolio consisting of the shares will deal with much less risk than the average risk. The latter case may lead to lower return since there is in general a tradeoff between risk and return.

As a whole, an answer for stock portfolio optimization requires building designs from existing readily available information to predicting the volatility and gratification of possessions in the future, in an effort to maintain the maximum portfolio for individual's choices. For many years financial theorists and investors have been working on developing techniques to cope with this concern. In fact, investors have accessibility to a broad variety of historic economic signs about the offered stocks as well as to a wealth of news repositories and social media platforms which may directly or indirectly affect stock prices. Hence the need for automated tools to guide investors as it is almost impossible to adapt a manual process to analyze the daily growing big volumes of available data.

1.1 Problem Definition

Stock markets are a major component of the world economy since they provide a large platform for companies to raise money efficiently and effectively. Although stock markets form one of the most well-documented and data-rich areas for research, yet there is no reliable tool developed which can effectively predict their trends. The vast and unmanageable amounts of data and external factors turn the stock market trends prediction task to be very large and complex. Rumors, local or international news, or announcements can change stock prices unpredictably and radically high or low within minutes. This problem can be as unpredictable as humans fears and reactions. Indeed this is the most natural behavior of humans who tend to react rather than carefully investigate and wisely act. Thus, in order to predict reliably, various sources of information must be taken into account.

Stock market prices are generally very dynamic and susceptible to quick changes because of lots of unpredicted factors such as the underlying nature of the financial domain, and in part because of the mix of known parameters and unknown factors. Choosing best option between stocks for investors in the stock market is regarded as one of the most difficult tasks. Indeed individual investors try to narrow their investment in one of few categories of stocks and try to build their own experience by trial and error or by depending on experienced brokers. Thus, developing automated tools would be of great help to individual investors who may try various options with the hope to maximize the income from their investments. An automated tool may incorporate the view or perspectives of multiple brokers, rank and prioritize them leading to more informative decision making. This latter multiple perspective based analysis is not considered in this thesis; it has been left as future work.

Economic and statistical models, in addition to machine learning and data mining techniques, have been proposed as heuristic based solutions with limited long-range success. Actually, there is a need to find the best correlation between companies by considering their anticipated risk and profit. It is expected that, as a best choice or may be the dream of every investor in the stock market, the risk should be low and the profit for the investors should be high. According to these constraints, the correlated companies and their shares will be suggested by the automated system to users directly. The more comprehensive the coverage of the system is the more realistic and trusted are the suggestions.

The complexity of stock market data yields a problem that is to develop an accurate program which will work for every case and situation to predict the movement of stock prices. It has been well established that this trend analysis for prediction is a multidisciplinary area of research being undertaken jointly by many disciplines, including finance, computer science and statistics.

Predicting the direction of stock market price index movement is regarded as a challenging task of financial time series analysis [14] & [30]. Predicting the movement of stock index is regarded as a challenging task in stock market trend analysis because the stock market data is dynamic, chaotic, nonparametric, complicated, and nonlinear. Also, stock market prices are quickly affected from economic, social and political events, death of celebrities, etc. and reactions for these statements in a negative or positive way can be easily seen in the stock price movements. Researches in this financial area show that this direction of movement can be used in high volatility market as a beneficiary for high returns. According to high accuracy results of prediction, investors may improve and increase their earning from their investments in a short term using daily stock prices. All these investments are indexed to the level of analysis conducted, to the completeness of the data used, and to the comprehensiveness of the factors integrated in the analysis. Such a complex process is hard and even impossible to accomplish manually. The mentioned gap may be filled as described in this thesis by a tool that maximizes the benefit from the available data sources for effective knowledge discovery leading to informative decision making.

1.2 Motivation

As described in the literature, a large number of research efforts have been described in the literature and lots of state-of-the-art approaches have been suggested by various researchers who closely work on investigating and understanding the stock market. Despite all these studies, there is no perfect solution for this topic and still researchers and practitioners continue their effort to find the best optimized solution that will cover every situation. Moreover, according to our research, there is no completed research so far that comprehensively utilizes data mining techniques with the purpose of recommending one of best combination of companies that show similarity in portfolio and risk of investment to non-professional investors. Realizing this need is the main motivation of our proposed approach to study the available portfolios of expert investors and predict the direction of movement in the daily stock market data using text and data mining techniques from different sources, including the actual stock market data, news and social media data. Our target is to recommend to non-professional investors the most appropriate portfolio that fits their expectations and plans.

1.3 Data: Dow Jones Industrial Average Index

The Dow Jones Industrial Average (DJIA) is one of several indices created by Wall Street Journal editor, Dow Jones and Company co-founder Charles Dow. JDIA is an index that shows how 30 large publicly owned companies based in the United States have traded in the stock market. All stocks traded in DJIA and the sectors that they belong to are listed in Appendix A.

1.4 Overview of the Proposed Solution

The block diagram of the proposed approach is shown in Figure 1.1. The goal of our research is twofold. First, we will semantically analyze the news and social media to predict market trend and investigate potential external factors that might be affecting the markets such as financial officers, even some political or military events that may affect the stock market trend. As a result of this analyze, we revealed that various news streams have a very strong correlation and an effect on the stock market price movement.

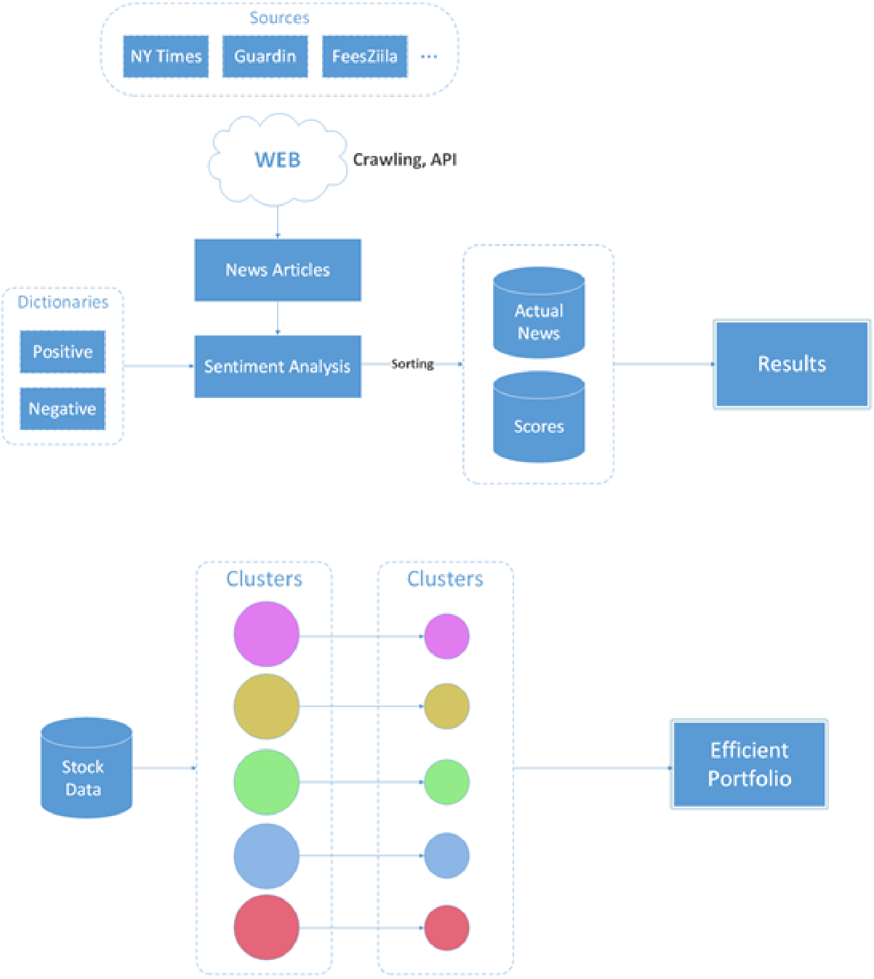

Figure 1.1: Block Diagram of the Proposed Framework

For the sentiment analysis portion of our research, news deemed related to the stock market have been collected from reliable sources such as Yahoo Finance, Yahoo News, Google Finance, Google News, etc. There are lots of reliable news sources which could be considered to extract some information that may have direct or indirect effect on the stock market trend. These news sources are accessible via their APIs. Using text mining techniques, the valuable information about companies or stock market data should be extracted. Then data mining techniques and sentiment analysis should be applied to classify the collected news into two classes: negative and positive.

We argue that understanding the correlation between stocks may give us new insight into modeling the behavior of financial markets. We aim to use data mining techniques such as clustering to finding out interesting patterns and the correlation between various stocks. These are expected to explain the behavior of the stocks and their relationships within one group or across the groups. The discovery will help in building portfolio suggestion system. In this thesis, several clustering algorithms have been applied to the Dow Jones Industrial Average. We decided on using a variety of techniques in order to avoid any bias particularly linked to specific a specific technique.

1.5 Contributions

The implications of our work were mainly envisioned as two components of a general framework which incorporate the following contributions:

First, we designed and developed a general framework that shows the effect of the daily news about companies and to testify and prove their predictability to take advantage of these statements in order to give a chance to the investors to increase their profit. We used open sources of news from reputable organizations such as Google News, Yahoo News, etc. To validate our approach, we will compare our rated scores for daily news with daily stock prices for a specific company from DJIA stock list and check for a significant correlation between two graphs.

Second, we aim to find a good clustering of the communities involving companies in the stock market using their price movement information. We have built a recommendation system that utilizes the discovered information to guide the process of making stock investment decision. Our tool will also perform portfolio optimization for users. That is, it will assist users in selecting a basket of companies to invest in in order to minimize the risk with the hope to produce positive return. For our portfolio we fixed the return equal to the return of equal weight portfolio because we do not want less return; instead we try to reduce the risk by adding more constraints. We observed the following result, the return of portfolio is the same but the risk is reduced by 17

1.6 Thesis Organization

Figure 1.1 shows the flow of content of the four main chapters includes in this thesis. In general, the rest of the thesis is organized as follows: In chapter 2, we provide the general background needed to understand the content of the thesis; then we review the most popular previous studies conducted on portfolio optimization and trend market analysis by incorporating sentiment analysis from different resources such as social media, news, etc. The news sentiment analysis method applied in this thesis is presented in chapter 3. The portfolio optimizer tool is covered in Chapter 4. Finally, conclusions and future work are discussed in Chapter 5.

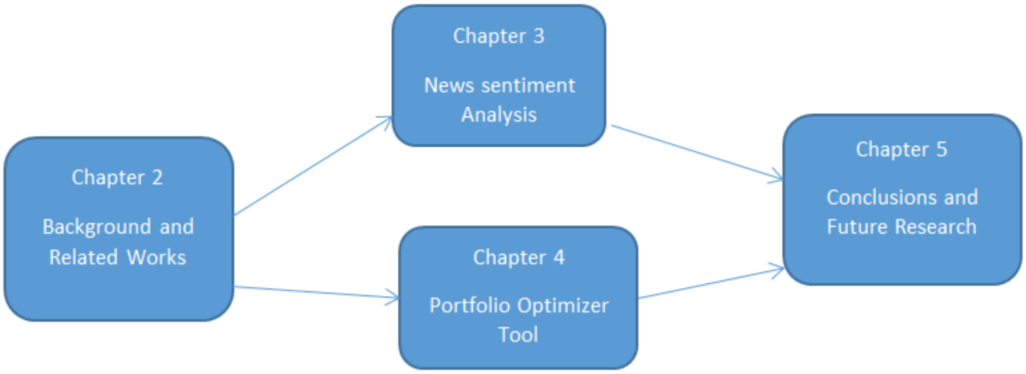

Figure 1.2: Flow of Content in the Thesis

Chapter 2

Background and Related Work

The rapid development in digital information capturing has led towards the fast-growing amount of information stored in databases, data warehouses, or other kinds of electronic data repositories [49]. Although valuable information may be concealing behind the data, the overwhelming data volume makes it hard for individuals to extract the hidden knowledge without non-traditional and powerful techniques. In other words, it has been well realized that traditional information retrieval and query processing techniques are no more effective for maximizing the benefit from the available data repositories. We argue that for long time researchers concentrated on how to store and retrieve the data without realizing the valuable nuggets hidden in the data. Such nuggets are not retrievable by traditional means. Only after the database construction and manipulation literature reached its fancy researchers started to investigate the possible techniques to dig more in the data and retrieve nontrivial information which is not explicitly stored in the repository. Such efforts has allowed for effective knowledge discovery which would easily lead to informative decision making.

In an effort to ease knowledge discovery in growing data repositories, a new discipline named data mining emerged, which devotes itself to extracting knowledge from huge amounts of information, using the assistance of ubiquitous modern computing products, particularly, computer system. Financial time series forecasting happens to be addressed since the 1980s. The objective is to defeat financial markets and win more revenue. Until now, financial forecasting is nonetheless regarded as the most challenging application of modern time series forecasting. Financial time series have really complex behavior, resulting from a huge wide range of aspects which could be economic, governmental, or psychological. They are naturally noisy, non-stationary, and deterministically chaotic [42].

The number of proposed methods in financial time series prediction is tremendously large. These methods rely heavily on using structured and numerical databases. In the field of trading, most analysis tools of the stock market still focus on statistical analysis of past price developments. But one of the areas in stock market prediction comes from textual data, based on the assumption that the course of a stock price can be predicted better by investigating news articles. In the stock market, the share prices can be influenced by many factors, ranging from news releases of companies and local politics to news of superpower economy [27].

Easy and quick availability of news information was not possible until the beginning of the last decade. In this age of information, news is now easily accessible, as content providers and content locators such as online news services have sprouted on the World Wide Web. Nowadays, there is a large amount of information available in the form of text in diverse environments, the analysis of which can provide many benefits in several areas. The continuous availability of more news articles in digital form, the latest developments in Natural Language Processing (NLP) and the availability of faster computers lead to the question how to maximize the benefit from all these nice developments and work out some technique capable of extracting more information out of news articles [46]. It seems that there is a need for extending the focus to mining information from unstructured and semi-structured information sources. Hence, there is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of unstructured digital data. These theories and tools are the subject of the emerging field of knowledge discovery in text databases, known as text mining.

Prior to 1950, investment decisions focused on individual stock returns with little consideration of portfolio risk. In 1952, Markowitz was the first to quantify the link that exists between portfolio risk and return through which he founded the Modern Portfolio Theory [41]. He demonstrated that portfolio risk came from the covariance of assets making up the portfolio. He empirically demonstrated a link between portfolio risk and cumulative asset returns, by measuring portfolio risk relative to the covariance among assets. His theory targeted to maximize the use of the investors terminal wealth [35]. In other words, this theory attempts to mathematically identify the portfolio with the highest return at each level of risk [15], [24] & [1].

Researchers have tried to find the correlation between stocks and the trade off, that is, between risk and portfolio since Markowitz proposed the modern portfolio theory in 1950s [25]. Fundamental analysis is trying to find some good attributes and values from companies return and profiles [37]. In general, a multiple criteria have to be considered for making investment decisions [41]. Although a great effort has been devoted to developing systems for guiding stock investment decisions, limited success has been achieved and reported.

In traditional decision making for investments, investors focus only on maximizing their expected return without considering the concept of investment risk [25]. On the contrary, it is important for investors to control and consider the risk to which they subject themselves while searching for high returns. For most cases, choices from which investors can make lots of money are showing high risk [23]. Because of this, there is always a trade-off between risk and return in the investment decision making process.

2.1 Data Mining

From the beginning of 1980s, data mining techniques are being actively applied to stock market such as predicting stock prices and indices, trend detection, portfolio risk management, etc. Analysis of stock market trend using these kinds of methods will help investors to discover hidden patterns from historical finance data that have apparent predictive potential in their investment opinions [6], [26], [38], [47], [29] & [18]. Data mining is a mathematical technique which is used to obtain the requested or hidden information from unprocessed raw data based on a variety of sources. Prior to the application of a data mining technique, it is not possible to turn the available data into fully useful and meaningful resource. For effective application of data mining techniques, the available data should be processed and put in a specific format. Figure 1.1 shows data mining models and tasks.

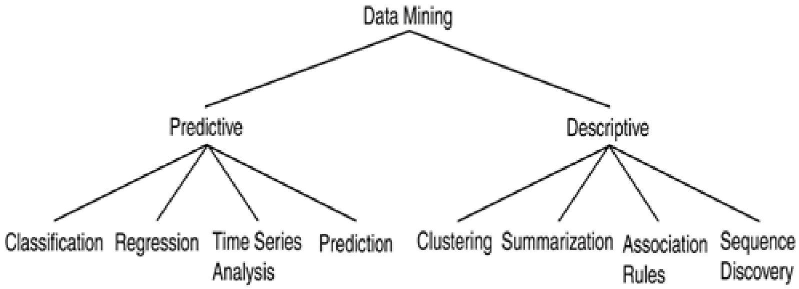

Figure 2.1: Data Mining Models and Tasks

In many real-world domains, an important part of data such as articles, books or web pages is stored as text in electronic format. As a result, structural data and useful phrases or hidden values from these electronically available text data and documents could be obtained using text mining, which is a sub-field of data mining. In addition, summaries, conclusions and inferences can be made from text based on natural language processing and the semantic definition of a pattern between documents.

2.1.1 Knowledge Data Discovery

The term Knowledge Discovery in Databases (KDD) refers to the broad process of finding knowledge in data, and emphasizes the ”high-level” application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. The whole process is summarized in Figure 2.2.

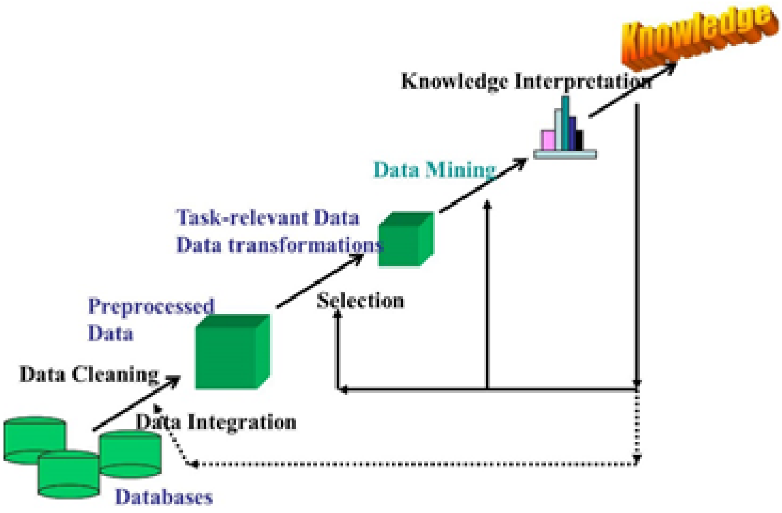

Figure 2.2: Knowledge Discovery in Database Process

2.1.2 Text Mining

Data mining focuses on the computerized exploration of huge amounts of data and in the finding of interesting patterns within them. Until lately computer scientists and information system specialists focused on the breakthrough of knowledge from structured, numerical databases and data warehouses. However, a good amount of details nowadays is available by means of text, including files, news, manuals, e-mail, etc. The increasing number of textual data features led to have knowledge finding in unstructured (textual databases) data understood as text mining or text data mining [46].

Text mining is an emerging technology for examining large collections of unstructured documents for the purposes of extracting interesting and non-trivial behavior or knowledge. Text mining has an objective to try to find habits in all-natural language text and to draw out corresponding information. The work described in [50] regards text mining as an understanding creation tool which provides effective possibilities for creating knowledge and relevance away from the massive quantities of unstructured information available in the Internet and business intranets.

Among the applications of text mining is finding and exploiting the relationship between the document text and an additional supply of details such as time stamped streams of information, namely stock market estimates. Forecasting the movements of stock prices based on the news articles is certainly one of several opportunities of text mining techniques. Information about companies reports or breaking news tales can dramatically affect the share price of a safety. Actually, there have already been numerous research efforts carried out to investigate the influence of news articles on stock market and the reaction of stock market to press releases. Researchers have shown that there is a strong connection between the time when the development stories are introduced and the time when the stock rates fluctuate. This led scientists to investigate a brand-new direction of research by predicting stock trend movement according to the content material of development stories. While there are many promising forecasting techniques to predict stock market movements based on numeric time sets information, only few predicting methods are available for text mining by utilizing news articles. This is true because text mining appears to be more complex than information mining as it requires working with text information that is inherently unstructured and fuzzy.

2.1.3 Clustering

Clustering is the grouping of objects based on the values of a prespecified set of features such that similar objects end up to be in the same cluster. Thus, clustering requires selecting a similarity function that could be applied to determine how close objects to be clustered are based on the values of their features. The similarity functions takes the features of two objects as input and return as output their degree of similarity.

Similarity may be expressed in terms of distance in case the features to be used have numeric values or can be mapped into numeric values. However, normalization of numeric features is recommended as a preprocessing step to avoid the dominance of any feature(s) with large values. While objects located within the same cluster show similar characteristics, objects in different clusters are expected to be less similar.

Clustering is an unsupervised learning method and a technique widely used for data analysis. It has roots in statistics. However, advanced machine learning based clustering techniques are more powerful than clustering techniques that depend merely on statistical concepts. This is true because statistics mainly work on small number of samples while machine learning techniques are expected to scale for large data. Furthermore, clustering has been successfully used in many fields such as pattern recognition, image analysis, information extraction, and bioinformatics, among others. In clustering, different sets are created for different parameters. Various applications of clustering can be found in many disciplines such as medicine, finance, marketing, astronomy, web, geographic information systems, computer local networks, health. Finally clustering has a potentially important contribution to stock market portfolio analysis.

Clustering actually refers to the analysis a data set without consulting a known class label. In other words, class labels are not present in the training data, as they are not known to begin with. Clustering is used to divide a data set into classes (by generating labels for them) using the principle of maximizing the intra class similarity and minimizing inter class similarity. Thus, clustering also facilitates taxonomy formation, i.e., organization of the observed objects into a hierarchy of classes that group similar things together.

There are various categories of clustering algorithms; each has its advantages and disadvantages. But most clustering algorithms require the users to specify the target number of clusters directly or indirectly. By saying indirectly, we mean the user is expected to specify certain parameters that lead to a specific value for the number of clusters. Our research group developed a multi-objective genetic algorithm based clustering approach which has the advantage of finding the number of clusters and the most compact clusters in an automated way.

2.1.3.1 Similarity Function

A similarity function takes as input two values and returns as output a value indicating their level of similarity ranging from very similar or totally overlapping (being the same object) or totally dissimilar. Distance measures form one main category of similarity functions. Based on a distance measure, two object are said to be similar when the distance separating them is close to zero and the two objects tend to be dissimilar when the distance measure between them approaches one based on a scale normalized in the interval [0,1].

Two types of distance measures are mainly used in clustering, namely Euclidean distance and non-Euclidean distance. While the non-Euclidean distance is mainly based on the characteristics of objects (like color, citizenship, etc.), the Euclidean distance is computed based on the spatial characteristics of objects (all the characteristics considered have numeric values representing the location of the object in the space). One of the most commonly used Euclidean distance measure is the Euclidean distance which is computed between two objects as the square-root of the sum of the squares of the differences between the values of corresponding characteristics in the two objects. It is expressed as follows:

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) of each object, p_k and q_k are, respectively, the k_{th} attributes (components) in data objects p and q.

2.1.3.2 K-Means Clustering

K-means is a partitional clustering algorithm in the sense that it partitions the data objects into a predefined number of clusters. Every object is forced in a cluster, i.e., k-means is not capable of identifying outliers in the data and hence it finds clusters of arbitrary shapes. It addition to not considering outliers, k-means expects the user to specify the number of clusters which may be hard to determine especially in a real environment where clustering is expected to be totally unsupervised and users may not know details of the given data.

Further, k-means suffers from local minima and the clustering result is very sensitive, i.e., depends highly on the choice of the initial seeds. Accordingly, there has been some research related to the proper choice of the initial seeds that might lead to better clustering.

In our approach to determine the number of clusters k, we are searching all the values in the interval [2,100]. In other words, we try different values of k, looking at the change in the average distance to centroid as k increases. The average falls rapidly until an appropriate value of k is reached, then it changes very smoothly, like parallel to the x-axis, which means all corresponding values of k will not introduce any further improvement. Steps of the k-means algorithm are given in the next section.

K-Means Algorithm

1. Define a similarity measure between any two points in the space that contains the objects to be clustered.
2. Choose k points as initial seeds or group centroids.
3. Apply the similarity measure between each object and every existing centroid.
4. Assign each object to the group/cluster which has the closest centroid.
5. When all objects have been assigned, recalculate the positions of the k centroids.
6. Repeat Steps 3, 4 and 5 until the centroids no longer move, that is every object settles in its hosting cluster. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Guessing Appropriate Value for Number of Clusters k Determining the appropriate value for the number of cluster k is a challenge in k-means. One possible approach is to try different values of k, and to look at the change in the average distance to centroid as the value of k increases as shown in Figure 2.3. The average falls rapidly until a reasonable value of k is found, then the change in the average distance to centroid becomes almost negligible.

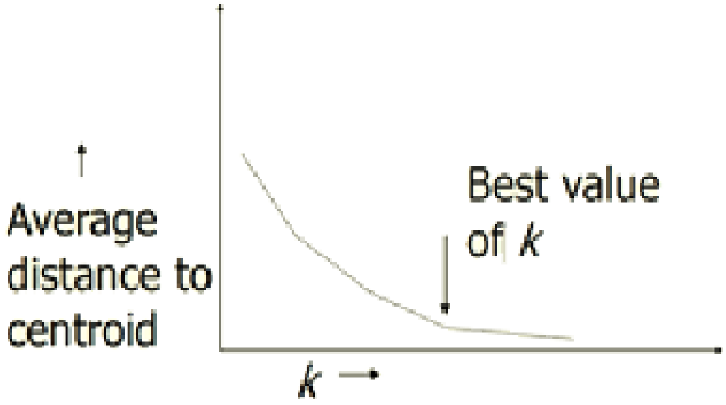

Figure 2.3: Deciding on the Number of Clusters

2.1.3.3 Hierarchical Agglomerative Clustering

In hierarchical clustering the data objects are not partitioned into a particular clustering solution in a single application of an algorithm as it is the case with k-means for instance. Instead, the hierarchical clustering process may work from whole to parts or from parts to whole, i.e., either top down or bottom up. In the top-down approach, the process starts from a single cluster containing all the given N objects then clusters at each level are recursively partitioned until we have N clusters each contains a single object. In other words, hierarchical clustering is subdivided into agglomerative methods, which proceed by series of fusions of the N objects into groups, and divisive methods, which separate the N objects successively into finer groupings. During the clustering process, we need to compute a correlation matrix, and for this case we use Euclidean distance. The details steps of the algorithm are given next.

Hierarchical Algorithm

1. Start by assigning each item to a cluster. Assume the distances (similarities) between the clusters are the same as the distances (similarities) between the items they contain;

2. Find the closest (most similar) pair of clusters and merge them into a single cluster;

3. Compute distances (similarities) between the current clusters, including the clusters produced from Step 2;

4. Repeat step 2 and 3 until all items are clustered into a single cluster of size N.

Step 3 in this algorithm requires computing the distance (or similarity) between clusters. This is known in the literature as separateness which could be determined in several ways based on whether the user is interested in considering only the centroids as representative of clusters or all the objects in each cluster are to be used in the computation; a mixture of centroids and all objects is another option.

Separateness Separateness measures how far two clusters are. It is mostly used to measure the dissimilarity between clusters. Hierarchical clustering benefits from the separateness measure to find the least dissimilar two clusters as candidates for merging them in the next step of the recursive process. There are various methods to measure the separateness of two clusters.

- Single linkage clustering (nearest neighbors technique) here the distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered (think of this like a bipartite graph between the two clusters). The distance between two clusters is given by the value of the shortest link between their objects. At each stage of the hierarchical clustering the two clusters for which the distance is minimum are merged.

- Complete linkage clustering (farthest neighbors) is the opposite of the single linkage, i.e., distance between groups is defined as the distance between the most distant pair of objects, one from each group. At each stage of the hierarchical clustering process the two clusters for which the complete linkage distance is minimum are merged.

- Average linkage clustering the distance between two clusters is defined as the average of the distances between all pairs of objects, where each pair is made up of one object from each group. At each stage of the hierarchical clustering process the two clusters for which the average linkage distance is minimum are merged.

- Average group linkage clustering with this method, once formed, groups are represented by their mean values for each variable, that is, their mean vector and inter-group distance are defined in terms of the distance between two such mean vectors. At each stage, the two clusters for which the distance is minimum are merged. In this case, the two clusters are merged such that the newly formed cluster, on average, will have minimum pairwise distances between the points in it.

- Wardś hierarchical clustering - Ward (1963) proposed a clustering procedure seeking to form the partitions P_1,..., P_n in a manner that minimizes the loss associated with each grouping and to quantify that loss in a form that is readily interpretable. At each step, the union of every possible cluster pair is considered and the two clusters whose fusion results in minimum increase in "information loss" are combined. Information loss is defined by Ward in terms of an error sum-of-squares criterion [16].

In the complete linkage method, the distance between two groups is set as the most distant pair of individuals in the groups, for example the distance between cluster I and II in Figure 2.4 is the maximum distance of the pairs of $d_1$3; $d_1$4; $d_1$5; $d_2$3; $d_2$4 and $d_2$5. So the distance between cluster I and II is $d_1$4, which is the most distant among all. Figure 2.5 shows a Simple example of dendgram tree.

2.1.4 Portfolio Optimization

Portfolio optimization is the process of choosing the proportions of various assets to be held in a portfolio, in such a way that the selected portfolio is better than any other portfolio

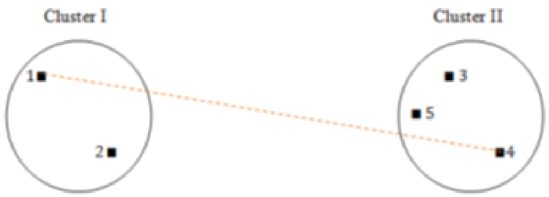

Figure 2.4: Distance between Two Clusters

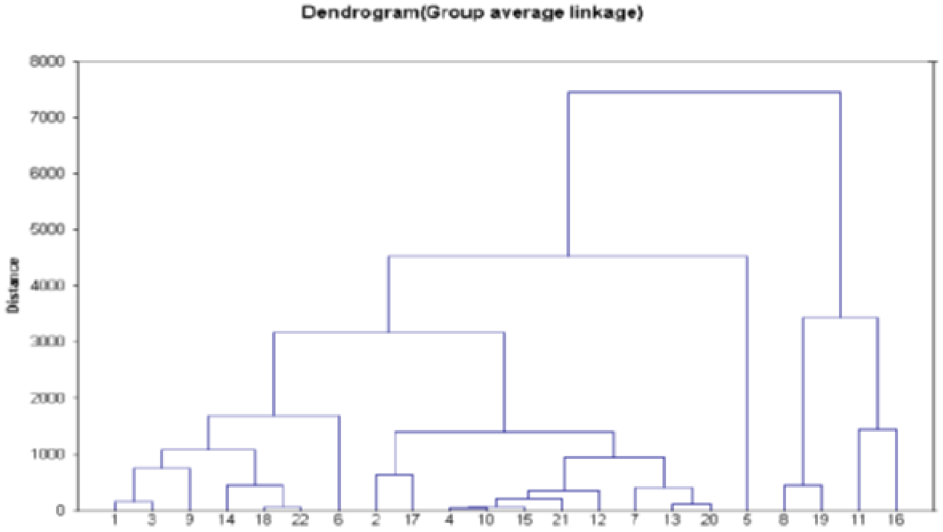

Figure 2.5: A Simple Example of Dendgram Tree
Adopted from [http://www.resample.com]

according to some criteria. The criteria will combine, directly or indirectly, considerations of the expected value of the portfolio's rate of return as well as of the returnś dispersion and possibly other measures of financial risk.

2.1.5 Efficient Portfolios

Modern portfolio theory, attributed to and fathered by Harry Markowitz in the 1950s, assumes that an investor wants to maximize a portfolio's expected return contingent on any given amount of risk, with risk measured by the standard deviation of the portfolio's rate of return. For portfolios that meet this criterion, known as efficient portfolios, achieving a higher expected return requires taking on more risk, so investors are faced with a trade-off between risk and expected return. This risk-expected return relationship of efficient portfolios is graphically represented by a curve known as the efficient frontier. All efficient portfolios, each represented by a point on the efficient frontier, are well-diversified.

2.1.6 Markowitz Portfolio Theory

It is not optimal for the investor to hold a single stock, the opposite of this, i.e., holding a group of stocks is more valuable. The portfolio is a group of assets that is generated based on the investors choices and interests. These choices form the risk range and the investors anticipated actual profit at the end. The most well-known portfolio theory is Markowitz portfolio theory. The main idea of this theory is to define the risk as standard deviation and based on that Markowitz defined an optimal portfolio that maximizes the return for specified risk or minimizes the risk for a given level of return. There are different optimization problems related to Markowitz theory, but here we focus on the two most important problems as follows:

The minimum risk portfolio;

$$Min\ w^T \sum w$$

Subject to:

$$w^T \mu = \overline{r}$$

$$w^T 1 = 1$$

With the above two optimization problems, we can define a set of efficient portfolios that offer the maximum return for a different defined level of risk or the lowest risk for a different defined level of return. These form a set of optimal portfolios that we can draw in a risk-return space called efficient frontier.

2.1.7 Related Work

In the work described in [3], the authors have two stages, in the first stage, they categorize the stock data based on zero growth, slow growth and fast growth using the k-means algorithm. In the second stage, they use the CIR algorithm to generate useful trends about the behavior of the stock market.

Currently, soft computing techniques are widely accepted in studying investment management and in evaluating market behavior [4]. Various techniques have been proposed for this purpose such as Neural Networks (NN), Genetic Algorithms (GA) and Support Vector Machines (SVM). Several researchers presented encouraging results on stock selection using these data mining techniques, etc., [35], [5], [28], [19], [34], [8], [10] &[17].

Also, in the last years, Social Network Analysis (SNA) started to be applied to find hidden relationships between companies and, they are suggesting the best combination of companies relies on similar correlation to investors in order to increase their investment and reduce their risk in the market compared with their encouraging results with Markowitz Model [21] & [20]. As in the aforementioned section, there is lots of work done for trend market analysis with semantic analysis and research is still going on under this topic. This is indeed a very promising research area.

Theussl et al. [44] presented a framework for sentiment analysis using text and data mining techniques. They collected their dataset from the New York Times for the sentiment analysis. Later, polarity calculation for annotated terms and sentiment score calculations have been made using some techniques detailed in their paper. In the background, kind of Map Reduce algorithm is used for text mining and analysis. According to these extracted words, they calculate positive and negative values of each text content and a score for each daily documents sentiment. They showed their sentiment relationship with daily stock prices.

In another state-of-the-art work, Godbole et al. [11] present a scheme for extracting sentiments from daily news articles and blogs. Also using daily news and prices, they made a scoring and provided a good theoretical background for their work. They showed and proved their algorithms results. But their project is good for taking an idea because they are not using a real-time data for their analysis. Using static data for this type of analysis is not enough to show accuracy of the system. Also, exact results of their algorithm and how it works are not clearly mentioned in their methodology and experiments section.

Goonatilake et al. [13] present a work which shows a relationship between the stock market and oil prices. They made a connection between news articles and stock market indices from Dow Jones Industrial Index (DIJA), S&P 500, and NASDAQ. They classify their extracted news data into four dimensional categories. Using a regression model, they try to measure their sentiment score on the daily stock data for detecting the movement of the stock price index.

Yu et al. [48] present a framework which is based on text mining to find the sentiment score of the news articles. They revealed their results for energy companies and collected articles about them. They proved a relationship between articles and their stock market prices.

Noticed onrush of Global Stock marketplace provides these new options for the effective expense. Conversely, the pushing problemthe effect of the marketplace growthis the selection of suitable shares or inventory choice. The monetary overall performance for the inventory returns of the majority of businesses is today provided in online software packages dedicated for stock evaluation. The process is intended to be simpler and faster to modify a portfolio to suit the desired design and the tastes associated with the buyer. The individual can modify a design to feature up to 2040 various signs at the same time and that can scan on general values, e.g., the shares aided by the greatest relative power, or on fundamental factors, like earning per share (EPS) energy. In a number of situations, such methods supply great outcomes according to the solitary deciding criterion [12], but as a whole, the inventory choice issue appears as having numerous requirements. In rehearse, a buyer deals with a collection of regional requirements based clearly or implicitly on the various monetary ratios and inventory return. Two different sources of information can be used for stock selection: namely financial ratios and stock prices. The final results of selection can be obtained with the use of teaching of the decision making system based on the comparison of firms financial performance with its success in the Stock Exchange. The teaching of the decision making system can be carried out with the use of optimization methods [41].

A big wide range of businesses make use of news analysis to aid them make much better company choices [43]. News analysis is the dimension of these different qualitative and quantitative characteristics of textual (unstructured information) development tales. Some of these characteristics are: belief, relevance, and novelty. Revealing development tales as figures and metadata allows the manipulation of daily details in a mathematical and analytical method. News analytics are found in monetary modeling, especially in quantitative and algorithmic trading [33]. Moreover, development analytics may be used to plot and define fast behaviors in the long run, and therefore produce crucial strategic ideas about opponent companies [2] & [45]. News analytics are generally derived through automatic text evaluation and they are used for electronic texts by making use of components from all-natural language control and machine mastering such as latent semantic analysis, support vector machines, ”bag of words” among various other methods.

Stock market forecasters focus on developing techniques to effectively predict list values or stock costs, intending at large earnings by making use of well identified trading methods. The main concept to effective stock market forecast is attaining greatest results by utilizing minimal necessary input data as well as the minimum complex stock market design.

Unquestionably, forecasting stock return is hard because of market volatility that should be captured in the utilized and implemented designs. Correct modeling requires, among various other elements, consideration of the phenomena defined, for example, by economic downturn or development times, and large- or low-volatility durations. Noticed volatility in stock market returns/prices occurs through the reality that desirable (required) rates of return are on their own extremely volatile, driven by cyclical along with other short-term variations in aggregate need. Current improvements in software computing methods provide helpful resources in forecasting loud surroundings like stock areas, by acquiring their particular nonlinear behavior. Choices in monetary areas frequently include good amounts of cash and thus place a considerable quantity of an investors economic money at threat. The entire process of generating transactional choices within the stock marketplace is frequently complex provided the quantity of details offered to every single specific buyer. To complicate things more, the rate and price of data arrival features are intensified because of the too technical innovations such as web based trading and development reports. In general information happens to be delivered significantly quicker to people as in comparison to just what ended up being years ago.

Behavioral research reports have found that buyers are susceptible to an affect heuristic; particularly when overwhelmed by details and it is necessary to make choices that are naturally high-risk. Affect in an emotional condition relates towards a state of feeling or mood, which may be either good or unfavorable. The work describes in [9] files that people usually have a double procedure of ideas; wherein thoughts perform a crucial role. Humans have a tendency to utilize affective thinking instead of analytical thinking whenever emotionally shocked.

Intellectual wisdom and the choice generating after that falters. Indeed, individual behavior usually has a tendency to irrationality whenever experienced with an extremely unfavorable mood or feeling and this can be impacted by the development and market environment.

If how feelings perform plays a component into the economic market, after that which emotions can be mentioned due to prominent persons who impact people? Recall that an investors primary function is to attain revenue. Expense revenue is made through creating decisions which maximize any gain, while minimizing any loss made. Through this, information technology can be suggested that buyers are afraid of huge loss and also at exactly the same time money grubbing for gains based on their portfolio. Concern and greed are extremely all-natural personal characteristics, and it will be postulated that they are the two primary psychological says that are embedded in the stock market at any provided time.

This tendency towards irrationality by buyers features some implications. Irrational trading behavior indicates that there will be often over response or under reaction in security prices. This deteriorates the pricing high quality or signals as sentiment fluctuates. Investor confidence features have already been examined extensively during the past decade [32]. The work described in [7] builds theoretical designs of where buyers are biased to their very own private signals, and underweight the public signal. These models of investor overall confidence point towards decreasing cost high quality, enhanced volatility, reduced trading earnings and therefore lower utility as investor overall confidence increases. In summary, we have actually over responses in prices, which later on reverses. Also, it provides another model where investor confidence is powerful; exclusive signals that are later on confirmed by public signals are enhanced, and vice versa, this leads to variants in buyer confidence. For example, if an investor's buy signal is later on confirmed by a public signal, his/her confidence is most likely to improve.

Chapter 3

News Sentiment Analysis

The rapid advancement in technology increased the availability of and facilitated easier accessibility to huge news repositories. Such news repositories contain valuable nuggets that require sophisticated techniques to analyze them for effective knowledge discovery leading to informative decision making. News articles cover a wide range of topics and could easily benefit a variety of application domains. However, in this thesis we concentrate on the analysis of news related directly or indirectly to stock market. Our target of applying sentimental analysis is to discover how news influence the trend in the stock market and which stocks are affected by what categories of news articles. Sometimes, the effect of the news may be realized directly and immediately in the stock market. However, some news articles may help in predicting the near or far future trend and hence allows for quick selling or purchase to make more profit. It is a matter of how quick an investor is able to act; sometimes it is almost nano seconds. Hence, the need for an automated tool that could guide the investors in the right direction and on the right time.

In this chapter we describe our methodology for sentimental analysis in support of stock market trend prediction. We report the results of our testing to demonstrate how the whole process is applicable and effective.

3.1 The Proposed Methodology

In this section, we will describe our approach to determine the stock price movement using sentiment score analysis. We will also present the details of the requirements for this analysis, including framework design, data gathering, data preprocessing and cleaning using text mining, sentiment analysis using data mining techniques as well as finding appropriate score, and making a connection with the stock market prices.

3.2 The Framework

In Figure 3.1, we have reflected our scheme for the system that has been developed by using the rich R library as the mathematical/statistical component and then interfacing with C# for the front end application design. In other words, we developed our approach using R and later C# for component design; we shifted and established a connection between them.

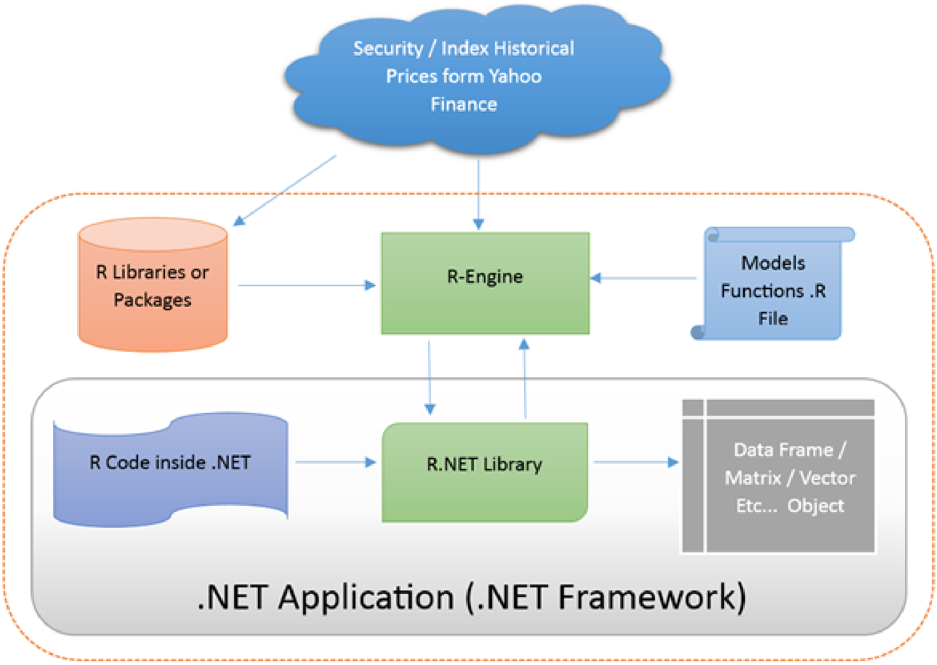
Security / Index Historical Prices form Yahoo Finance
R Libraries or Packages
R-Engine
Models Functions .R File
R Code inside .NET
R.NET Library
Data Frame / Matrix / Vector Etc... Object
.NET Application (.NET Framework)

Figure 3.1: Framework Design

3.2.1 The Data Utilized in the Process

The first step in our method is to create and collect news articles from various trusted news sources such as Yahoo News, Yahoo Finance, Google News and Google Finance. We concentrated on the data between April 1, 2013 and April 28, 2014; we combined together all the crawled data. In Table 3.1, we list the crawled sources and the number of news components extracted from each source. We did our tests for one company, namely Microsoft, in order to demonstrate the effectiveness and applicability of our approach. The same process may be applied to analyze the news related to any other available stock. Of course the process is the same but the outcome would be different based on the characteristics of the specific stock to be investigated. Also, we extracted related to Microsoft financial data from Yahoo for the same period.

News Source	**Quantity**
Google News	216
Google Finance	42
Yahoo News	189
Yahoo Finance	37
Total	**484**

Table 3.1: Collected Data from News Resources

3.2.2 Preprocessing

Our collected data is already structured but it is not ready enough for further analysis. Thus some preprocessing may be helpful before we proceed. The following steps are taken to preprocess the news articles before they are ready for actual knowledge discovery by our method.

- All letters are converted to lowercase because the analysis is case sensitive.
- URLs are removed.
- Frequently used meaningless words are cleaned, e.g. stop words such as the, in, a, vb.
- Numbers are removed from the text because they do not have useful meaning for the processing within the context of our study.
- Removing punctuations.

3.3 Sentiment Analysis

Our algorithm starts by extracting data from real time news sources about a specific company. Also, the algorithm takes the financial data from Yahoo finance for the same company.

After collecting all news pieces related to the company to be investigated, we polarize key words of the sentences. In the text given below, there are two news headlines and their polarization is highlighted to demonstrate the negative and positive polarization cases. Notice the keywords colored in red or green as indicators of the negative or positive polarization, respectively.

- ”Google’s **threat** To Microsoft, Chromebooks Are Now 21% of Notebooks And 10 ... - Forbes” ⇒ Negative Explanation
- ”Microsoft **releases** three classic Windows time-wasters for Windows Phone 8 - Polygon” ⇒ Positive Explanation

For the analysis, we have used Natural Language Processing (NLP) tools for sentence analysis. After that, we have worked on cleaning garbage words from the given sentences. Also, taking benefit from the text, we are using the headlines and the description for news not just only the main text section. We compared each of these main texts, headlines and description analysis in the experimental analysis section.

3.3.1 Sentiment Words

For the sentiment analysis, we used a dictionary that has already been used for the previous state-of-art-works. This dictionary contains positive and negative words lists. According to these files, we are comparing our words and counting how many positive and negative words are contained in one sentence. Later, we conduct our analysis with the whole document score. This calculation is explained in the next section in details. Some details of the used dictionary are given in Table 3.2 [17].

Dictionary Type	**Number of words**
Positive List	19093
Negative List	44759

Table 3.2: Dictionary Details

3.3.2 Sentiment Scoring

In the previous section, we mentioned that we have used the dictionary for our analysis. And we calculated the scores for the sentences based on the following rules.

Rule 1: An instance is classified as positive if the count of positive words is greater than or equal to the count of the negative words. Similarly, an instance is classified as negative if the count of negative words is greater than the count of the positive words.

$$s_A = \sum_{i}^{k} sing(n_p - n_n)$$

According to this formula, S_A represents a score for a news article, n_p is the number of positive words, n_n is the number of negative words, k denotes sentence number. In others words, the score is computed as the sum over k sentences.

Rule 2: There is a need for normalization between 0 and 1 to represent a graph. Assume that there are n rows in the data with two variables A and B. We use variable A as an example in the calculations below. The remaining variables in the rows are normalized in the same way.

The normalized value of a_i for variable A in the i^th row is calculated as:

$$Normalized(a_i) = \frac{a_i - A_{min}}{A_{max} - A_{min}}$$

Where A_{min} is the minimum value for variable A, A_{max} is the maximum value for variable A.

This function helps us to analyze some text and classify it in different types of emotion: anger, disgust, fear, joy, sadness, and surprise. The classification has been performed using two algorithms: the first is a naive Bayes classier trained using Carlo Strapparava and

Alessandro Valitutti emotions lexicon [40]; the second is just a simple voter procedure. This task is intended as an exploration of the connection between lexical semantics and emotions. All words can potentially convey affective meaning. Every word, even those that are apparently neutral, can evoke pleasant or painful experiences due to their semantic relationship with emotional concepts or categories. While some words have emotional meaning with respect to an individual story, for many others the affective power is part of the collective imagination (e.g., words such as "mum", "ghost", "war"). These latter groups of words are particularly interesting because their affective meaning is part of common sense knowledge and can be detected in the linguistic usage. For this reason, we believe it is important to study the use of words in textual productions, and possibly their co-occurrence with words in which the affective meaning is explicit.

Several previous studies in linguistics and psychology have considered research issues related to the affective lexicon. For example Ortony et al. [31] distinguish between words directly referring to emotional states (e.g., "fear", "cheerful") and those having only an indirect reference that depends on the context (e.g., words that indicate possible emotional causes such as "killer" or emotional responses such as "cry").

To explore the connection between emotions and lexical semantics we propose to focus on the emotion classification of news headlines extracted from news web sites. The news headlines typically consist of a few words and are often written by creative people with the intention to "provoke" emotions, and consequently to attract the readers' attention. These characteristics make the news headlines particularly suitable for use in an automatic emotion recognition setting, because the affective/emotional features (if present) are guaranteed to appear in these short sentences.

In contrast to the classification of emotions, polarity classification allows us to classify some text as positive or negative. In this case, the classification can be done by using a naive Bayes algorithm trained on Janyce Wiebes subjectivity lexicon [39]; or by a simple voter algorithm [10].

3.4 Experiments

In this section, we report the results of some testing to demonstrate the applicability and effectiveness of the proposed approach. This is indeed a trend analysis tool and we conducted a short time series analysis because of a short time analysis, there is no need to train the system to predict the next movement to determine what is anticipated to happen. It is possible to use the system with a real time data to do this kind of analysis. In the previous section, we mentioned how we gathered our date and time dilation.

Figure 3.2 represents the histogram of the news resource. It is obvious that for the four different resources positive, negative and neutral scores are close to each other, respectively. In other words, for the four resources each of the three scores is almost indistinguishable. This is so good for the analysis and we can trust our analysis results; in fact it is not a chance and coincidence. Correlation is represented between these news resources.

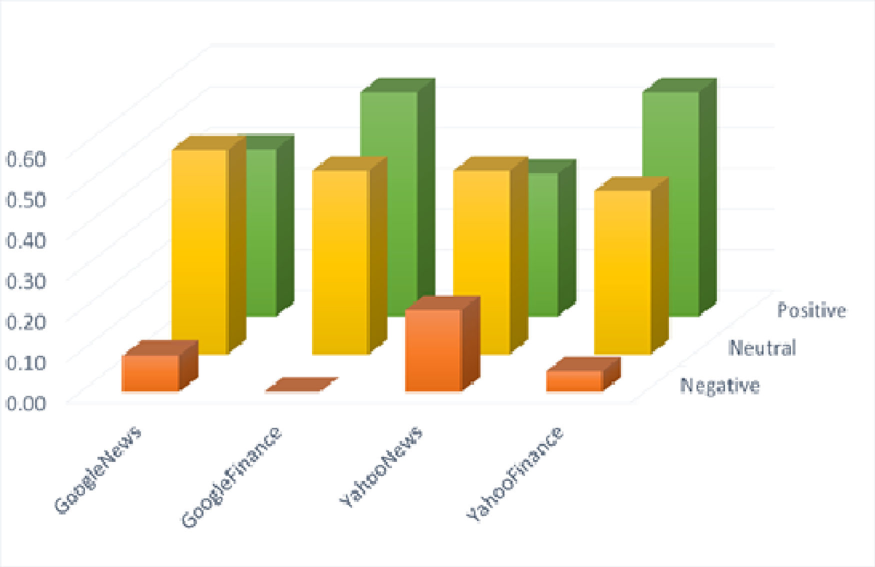

Figure 3.2: Histogram of News Resources

Figure 3.3 shows the correlation in daily finance and news data. Actually, we investigated the data in three sections: (1) by considering directly the text content of the news; (2) by considering headlines; and (3) by concentrating on the description of the news. The results related to general text contents, headlines and descriptions are shown in Figure 3.3, Figure 3.4 and Figure 3.5, respectively; they are compared with daily finance data. Black lines point to the sentiments scores for news and red lines represent the daily returns of stock data. It can be clearly realized that there is a correlation between stock market prices movements and news sentiment scores. According to todays data, we can say that tomorrow or one day later prices may go up again. Daily returns as shown today and the day before reflect prices movements down. In addition, these results are affected from bad news about Microsoft Company. In Figure 3.5, we combined all these three types of input and took the average of them.

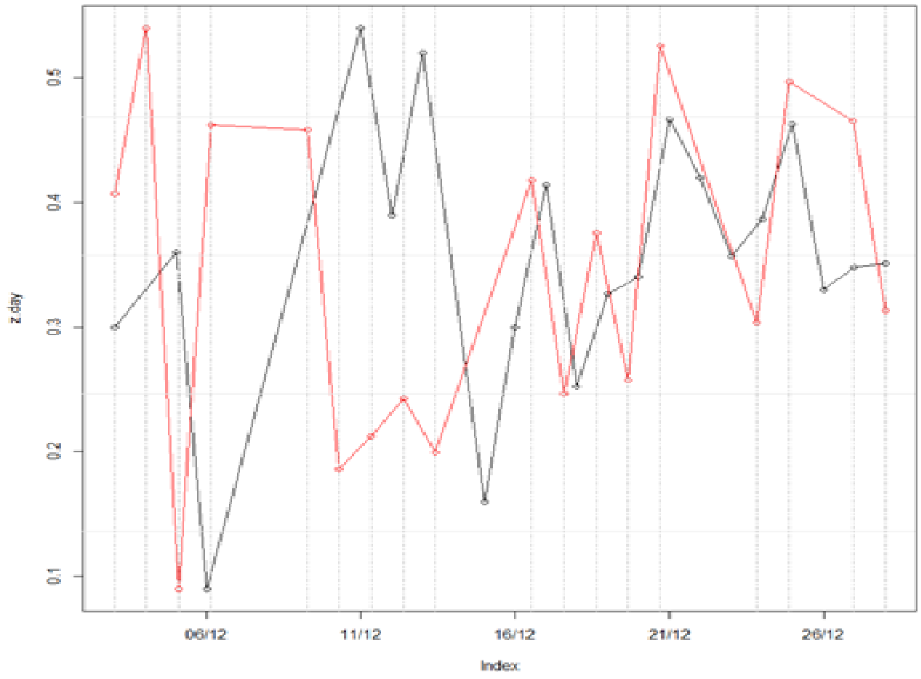

Figure 3.3: Showing Correlation between Main Content and Daily Return

In Figure 3.7, we represent an emotional analysis for text contexts. Emotions are classified

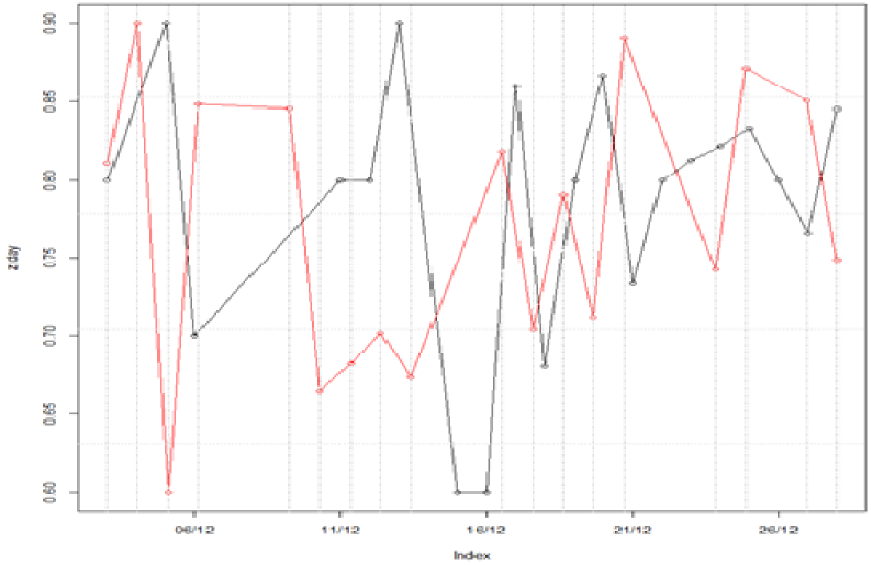

Figure 3.4: Showing Correlation between Headlines and Daily Return

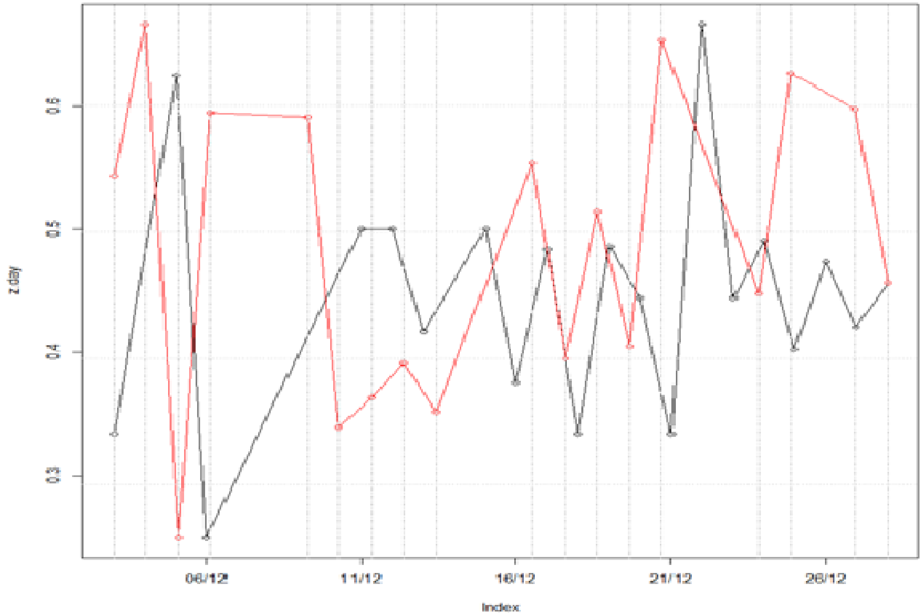

Figure 3.5: Showing Correlation between Description and Daily Return

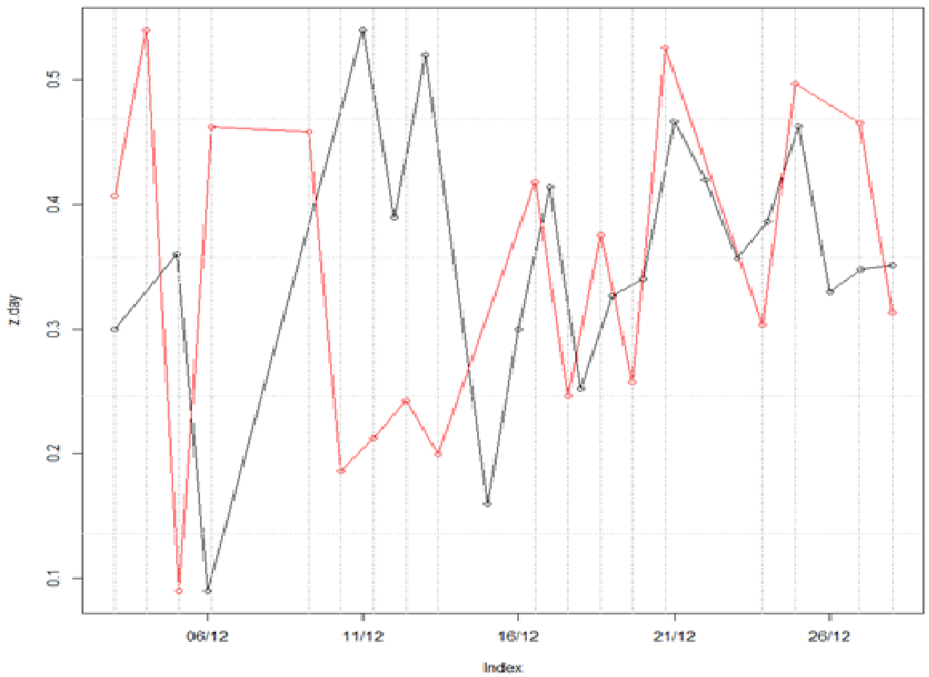

Figure 3.6: Showing Correlation between All and Daily Return

in seven clusters. For the testing, we used a Bayes classifier that is trained from a dictionary and used in some previous works. The case reflected in Figure **??** outlines our distribution for positive, negative and neutral words. In Figure 3.8, we again show the daily number of published online news counts.

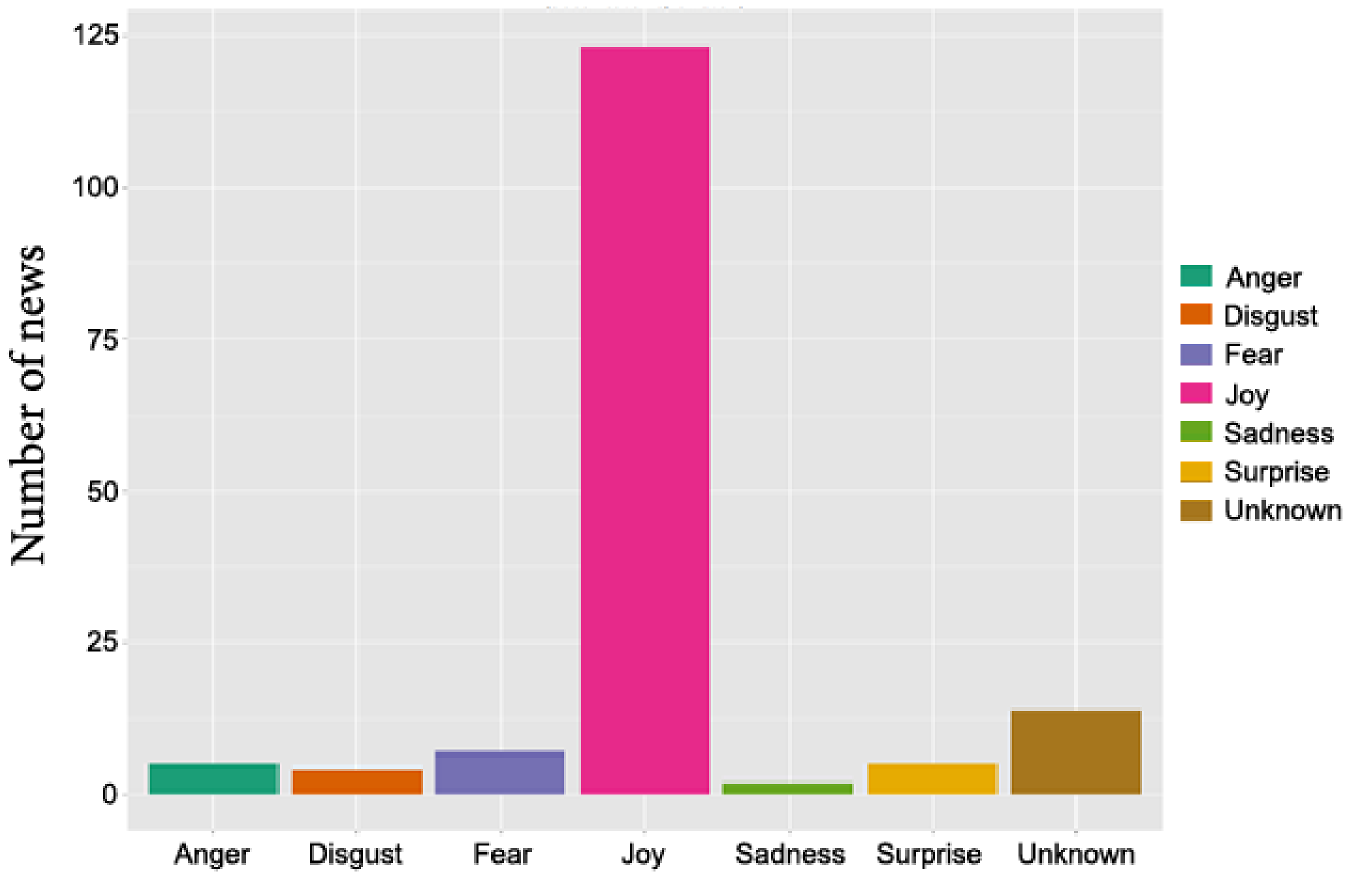

Figure 3.7: Emotional Analysis for Text Contexts

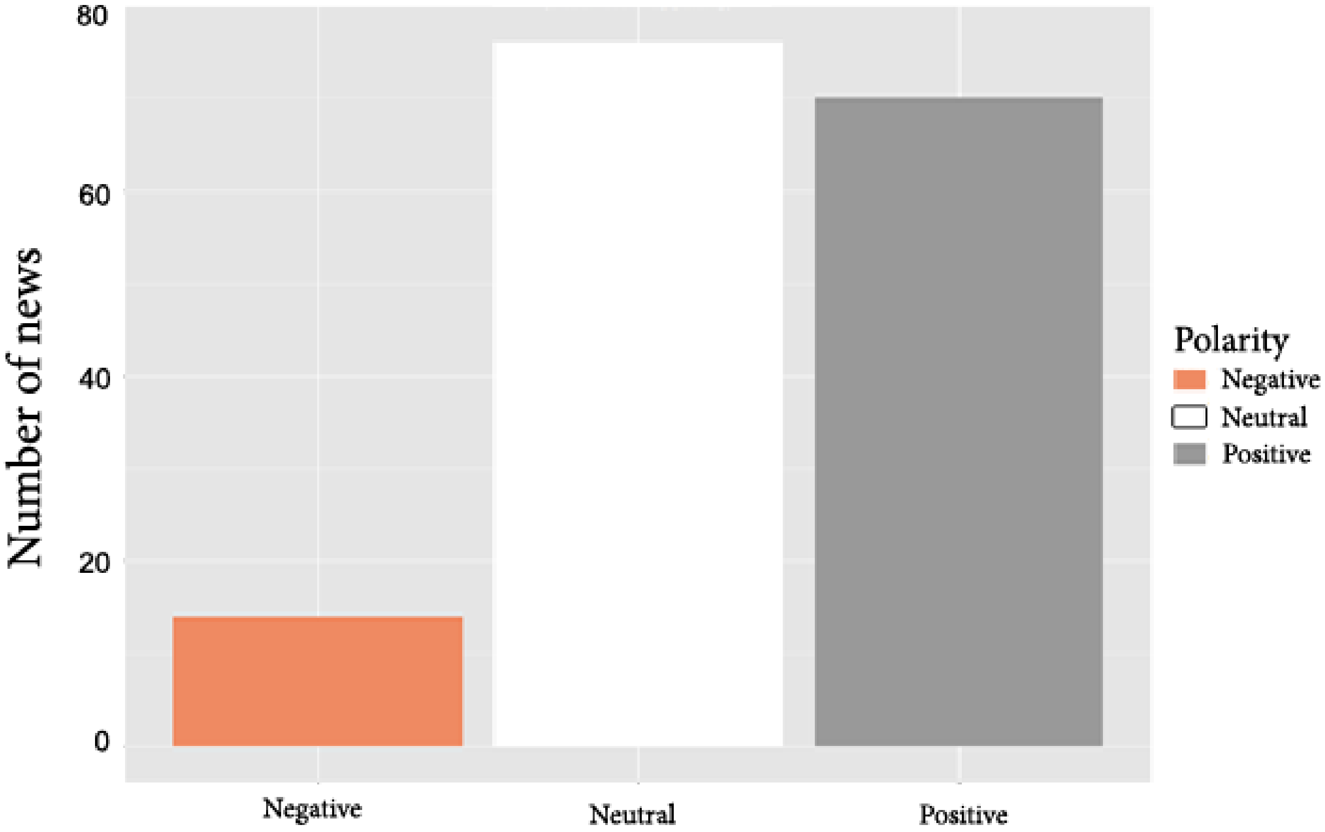

Figure 3.8: Daily Number of Published Online News Counts

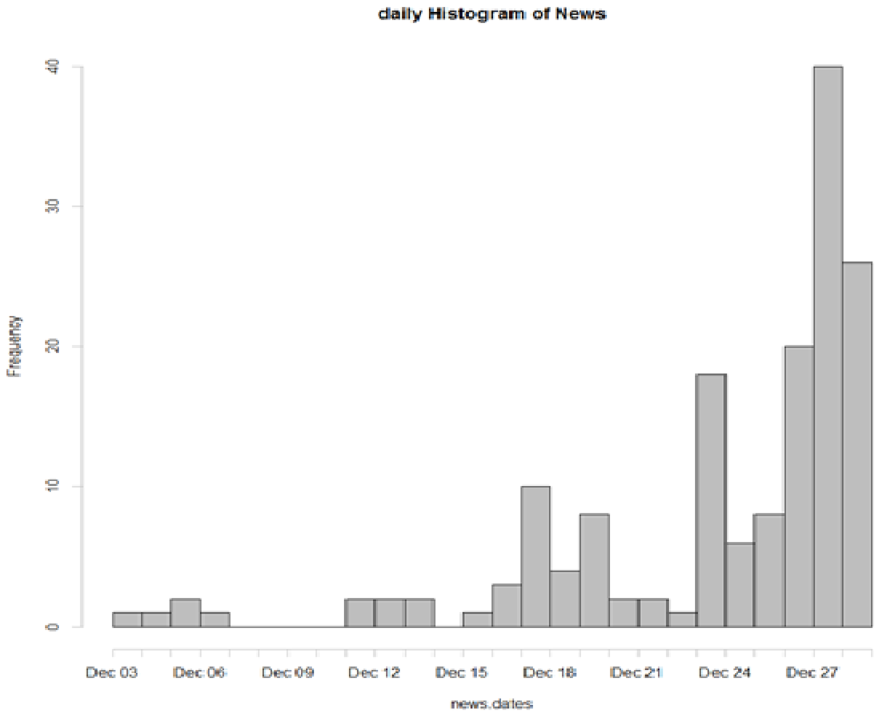

Figure 3.9: Showing Correlation between All and Daily Return

Chapter 4

The Portfolio Optimizer

Investing in the stock market though attractive is a complicated process that should be approached with care. There is always a risk to lose may be of higher probability that the opportunity to gain especially for investors who try to rush without knowing the bigger picture and hence may get trapped. Big investors may have the tendency to take more risk as they are interested in more return. However, small investor and even cautious big investors may look into diversifying their investments in order to reduce the risk when some stocks suddenly invert the trend from up to down. To facilitate more systematic way for diversifying the investment, it is necessary to develop an automated process that could investigate the whole situation of the stocks globally to group them such that each group contains stocks which mostly move together either up or down. Once the informative grouping is obtained, investors may depend on that to decide on the stocks to invest in. This may also be used as a recommender to investors to guide them by providing packages of stocks with varying level of risk ranging from low risk to high risk. In other words, a systematic approach should be capable of showing the investor the big picture based on the level of risk the investor is willing to take.

The approach described in this chapter addresses the diversification problem by handling the portfolio optimization process. The reported test results are encouraging and show the great potential of the proposed approach.

4.1 Methodology

We started by downloading the adjusted daily price of stocks and calculated the daily return of each stock. For the second step in this section, we calculated the correlation coefficient between each pair of stocks in the related indices. We calculated the correlation coefficient, because correlation is a way of finding the similarity or dissimilarity between objects. The value of the correlation coefficient varies between -1 to +1. The -1 value means that the pair of stocks react completely differently to specific financial events, while the +1 value of correlation between a pair mean they are moving in exactly the same direction when there is a financial event, and 0 value means the two stocks are uncorrelated.

We need the distance between each pair of stocks in order to construct the dendrogram tree structure of objects and hence produce the complete clustering options. We cannot use the correlation matrix to do the hierarchical clustering analysis because it does not satisfy the following three requirements of distance matrix:

$$d\left(i,j\right) = 0 \qquad iff\ i = j$$

$$d\left(i,j\right) = d(j,i)$$

$$d\left(i,j\right) \leq d\left(i,k\right) + d(k,j)$$

Therefore we converted the correlation matrix to distance matrix by the following formula:

$$d\left(i,j\right) = \sqrt{2(1-\rho_{ij})}$$

We computed the k-means clustering of our dataset, 30 stocks, with the supposed optimal number of clusters. We have to note that the number of different sectors in this sample is 9. After that we calculated the Entropy of each Sector.

Then we implemented Hierarchical Cluster Analysis. After that we constructed our portfolios with the analysis from Hierarchical Clustering and K-means. Finally, we compared our portfolios with the optimal portfolios.

Our framework is presented in Figure 4.1.

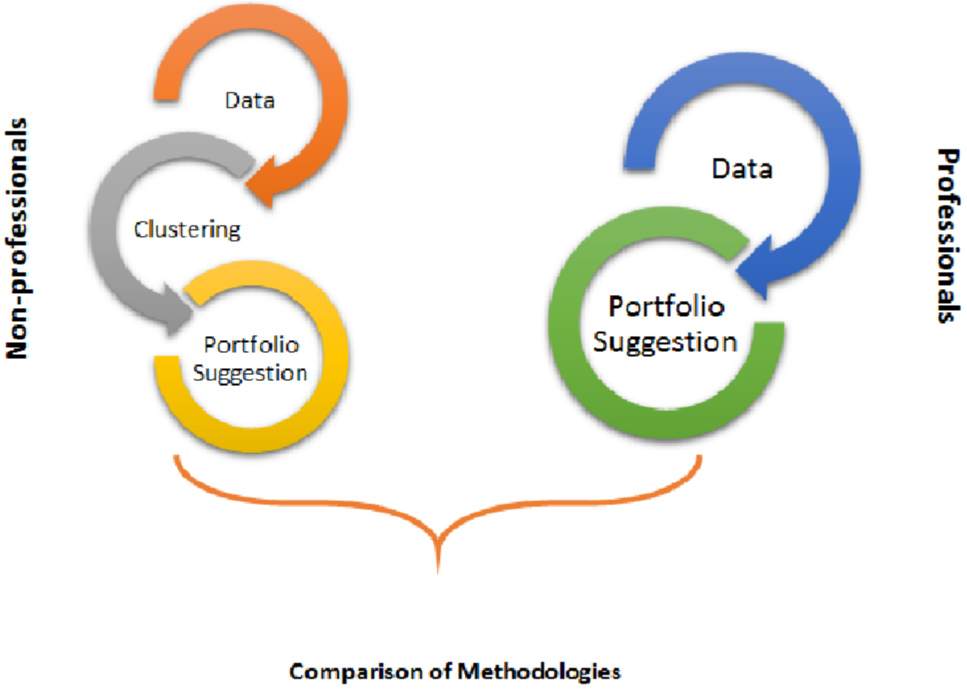

Figure 4.1: Framework Design

4.2 Evaluation

4.2.1 Partitioned clustering (K-means)

In the example given below in Figure 4.2 we can clearly see that there is a knee on the curve, and therefore the optimal number of clusters for this dataset is 9. In other words, our number of sectors is 9 and hence we can give the cluster number according to the sector number.

The Visualization for k-means is shown in Figure 4.3. The red bold edge connecting the two nodes (stocks) XOM and CVX means these two stocks have high correlation. Accordingly, we will give the two stocks XOM and CVX zero weight when we will create the constraints for our portfolio. Also, all the stocks in the technology, transportation, and energy are clustered based on the sector. Therefore, we will add other constraints that will limit the maximum weight of these sectors in the portfolio.

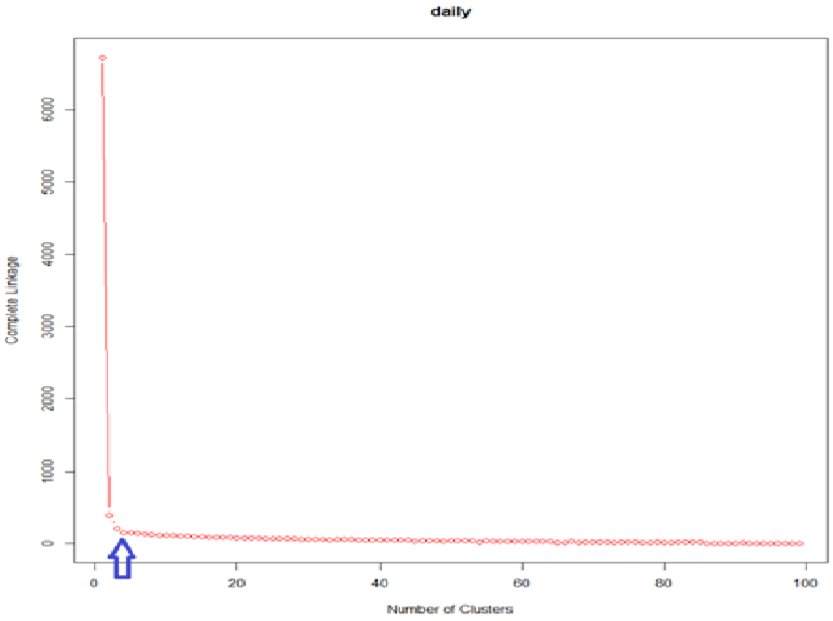

Figure 4.2: Optimal Cluster Number

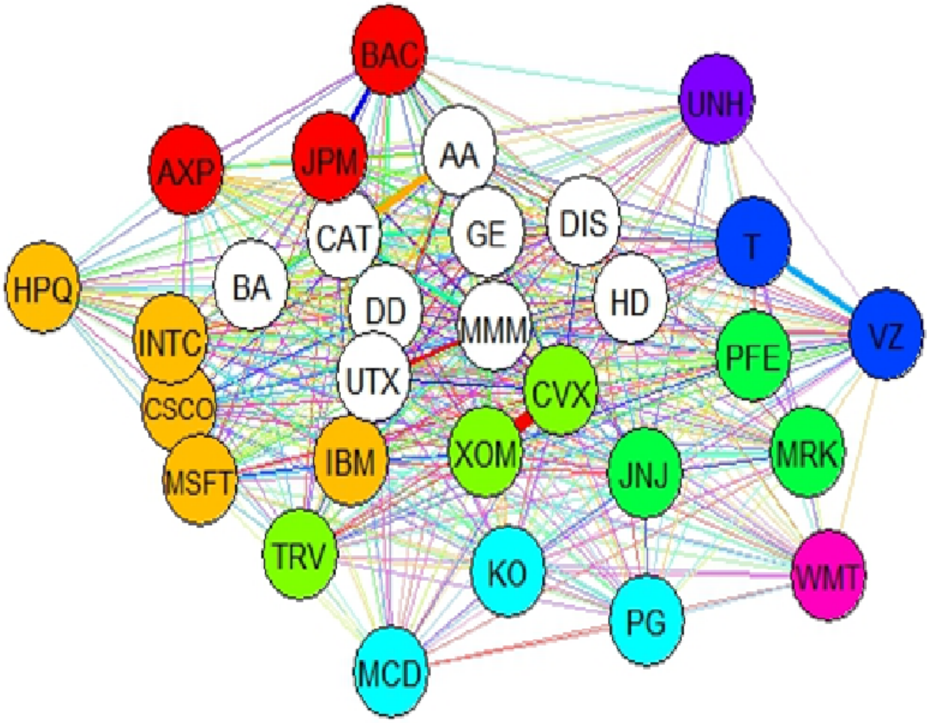

Figure 4.3: Visualization of K-Means in the Graph

We use the entropy to evaluate which of the sectors were easy to evaluate, which were homogeneous, and which were heterogeneous as shown in Table 4.1.

Basic Industries	**Capital Goods**	**Energy**	**Finance**	**Health Care**
1.58	1.50	0.00	0.00	1.50
Miscellaneous	**Public Utilities**	**Technology**	**Transportation**	
1.50	1.58	0.00	0.00	

Table 4.1: The Entropy for each Sector

We know that the higher the entropy is the less efficient the K-means algorithm is to predict the corresponding class. We can see that the Public Utilities sector and the Basic Industries sector have high entropy. Based on this, we can deduce that the companies within this sector are quite different; at least these companies have stocks with different behavior. Comparatively, the Technology, Energy, Transportation, and Finance sectors have lower entropy which means the companies within these sectors are more similar. This makes sense when we consider the fact that companies in these sectors are considerably large companies and very rare in their market.

According to the economy theory, the more different and heterogeneous the stocks contained in a portfolio are the less risky the portfolio is.

4.2.2 Hierarchical Clustering

We tried the three types of hierarchal clustering as can be easily seen in the results reflected in the Figures 4.4, 4.5 & 4.6 below. For our data the best graph-based approach is to use Wards method. We have also computed a complete-link clustering dendgram for our data set. It is not as clearly defined nor as optimal as the clustering outcome obtained using Wards method. In Wards method it is possible to see the stocks which are grouped together actually belong to the same sector; for example all the four technological companies (Microsoft, GOOG, IBM and CISCO) are grouped together. The complete clustering outcome allows to easily identify cases which are outliers and less susceptible to noise.

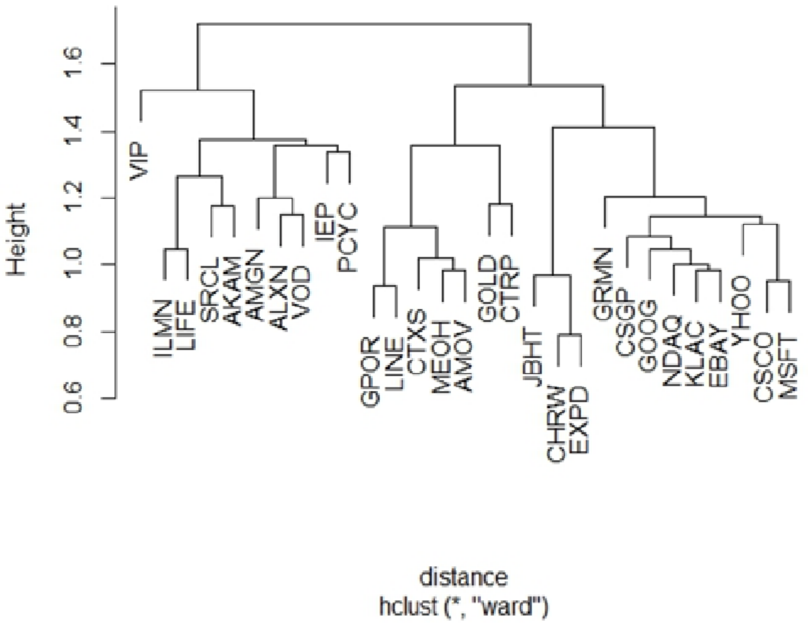

Figure 4.4: Hierarchical Cluster - Ward Method

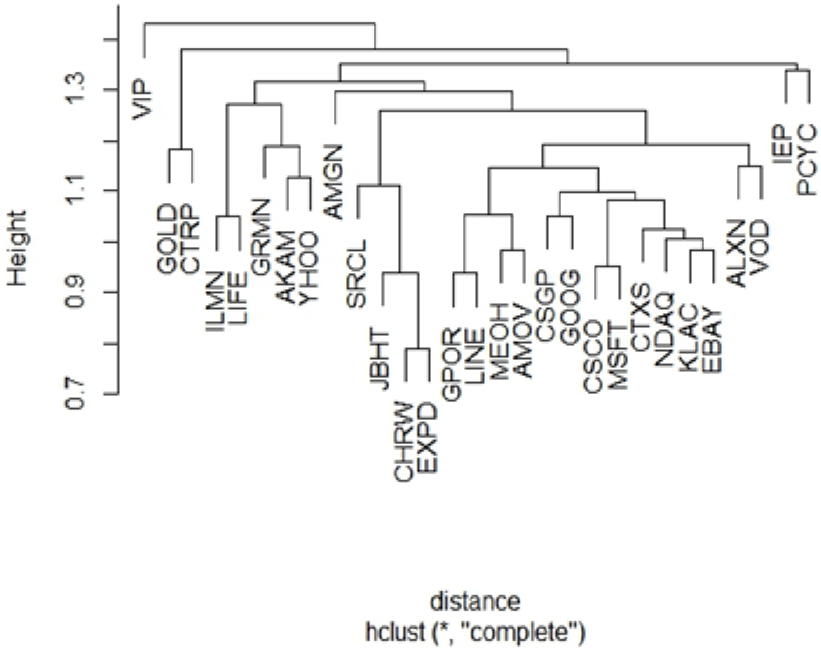

Figure 4.5: Hierarchical Cluster - Complete Method

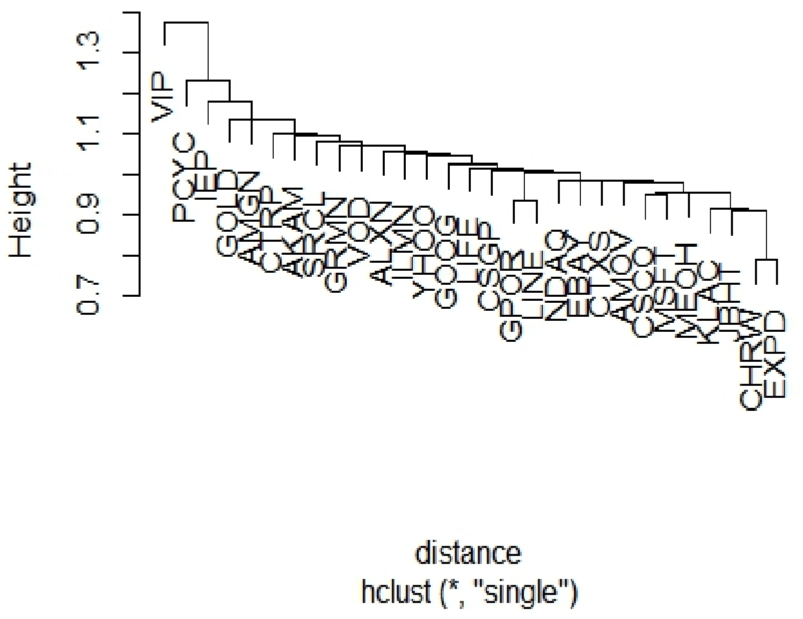

Figure 4.6: Hierarchical Cluster- Single Method

We construct our portfolio shown in Figure 4.7 with information that we pull out of the dendgram. We will try to reduce the risk by adding more constraints. Here are the constraints:

- The minimum weight for each stock is zero. This constraint is the long selling only constraints, short selling is not allowed.
- Stocks with symbols VIP, PCYC, and IEP should have at least 5% share of investment each. These stocks which have highest distance with other stocks and are only single member clusters will help a lot to diversify the portfolio. Based on this, we can emphasize that one should invest at least 5% or much more of his investment in each of these stocks.
- Do not invest any funds in CHRW and EXPD. These two stocks have the highest correlation (lowest distance) and both are in exactly the same sector and sub sector industry so we omit them.

- The maximum sum of investment in the technology sector is 20% or less. This is to impose upper limit on the four companies in the Technology sector. All these stocks in the Technology sectors should not have together more than 20% of all the investment.

- The maximum sum of the investment in the two sectors Basic Industries and Energy should not be more than 10% of all the investment.

The return, risk and VaR of the constructed portfolio are specified as:

r = 0.0043; Sigma = 0.0163; VaR = 0.0237

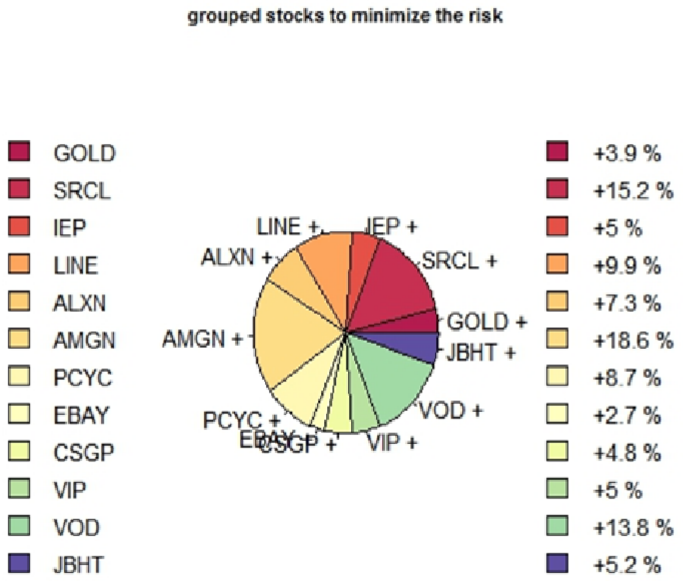

Figure 4.7: Grouped Stocks to Minimize the Risk

We construct an equal-weight portfolio of all the 30 stocks. Each stock has weight of 0.0333. The return, risk and VaR of the constructed portfolio are:

r = 0.0043; Sigma = 0.0198; VaR = 0.0291

It can be easily seen that the return of the portfolio is the same but the risk is reduced by 17% and the VaR is reduced by 18%. Further, the pie chart shown in Figure 4.7 reflects the weight of the portfolio that satisfies all of our constraints.

It is possible to use our visualization graph for k-means and the dendgram tree as the source of information from stocks, and by analyzing them carefully we construct a better portfolio from stocks. Having the same target risk has equal weight portfolio shown in Figure 4.8, but portfolio with lower risk is what every investor with different investment strategy of risk would appreciate.

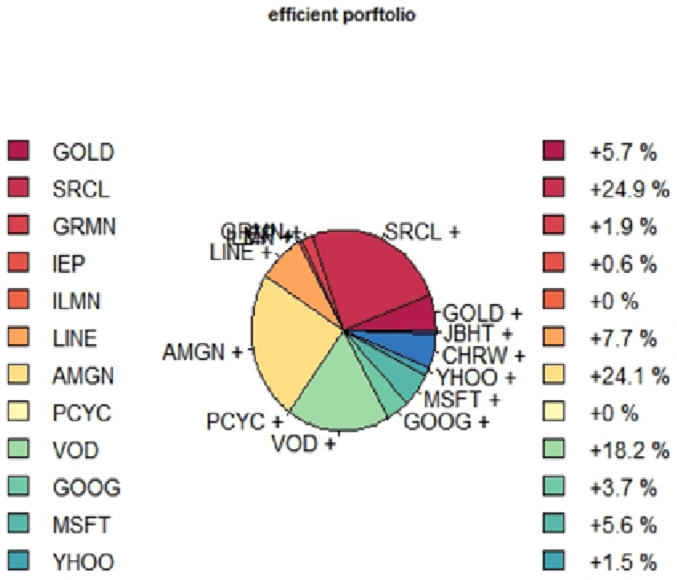

Figure 4.8: Efficient Portfolio

Chapter 5

Conclusions and Future Work

The Stock market is one of the main attractions for investors, especially after the widespread of the technology including handheld devices which increased the availability of data and resources. An investor may stay connected and watch the market movements on the spot since the market opens until it closes regardless of his/her location and time zone. Wherever there is internet connection it is possible for an investor to watch the market and act accordingly. However, watching the market alone is not enough for informative decision making. Several other factors should be considered to avoid surprises that might lead to loss due to selling at an inappropriate time or missing the opportunity to buy at the appropriate time. Thus, a comprehensive framework capable of integrating historical data analysis with social media and news analysis would lead to more informative decision making process and may reduce the risk. Further, optimizing the portfolio is another important issue to consider when it comes to the level of risk an investor is willing to consider. All these have been incorporated in the framework described in this thesis.

5.1 Conclusions

In this study, we made a sentimental analysis for stock market trend analysis to help in predicting actual stock price index movement. This is an important aspect that has attracted the attention of researchers. According to the accuracy of prediction results, the investors may make more gain or lose. In this study we demonstrated how it is possible to make a better prediction by benefiting from different resources, such as social media analysis or sentiment analysis from articles. But, these tasks are highly complicated and very difficult. We need to use multidisciplinary approach to get a high rank by considering techniques from different research areas.

We have been able to see and realize clearly that there is a strong relationship between daily return prices and news articles. This is well supported by the reported experimental results.

Away from sentimental analysis, portfolio generation is an important issue for making suggestions by finding the best combination of stocks that meet the expectations of investors who are willing to go for a certain level of risk. However, it is a complicated problem involving many factors to deal with, such as risk and profit. Professional investors consider lots of criteria to find the best combination but non-professional investors may get lost because they mostly do not have any idea how to make appropriate and guided investment and what are the best criteria to consider. In this study, we demonstrated how an automated process is capable of providing guidance for investors by producing appropriate portfolios that match their expectations and needs.

Being able to recognize the difference between the assets and to be able to pick a very heterogeneous group of stocks is a motivation of every financial manager. In this thesis, we applied data mining and machine learning Techniques to differentiate several clusters of assets and proved that constructing a portfolio using cluster analysis is efficient and effective.

We computed the entropy of the different sectors in comparison with a K-means clustering on a dataset of DJIA stocks and we ended with different levels of entropy regarding the sectors. This means sectors are predictable even though some are more heterogeneous than others. Then, we implemented Hierarchical Cluster Analysis to classify the stocks in groups. Finally, we constructed our portfolios based on the data extracted from the analysis which explained the stocks behavior with respect to each other in the group that they belong to and in comparison to other groups.

5.2 Future Work

The work described in this thesis forms a first step in the right direction. It helped us in getting involved in the research related to stock market analysis and portfolio optimization. The obtained results are encouraging and we could build on what has been achieved so far to produce a more sophisticated system with more functionalities. For further steps and improvements, we are planning to improve our dictionary by considering resources that can provide more accurate results; alternatively we will try to implement another solution like rule based model. We also plan to increase our news resources in order to increase the daily prediction score by considering for instance more financial news resources related directly to the financial area. Another area we plan to investigate further is that we want to conduct social network analysis to predict hidden values between companies and this will help us to suggest best combinations of stocks to investors. Also, we plan to compare to our suggestion results with the Markowitz model. Finally, we will also develop a multi-agent based approach for sentimental analysis and portfolio optimization. We want to incorporate in the framework multiple perspectives from various available brokers and we will try to rank and prioritize the available brokers in order to provide the investors with the knowledge that may be needed for a more informative decision making.

Bibliography

[1] N. Amenc and V. Le Sourd. *Portfolio theory and performance analysis.* John Wiley & Sons, 2005.

[2] N. Analytics. http://www.newsanalytics.net, October 2013.

[3] R. V. Argiddi and S. Apte. Future trend prediction of indian it stock market using association rule mining of transaction data. *International Journal of Computer Applications*, 39, 2012.

[4] G. S. Atsalakis and K. P. Valavanis. Surveying stock market forecasting techniques–part ii: Soft computing methods. *Expert Systems with Applications*, 36(3):5932–5941, 2009.

[5] Y. L. Becker, H. Fox, and P. Fei. An empirical study of multi-objective algorithms for stock ranking. In *Genetic Programming Theory and Practice V*, pages 239–259. Springer, 2008.

[6] S. W. Chan and J. Franklin. A text-based decision support system for financial sequence prediction. *Decision Support Systems*, 52(1):189–198, 2011.

[7] K. Daniel, D. Hirshleifer, and A. Subrahmanyam. Investor psychology and security market under-and overreactions. *the Journal of Finance*, 53(6):1839–1885, 1998.

[8] D. Enke and S. Thawornwong. The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with applications*, 29(4):927–940, 2005.

[9] S. Epstein. Integration of the cognitive and the psychodynamic unconscious. *American psychologist*, 49(8):709, 1994.

[10] A. Fan and M. Palaniswami. Stock selection using support vector machines. In *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*, volume 3, pages 1793–1798. IEEE, 2001.

[11] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. *ICWSM*, 7, 2007.

[12] S. C. Gold. Computerized stock screening rules for portfolio selection. *Financial services review*, 8(2), 1999.

[13] R. Goonatilake and S. Herath. The volatility of the stock market and news. *International Research Journal of Finance and Economics*, 3(11):53–65, 2007.

[14] M. Hagenau, M. Liebmann, and D. Neumann. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, 55(3):685–697, 2013.

[15] D. Hillier, S. A. Ross, R. W. Westerfield, J. Jaffe, and B. D. Jordan. *Corporate finance: 1st european edition*. Number 1st Edition. McGraw-Hill, 2010.

[16] A. K. Jain and R. C. Dubes. *57Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[17] G. H. John, P. Miller, and R. Kerber. Stock selection using rule induction. *IEEE Intelligent Systems*, 11(5):52–58, 1996.

[18] Y. Kara, M. Acar Boyacioglu, and Ö. K. Baykan. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert systems with Applications*, 38(5):5311–5319, 2011.

[19] K.-j. Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1):307–319, 2003.

[20] N. Koochakzadeh, F. Keshavarz, A. Sarraf, A. Rahmani, K. Kianmehr, M. Rifaie, R. Alhajj, and J. Rokne. Stock investment decision making: A social network approach. In *Emerging Intelligent Technologies in Industry*, pages 47–57. Springer, 2011.

[21] N. Koochakzadeh, K. Kianmehr, A. Sarraf, and R. Alhajj. Stock market investment advice: A social network approach. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 71–78. IEEE Computer Society, 2012.

[22] W. R. Lasher. *Practical financial management*. Cengage Learning, 2013.

[23] W. R. Lasher. *Practical financial management*. Cengage Learning, 2013.

[24] H. Markowitz. Portfolio selection*. *The journal of finance*, 7(1):77–91, 1952.

[25] H. M. Markowitz. *Portfolio selection: efficient diversification of investments*, volume 16. Yale University Press, 1970.

[26] A. Moreo, M. Romero, J. Castro, and J. M. Zurita. Lexicon-based comments-oriented news sentiment analyzer system. *Expert Systems with Applications*, 39(10):9166–9180, 2012.

[27] A. Ng and A. W.-C. Fu. Mining frequent episodes for relating financial events and stock trends. In *Advances in Knowledge Discovery and Data Mining*, pages 27–39. Springer, 2003.

[28] A. Nicholas Refenes, A. Zapranis, and G. Francis. Stock performance modeling using neural networks: a comparative study with regression models. *Neural Networks*, 7(2):375–388, 1994.

[29] A. Nikfarjam, E. Emadzadeh, and S. Muthaiyah. Text mining approaches for stock market prediction. In *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*, volume 4, pages 256–260. IEEE, 2010.

[30] P. Nizer and J. C. Nievola. Predicting published news effect in the brazilian stock market. *Expert Systems with Applications*, 39(12):10674–10680, 2012.

[31] K. Oatley and P. N. Johnson-Laird. Towards a cognitive theory of emotions. *Cognition and emotion*, 1(1):29–50, 1987.

[32] T. Odean. Volume, volatility, price, and profit when all traders are above average. *The Journal of Finance*, 53(6):1887–1934, 1998.

[33] R. Pack. http://www.sentimentnews.com, October 2013.

[34] P.-F. Pai and C.-S. Lin. A hybrid arima and support vector machines model in stock price forecasting. *Omega*, 33(6):497–505, 2005.

[35] A. Refenes, M. Azema-Barac, and A. Zapranis. Stock ranking: Neural networks vs multiple linear regression. In *Neural Networks, 1993., IEEE International Conference on*, pages 1419–1426. IEEE, 1993.

[36] G. D. Samaras, N. F. Matsatsinis, and C. Zopounidis. A multicriteria dss for stock evaluation using fundamental analysis. *European Journal of Operational Research*, 187(3):1380–1401, 2008.

[37] G. D. Samaras, N. F. Matsatsinis, and C. Zopounidis. A multicriteria dss for stock evaluation using fundamental analysis. *European Journal of Operational Research*, 187(3):1380–1401, 2008.

[38] R. P. Schumaker, Y. Zhang, C.-N. Huang, and H. Chen. Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3):458–464, 2012.

[39] SemEval. http://mpqa.cs.pitt.edu, October 2013.

[40] SemEval. http://www.cse.unt.edu/ rada/affectivetext/, October 2013.

[41] P. Sevastjanov and L. Dymova. Stock screening with use of multiple criteria decision making and optimization. *Omega*, 37(3):659–671, 2009.

[42] F. E.-H. Tay, L. Shen, and L. Cao. *Ordinary shares, exotic methods: Financial forecasting using data mining techniques*. World Scientific, 2003.

[43] P. C. Tetlock. Does public financial news resolve asymmetric information? *Review of Financial Studies*, 23(9):3520–3557, 2010.

[44] S. Theußl, I. Feinerer, and K. Hornik. Distributed text mining with tm. In *The R User Conference*, 2009.

[45] E. S. Tools. http://www.eventstudytools.com, October 2013.

[46] A. H. Van Bunningen. Augmented trading-from news articles to stock price predictions using syntactic analysis. 2004.

[47] B. Wang, H. Huang, and X. Wang. A novel text mining approach to financial time series forecasting. *Neurocomputing*, 83:136–145, 2012.

[48] W.-B. Yu, B.-R. Lea, and B. Guruswamy. A theoretic framework integrating text mining and energy demand forecasting. *IJEBM*, 5(3):211–224, 2007.

[49] Z.-H. Zhou. Three perspectives of data mining, 2003.

[50] P. Zorn, M. Emanoil, L. Marshall, and M. Panek. Finding needles in the haystack: Mining meets the web. *Online*, 23(5):16–18, 1999.

Appendix A

DJIA Stocks

Stock Symbol	Company	Sector
MCD	MC Donalds	Consumer Goods
KO	Coca Cola	Consumer Goods
PG	Procter & Gamble	Consumer Goods
PFE	Pfizerc Inc.	Health Care
MPK	Merk and Co Inc.	Health Care
JNJ	Johnson & Johnson	Health Care
UNH	United Health Group	Health Care
VZ	Verizon Communication Inc.	Telecom
T	AT & T	Telecom
INTC	Intel Co	Technology
IBM	International Business Machine	Technology
MSFT	Microsoft	Technology
CSCO	Cisco System	Technology
HPQ	Hewlett- Packard	Technology
UTX	United Technologies Co	Industrial Goods
CAT	Caterpillar	Industrial Goods
AA	Alcoa	Basic Material
DD	Dupont	Basic Material
MMM	3M	Conglomerates
GE	General Electric	Conglomerates
DIS	Walt Disney	Entertainment
XOM	Exxon Mobile	Oil & Gas
CVX	Shevron Co	Oil & Gas
BA	Boeing	Aerospace
HD	The Home Depot	Service
WMT	Wall-Mart	Service

Continued on next page

Table A.1 – continued from previous page

Stock Symbol	Company	Sector
JPM	JPMorgan	Financial
BAC	Bank of America	Financial
AXP	America Express	Financial
TRV	Travelers	Financial

Table A.1: DJIA Stocks List