DOW JONES INDEX CHANGE PREDICTION USING TEXT MINING

Marcos Neves do Vale

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Civil.

Orientador: Nelson Francisco Favilla Ebecken

Rio de Janeiro

Junho de 2018

DOW JONES INDEX CHANGE PREDICTION USING TEXT MINING

Marcos Neves do Vale

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA CIVIL.

Examinada por:

_____
Prof. Nelson Francisco Favilla Ebecken, D.Sc.

_____
Prof. Afonso Celso de Castro Lemonge, D.Sc.

_____
Prof. Lucio Pereira de Andrade, D.Sc.

_____
Profª. Solange Guimarães, D.Sc.

_____
Profª. Regina Celia Paula Leal Toledo, D.Sc.

_____
Prof. José Luis Drummond Alves, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
JUNHO DE 2018

Vale, Marcos Neves do

Dow Jones Index Change Prediction Using Text Mining/Marcos Neves do Vale. - Rio de Janeiro: UFRJ/COPPE, 2018.

XII, 117 p.: il.; 29,7 cm.

Orientador: Nelson Francisco Favilla Ebecken

Tese (doutorado) - UFRJ/COPPE/ Programa de Engenharia de Civil, 2018.

Referências Bibliográficas: p. 96-109.

1. Finanças. 2. Mineração de Textos. 3. Mineração de Dados. I. Ebecken, Nelson Francisco Favilla. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

Acknowledgments

To my family, who always trusted me and gave me inspiration and motivation to move forward.

To my friend Marcelo Beckmann which helped me a lot in this work, giving me good insights and expertise that lightened my way.

To my friend Carlos Alberto Lanzarine Casa whom I have known for many years and contributed by sharing with me his economic knowledge.

To my friend Gustavo Medeiros who helped me with some good insights that saved me some time.

To my American friend Brooke Boening who helped me a lot by correcting the English errors. She did an excellent job in such a short time.

To my teachers, especially Nelson Francisco Favilla Ebecken of COPPE/Federal University of Rio de Janeiro, who is not only a great teacher but also a great person always willing to help.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

PREDIÇÃO DA VARIAÇÃO DO ÍNDICE DOW JONES UTILIZANDO MINERAÇÃO DE TEXTOS

Marcos Neves do Vale

Junho/2018

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

Os recentes avanços nas técnicas de mineração de dados e textos permitem novas pesquisas sobre a previsão do mercado financeiro (TMFP). O objetivo deste trabalho é apresentar um novo modelo de previsão para as tendências do índice Dow Jones ao longo do dia. O modelo foi desenvolvido utilizando o RapidMiner juntamente com scripts SQL. O modelo utiliza-se de processos de mineração de texto existentes e uma nova técnica de alinhamento de dados obtida, de modo geral, a partir da coleta das notícias publicadas pelo *Yahoo Finance* e pelo *Google Finance* correspondentes às 5 ações com maior volume de negociações por minuto. A qualidade do modelo é medida pelos índices *Precision*, *Recall* e *F-Measure*. Os resultados obtidos foram excelentes e superam as técnicas descritas até o momento para esse fim. Além disso, o modelo mostrou-se robusto e eficiente, demonstrando que a utilização de técnicas de mineração de texto juntamente com a estratégia correta aplicada no mercado financeiro é uma alternativa a ser considerada e contribui para o estado da arte nessa área de pesquisa.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

DOW JONES INDEX CHANGE PREDICTION USING TEXT MINING

Marcos Neves do Vale

June/2018

Advisor: Nelson Francisco Favilla Ebecken

Department: Civil Engineering

The recent advances in data and text mining techniques are enabling new research on financial market prediction (TMFP). The purpose of this work is to present a new prediction model for Dow Jones index trends throughout the day. The model was developed using RapidMiner along with SQL scripts. The process uses existing text mining processes and a new alignment technique that is briefly made by picking up the news published by YahooFinance and Google Finance corresponded to the 5 stocks with highest trading volume in each minute. The quality of the model is measured by Precision, Recall and F-Measure indices. The results obtained were excellent and surpass existing techniques today and also in the literature for this purpose. The model proved to be robust and efficient, demonstrating that the use of text mining techniques along with the correct strategy applied in the financial market is an alternative to be considered and contributes to the state of the art in this area of research.

Table of Contents

List of Figures

List of Tables

CHAPTER 1 – Introduction

Internet has become more popular and is a large part of many people's lives. Each day, (the) dependency on it increases in many areas such as education, technology and the financial markets. The speed at which data is produced has increased to a degree/at a rate that is impossible to process and it has encouraged research in many areas, including data mining and text mining. These two areas have emerged in the last decade mainly due to research in artificial intelligence, machine learning and inferential statistics. These recent studies have contributed significantly to the financial market, which, through the constant need to understand the market movement and predict future movements, shows a very promising area for the application of these new techniques.

There are many resources available to financial market players in investment decision making, with the results reported by the Stock Exchange Indexes being the most reliable indicators in the world both for the representation of market movements over a given period and for the decision-making moments of the stock market agents that rely on their performance to invest.

Considering that the expansion of the connections between the international financial markets is a phenomenon which has intensified since the twentieth century and that the impact generated by its contagion effect affects both the global economy and societies (VARTANIAN, 2012), understand and predict the behavior of the series of shares and the indexes that compose the market is an important tool to be used as a way to choose better investments and generate profit (ADRIÃO, 2009).Thus, the principle of pricing financial products based on their inherent risk characteristics has dominated theories of valuation in the last 50 years (WRIGHT *et al.*., 2011).

The theory of Efficient Market Hypothesis (EMH), for example, states that the equity prices, always fully reflects all available information (FAMA, 1965). It represents the standard view of economists about the stock market. It says that for efficient markets, when new information becomes available, the stock market changes, behaving like the economists say as a "random walk". However, nowadays, this theory is not supported by all economists.

On the other hand, the Dow Theory that had its origin in late 19th century with Charles H. Dow relies on measure of overall business conditions within the economy and by analyzing those conditions one could identify the direction of major market trends. Dow first used his theory to create the Dow Jones Industrial Index (DJIA) and the Dow Jones Rail Index (DJRI). He believed that these indexes were an accurate reflection of the business conditions within the economy (BROWN, *et al..*, 1997).

However, according to the data POON and GRANGER (2003) in the period from 1976 to 2002, in view of the diversity of results obtained by these models, it is assumed that, to date, there is no significant method to predict the volatility of index numbers, thus configuring a wide field for research and discussions.

When Google first launched their Initial Public Offering (IPO) in August 2004, for example, it increased 18% in just one day. At the end of 2004 the price had reached over than 100% of its initial value. The next three years were pretty impressive, making Google the fifth largest company in the US. In 2008 the share price dropped 40 percent. Only in late 2011 was the company able to recover much of its value. This episode is not well supported by any of the theory mentioned above.

Each theory has one view of the proble, which helps us to understand one aspect of the market movement but fails when predicting certain scenarios. That is where data mining techniques can be used in order to achieve better results by using computer techniques to gather and process all current and historical information which is humanly impossible, in an attempt to locate the correct correlation that can identify the direction (of the market). Thereby, the purpose of this work is to present a new prediction model for Dow Jones index trends throughout the day. This model is based on text mining techniques applied to the news from companies that compose the index. Briefly speaking, the process developed cleans the data, labels, classifies and simulates future trends for the Dow Jones index.

This thesis is organized as follows: Chapter 2 presents the DJIA index; Chapter 3 describes the Data Mining concepts and techniques used; Chapter 4 presents the Text Mining techniques; Chapter 5 presents the new methodology developed in this work; and Chapter 6 presents the experiments. Finally, Chapter 7 presents the conclusions.

CHAPTER 2 – Dow Jones Industrial Average (DJIA) Index

2.1 Index Numbers

In general, index numbers are defined as statistical indicators used to measure relative impairment between variables that are related to each other, whether qualitative or quantitative, over a given period of time or across the space (regional comparisons) thus obtaining an overall expression of data that has different degrees of importance in different fields (MERRILL and FOX, 1977, FONSECA *et al..*, 1985, CARVALHO, 1975).

In the specific case of the economic approach, the calculation of index numbers is closely associated with the need to infer about the welfare of a collectivity, being used to observe the evolution of the price of a certain product in relation to others traded in the same market, as well how to measure the real income of the economy and the cost of living (CARVALHO, 1975). For FONSECA *et al.* (1985, p. 158), knowledge about index numbers is indispensable to any economist, whether in macro or microeconomics.

The contribution of economic statisticians over the years has been relevant for advancing the improvement of technical aspects of the construction of index numbers. The earliest research records on price indexes in history date back to the early eighteenth century, with estimation attempts made independently, and always with the same goal: to calculate the variations in the purchasing power of different currencies. According to DIEWERT (1993), the first proposal for a method of calculating a price index was presented by the Bishop of Ely, William Fleetwood, in 1707, from the analysis of the change in the value of money by comparing the expenses of an Oxford student in 1707 published in his book Chronicon Preciosum.

Over the years, other initiatives to construct price indices have been recorded in the literature, such as the analysis presented by Dutot (1738) and Gian Rinaldo Carli (1764)

(DIEWERT, 1987). However, weighting criteria was not considered by the majority of the proposed indices at the end of 18th century. Thus, at the beginning of the following century, Joseph Lowe stands out for his pioneering work in introducing the concept of weighting the indices when using a fixed basket to analyze the price variation DIEWERT (1993). In the development of its index, however, Lowe does not specify how the vector of quantities would be formed, resulting, therefore, in the limitation of its use.

Years later, failure to specify how the quantity vector would be determined in the index proposed by Lowe was solved by Etienne Laspeyres' subsequent work in 1871 and Hermann Paasche in 1874 (MERRILL and FOX, 1977, CARVALHO, 1975, DIEWERT, 1993). Lastly, MERRILL and FOX (1977) highlight Irving Fisher's study of the Napoleonic wars and their effects on paper money, creating a calculation that made its index known as the ideal index, since it is the result of the geometric mean of the product between the Laspeyres and Paasche indices.

The elementary comparisons made with the use of index numbers in the economy revolve around prices and quantities. With regard to common markets for goods and services, it is easier to see its applicability, but it extends to more complex markets, such as the financial market, and capital markets.

2.2 Capital Markets and the Stock Exchange

Capital Markets can be defined as:

> "a set of institutions and instruments that deal with securities, aiming at channeling the resources of the buying agents to the selling agents. That is, the capital market represents a system of distribution of securities that has the purpose of enabling the capitalization of the companies and give liquidity to the securities issued by them". (PINHEIRO, 2009, p. 174).

Additionally, according to NETO (2010, p.69) the capital market is the main provider of financial resources for the financing of the economy, which in turn is directly dependent on the growth of companies (ZOTTE *et al.*., 2012). In this light, the Capital Market can be understood as a virtuous circle, where: the more investment in

the Market, the greater the economic growth and, consequently, the greater the income power to be directed again to the Market itself. For this reason, developed countries have proportionally more dynamic capital markets, more diversified and stronger.

Also, inserted in the Capital Market, there is the Stock Market, in which negotiations of credits issued by corporations occur. The trading of securities, in turn, takes place on two separate stages in two other independent markets: Primary Market and Secondary Market (FORTUNA, 1997). The Primary Market is characterized as the environment of the Stock Market in which the company itself issues shares or debentures, which are offered through a bank, in order to obtain resources for its ventures (FORTUNA, 1997, PINHEIRO, 2009). The Secondary Market is the one where shareholders wish to divest the assets they already own to recover the capital invested in them, as well as investors who wish to commit funds to shares already in circulation in the Market FORTUNA (1997); that is, it represents the transaction between buyers and sellers of shares through an entity other than the one that issued the bond in the primary market: the Stock Exchange.

The Stock Exchanges are therefore the main environment of the Secondary Stock Market, whose function is to promote appropriate conditions for the purchase and sale of securities, such as: preservation of ethical values in the negotiations, quick and efficient dissemination of results, security of transactions, registration of operations and enforcement of the normative instruments that regulate the operation of the exchanges (NETO, 2010, p. 188). FORTUNA (1997) also emphasizes that the Stock Exchanges are not financial institutions but rather non-profit associations that the brokerage firms seek to provide only the infrastructure necessary for the functioning of the Stock Market.

On a daily basis, on the Stock Exchange, there are so-called "trading sessions", where traders from the stock exchange meet to execute the orders to buy and sell securities given by investors to the brokerage firms they have hired (NETO, 2010). Open to trading on the Stock Exchange, the stock price in the market is defined according to the interaction of the demand and the offer presented in it, where the bigger the demand for a share, the greater the influence of its price increase / increase and vice versa. Thus, we can say that investors' behavior directly influences stock prices, and this behavior is influenced, among other factors, by specific index numbers, here called stock indexes or stock market indexes.

2.3 Stock Indices

The Stock Indexes indicate to the agents negotiating in the "trading sessions" the trends and expectations of the Stock Market, as well as the continuous performance of the assets. This information helps traders make decisions regarding how much, when and where to invest their (available) resources to generate more value. Risking resources without previous study is not rational and it is also not part of the standards followed by Capital Markets. In this way, the Stock Indexes are considered a valuable instrument for comparative performance and performance evaluation for managers and investors. FORTUNA (1997) emphasizes the analysis of the stock performance movement in relation to the stock market index in which it is traded as the first of the main direct indicators that influence investors in buying or selling stocks.

PINHEIRO (2009, p. 254) adds that stock indices comprise mostly weighted temporary numbers representative of the entire market and therefore can be considered a significant indicator of the price variation or market quotations used to monitor the behavior of the main shares traded on a stock exchange:

> "By comparing the indexes determined successively by stock exchanges, one can know if the market is up, stable or low, which guides investors in their applications in the near future. The monitoring of the index is usually done through a simple chart that records its evolution over time: a year, a month, a week or even a day."

However, it is known that for an index to be effectively used as a tool for evaluating the performance of markets and exchanges it must be composed of a portfolio of investments, that is, a set of securities, in order to benefit from the returns of this market and at the same time diversify its risks. According to FARIAS and SANTOS (2016), for analysts what is relevant is not the total value of the portfolio, but the changes in the index over a period of time, which when positive, represents profitability. When it comes to an investment portfolio that indicates the index will rise, NETO (2010, p.193) affirms that this should be composed of shares that correspond to the behavior of the market, prioritizing those most representative for each Stock Exchange, according to their respective percentage in the volume transacted in

the Market, and for each included share, one must attribute a weight according to its importance in the business of the theoretical investment portfolio. However, this broad concept of construction opens space for the adoption of different methodological characteristics in the preparation of financial market indices. According WEISS (2000, p. 38), these methodologies may differ as to: the calculation algorithm (arithmetic or geometric mean), the weighting (market value, prices, transaction volume, etc.), or the scope (universe of the assets or sample).

According BELLI (2002), the different existing stock price indices can be divided, in general, into four methodological generations, according to the main market value characteristics that incorporate them: the first was composed of indexes formed by samples and weighted by prices such as the Dow Jones and the Nikkei; the second includes samples of the pool of available assets, weighted, however, by market value (examples of this phase are the London FTSE-100 and the USA S & P 500); the third broadens the scope of the assets covered to the universe, weighted by the market value, as is the case of the NYSE, which includes the shares of the New York Stock Exchange and Topix Japan; and the fourth maintains the scope of the universe of shares, but instead considers the float instead of the market value (this index includes indexes created by financial consultants such as TSE300 and the SBWEI index family, produced by a group controlled by the company itself.

Recently, MELLAGI and ISHIKAWA (2007 p. 257-258) also mention a fifth category of indexes where the investment portfolio is based either on the volume traded on the cash market (liquidity) or on the accounting value of the issuing company. In this group, we can cite indices such as Ibovespa in Brazil and Merval in Argentina. However, regardless of the criterion used, in order for an index to be considered methodologically well prepared and accepted by market participants, it must follow some primary requirements (WEISS, 2000, p. 39):

- Relevance: reflect markets and assets of investors' interest;
- Scope: should include all available opportunities; x selection criteria; objectives: the rules of inclusion and exclusion of assets should be clear, simple and predictable and have the approval of investors;
- Subject to investment: market participants should be able to match index returns by acquiring their components so that comparisons

with the index are fair and performance rates over and above it are deserved; the CDI, for example, a common benchmark for fixed income investors, is only amenable to investment by treasuries, and not by individual or institutional investors;

- Component stability: a benchmark should reflect the market, adapting to its changes, but should have low turnover (turnover) of its components;

- Investment style and risk profile: should be clearly defined and informed to investors;

- Investment style and risk profile: wide and prompt disclosure of returns of the index, its components and methodology is crucial for its acceptance, as are the quality of the means of disclosure and the costs of obtaining it;

- Reliability: Errors in data processing and judgment regarding the inclusion or exclusion of components in the event of periodic rebalancing or corporate acts, as well as misinterpretation of rules or possible unwritten rules, may lead to loss of credibility of the index;

- Characteristics of the available components: information such as prices of assets and corporate events that affect it, the number of shares in the theoretical portfolio, the factor applicable to the market value or applicable float must be accessible to users; and

- Historical information: must be available".

2.4 DJIA Index

The Dow Jones Industrial Average (DJIA), which has existed since the late 19th century, is the best-known stock market index in the world, highlighting its remarkable ability to influence economies around the world. In this way, the DJIA is also the most widely published and discussed index.

The Dow Jones index was introduced in 1884 by the American Charles Henry Dow, a forerunner of the stock analysis, along with his partner Eduard Jones - the first editors of The Wall Street Journal- with the goal of setting up a company that would disclose stock quotes and economic news from the New York market (VIANA, 2009, LOPES, 2006), on the occasion/in response to(?) of a financial crisis attributed to an excess of investments in companies of the railroad that attacked the United States that same year (FARIAS and SANTOS, 2016).

Influenced by this scenario, Dow (1851 - 1902) concluded that "stock prices were in line with stock market trends as a whole, and thus introduced the use of averages, or indices, in that market in order to know the trends" (FARIAS and SANTOS, 2016, p. 493). Dow Jones & Company was born two years later (1896), the first index of the industrial sector was published: the Dow Jones Industrial Average (DJIA). In the beginning it was first composed by the 12 most important American companies (BEATTIE, 2017) mostly railroad builders, following the same simple arithmetic mean calculation of the first index created by Dow. Nowadays this index is composed by an average of 30 companies that are sometimes replaced according to market changes, in order to monitor the evolution of business on the New York Stock Exchange (NYSE).

BELLI (2002) explains that this index is formed by the average variation of the three roles that compose it. "It is considered because the choice of the companies that make up the Dow Jones index is based on criteria of size, economic tradition and solidity in the market," he concludes. However, despite the expansion of its investment portfolio, the methodology of the index remained almost unchanged. WEISS (2000, p.36) says:

> “Because their weighting is based on asset prices, those with higher prices have a greater influence on DJIA behavior. In addition to the low number of roles represented, other criticisms about its methodology are that it does not represent the most dynamic sectors of the economy throughout its history and that the shares of a company have a bonus or a split have its importance in the automatically reduced index as a consequence of this event”.

LEITE and SANVICENTE (1994, p. 14) define split as when the stock is divided into one or two parts, so that its quotation remains close to the average quotation of all the shares traded in a certain stock exchange. The objective of this stock split is to avoid the elitism of the market with high-performing shares, and this measure in no way diminishes the shareholder wealth. Faced with the split in shares that made up the Dow Jones index, it was necessary to adjust the average of the quotes in a way that did not reproduce the effects of variations caused by it.

Thus, according to BEATTIE (2017) the DJIA is now calculated by using the Dow divisor. This divisor is adjusted according to structural changes, splits and spin-offs in order to assure that the index will not be changed. It also states that the DJIA

index is not sensitive to how much a stock change affects a company and it is not composed of all companies. Notwithstanding those flaws, it is the index that best describes the American economy today.

The Dow index is calculated in US dollars and is held/operated under the responsibility of a Committee of Averages, which also makes rare changes in the components of the theoretical portfolio, in order to ensure continuity of the index. The criteria used for the selection of stocks that make up the portfolio are: the company's excellent reputation in the face of the market, sustained growth, high investor interest in its roles and prominence in its industry. The weighting of the DJIA is based on the price of the shares that comprise it, the greater weight being the one with the highest price in the market, and its basic reference date for the calculation is the first time it was disclosed on May 26 1896 (FARIAS and SANTOS, 2016).

Currently, there are more than twelve indices calculated by the company founded by Dow and Jones, with DJIA still retaining most of the characteristics of the first to be created and, even after hundreds of publications, it maintains a high reputation among the other indices which have emerged within the international stock market. How do they quote LEITE and SANVICENTE (1994, p. 61):

> "The century-old tradition that made Dow Jones Industrial the most important market indicator of all time in the world has resulted from the reliability it has gained among the investors who have recognized it and recognize a fair market representation, despite the limited sample based system and the exotic weighting system adopted by it".

In summary, the Dow Jones index is one of the most reliable indicators in the world and has since become an important reference for investors and managers of foreign resources, who rely on their performance to make investment decisions. Thus, the implementation of time series analysis techniques are extremely important in trying to understand and predict the behavior of the stock market (ADRIÃO, 2009) and could potentially be a decision support tool to help portfolio managers or traders of options on futures time the market. For instance, portfolio managers of mutual funds or institutional pension funds have to invest millions of dollars over a period as short as one week. They typically invest an equal amount of money each day (this is known as dollar cost averaging). However, if they have a prediction that the stocks are rising and they may themselves have the hunch that stocks are appreciating today, then they can

try timing the market. On certain days they would delay their investment (the stocks are expected to weaken but the market starts steady or strong); whereas on other days they might invest more and earlier in the day (when the closing value is expected to be up and the market starts weak (WÜTHRICH, 1998).

CHAPTER 3 – Knowledge Discovery in Databases

With the explosive growth of online data and the widespread use of databases in recent years, solutions based around their processing and application for decision support have become increasingly crucial, as well as the development of methodologies capable of intelligently and automatically transforming the processed data into useful information and knowledge (FAYYAD, 1996, FRAWLEY *et al.*., 1991, SILBERSCHATZ, 1995). This is because, contained within this large amount of data, there lies precious information at managerial and strategic level that can't be discovered by traditional database management systems.

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

The term KDD has been proposed as the most appropriate name for the general process of Knowledge Discovery (KD) (REINARTZ, 2002, CIOS & KURGAN, 2005) and concerns the entire process of knowledge extraction, including how data is stored and accessed, how to develop efficient and scalable algorithms that can be used to analyze massive data sets, how to interpret and visualize the results, and how to model and support the interaction between human and machine (FAYYAD *et al.*., 1996c, KLOSGEN & ZYTKOW, 1996). It is also defined as a nontrivial process of identifying valid, new, potentially useful, and ultimately understandable patterns of data (FAYYAD *et al.*., 1996b).

This definition is the most popular among the KD community, developed by reviewing the original definition published by FRAWLEY *et al*. (1991). It generalizes the application of the process to sources that are not database-based, although it emphasizes them as a primary source of data.

Such information obtained from the data analysis can be used for several applications, ranging from business management, and production control and market analysis to the engineering and scientific exploration project. For FAYYAD *et al*. (1996c), the convenience is not in archiving large volumes of data, but in our ability to derive useful reports to spot relevant events and trends, in order to justify decisions and policies made by companies and scientists based on statistical analysis and inference. As a result, researchers motivated by the challenge of transforming information into knowledge encounter Knowledge Discovery in Databases (KDD) early, focusing on the data mining (DM) application.

The expression KDD has emerged as a more appropriate alternative to naming the general process of Knowledge Discovery (KD) (REINARTZ, 2002, CIOS & KURGAN, 2005) and refers to any process of knowledge extraction, ranging from how data is accessed and stored, to the creation of dynamic algorithms capable of analyzing massive data set, predicting results and lead the interaction between human and machine (FAYYAD et al., 1996c, KLOSGEN & ZYTKOW, 1996). It is also defined as a nontrivial process of identifying valid, new, potentially useful, and ultimately understandable patterns of data (FAYYAD *et al.*, 1996b).

This definition follows the original description published by FRAWLEY *et al*. (1991) and is pointed out by CIOS & KURGAN (2005) as being the most popular among the KD community.

The KDD consists of a series of defined steps, each destined to the conclusion of a determined task of discovery, and realized by the application of a method of discovery (KLOSGEN & ZYTKOW, 1996):

1. **Data Integration:** data from different sources are collected and put together;
2. **Data Selection:** discard data that might not be useful for your research;

3. **Data Cleaning:** clean data by applying different techniques to deal with data errors, missing values, noisy and inconsistent data;
4. **Data Transformation:** data may need to be transformed into something appropriate for your model in order to speed up process, associate types, normalize values, etc.;
5. **Data Mining:** data discovery process to finding patterns. At this step, a data mining model will be applied into the data;
6. **Pattern Evaluation:** step to visualize the patterns generated; e
7. **Decision:** step responsible for helping user to understand the results in order to take better decisions.

This sequence comprises the cycle that the data travels until it becomes useful knowledge, as shown in Figure 1.

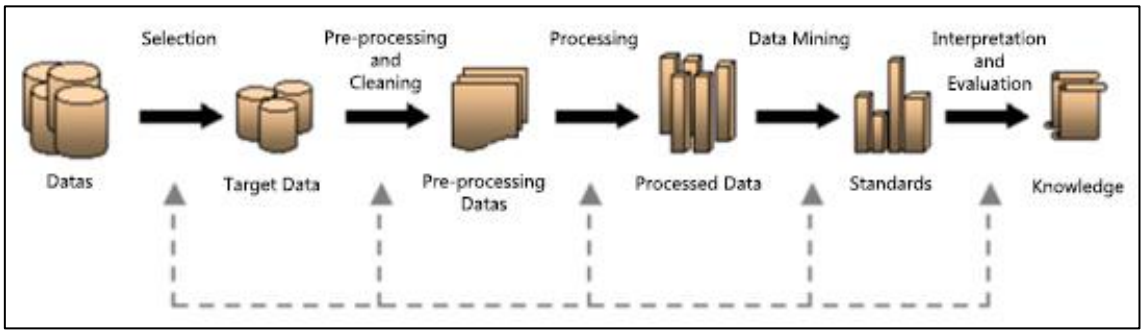

Figure 1- KDD process cycle (FAYYAD, 1996, p. 41)

Although all stages of the KDD process cycle must occur in the best possible way in order to achieve the desired result and thus must be/ are equally important for the transformation of the information into useful knowledge, the data mining phase can be considered the core of the whole process (PRASS, 2004). For many researchers, the term DM is also used synonymously with KD (REZENDE, 2005, WANG, 2005). However, it is worth emphasizing that for the location of the patterns to occur in a desired way, it is necessary to correctly perform the previous phases, and for each problem, there is a technique or algorithm that best fits in a given situation.

3.1 Data Mining

DM, a.k.a. knowledge extraction, data pattern processing, information discovery or data archeology (FAYYAD et al., 1996c), emerged in 1989 and consists of techniques and algorithms designed to analyze data or to extract patterns in specific categories of data (KLOSGEN & ZYTKOW, 1996).

Because it is considered multidisciplinary, the definitions about the term Data Mining vary with the field of action of the authors. In HAND *et al.* (2001), for example, the definition of DM is given from a statistical perspective: “Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner”.

In turn, FAYYAD *et al*. (1996) defines DM from a machine learning perspective as a step in the process of KD that consists in performing the data analysis and the application of discovery algorithms that, under certain computational limitations, produce a set of patterns of certain data. In addition, a third definition is presented by CABENA *et al.* (1998). For the author, DM joins knowledge machine techniques, pattern recognition, statistics, database and visualization, to be able to extract information from large databases.

DM knowledge of large databases has been recognized by many researchers as a key research topic in database systems and machine learning and by many industrial companies as an important area with great revenue opportunity. In addition, several applications emerging in information delivery services, such as data warehousing and online services over the Internet, also require various data mining techniques to better understand user behavior, improve service delivery, and increase opportunities of business (CHEN, 1996).

In these areas, the monitoring of data collected over time is used to make processes more efficient, effective, predictable and profitable. One of the biggest difficulties in extracting knowledge from data revolves around the need to invest in different areas such as statistical research, machine learning, optimization, data visualization, pattern recognition and high-performance computing in order to provide advanced business

intelligence and develop relevant analyses for business decision-making. This view is based on WIEDERHOLD (1996), who states:

> "... Knowledge Discovery is the most desirable end-product of computing. Finding new phenomena or enhancing our knowledge about them has a greater long-range value than optimizing production processes or inventories and is second only to task that preserve our world and our environment. It is not surprising that it is also one of the most difficult computing challenges to do well ..." (WIEDERHOLD, 1996).

Currently, different models are proposed to define and standardize the phases and activities of DM. In spite of the particularities, all in general contain the same structure. In this thesis, we chose CRISP-DM (Cross-Industry Standard Process of Data Mining) (CHAPMAN, 2000) as a model, due to the vast literature available and the accepted standard (HAND, 2001, LAROSE 2005).

3.1.1 CRISP-DM Methodology

The CRISP-DM work model emerged in 1996 from the initiative of data mining professionals and sought to develop a process model capable of operating in any type of industry, free and non-proprietary, able to fill this gap. With the aim of encouraging best practices and offering organizations the structure needed to realize better and faster results from data mining, the CRISP-DM operates using information from more than 200 DM users and tools, as well as service providers (SHEARER, 2000). For this, the model organizes the data mining process into six phases organized in a cyclical way (Figure 2), that help organizations understand the DM process and provide a road map to follow while planning and carrying out a project (SHEARER, 2000, OLSON and DELEN, 2008). Also, each of the six phases of CRISP-DM includes its own tasks, presented in Table 1. The step-by-step CRISP-DM reference model is summarized below.

1. **Business Understanding:** In this first step one must understand the goal he or she aims to achieve with data mining . This knowledge can then be converted into a data mining problem, and a preliminary plan designed to

achieve the objective can be developed (SHEARER, 2000, OLSON and DELEN, 2008).

2. **Data Understanding:** Once the DM objectives are defined, the analyst needs to be familiar with the data in order to: be able to clearly describe the problem, identify the relevant data for the problem, and make sure that the variables relevant to the problem are not interdependent (SHEARER, 2000, OLSON and DELEN, 2008).

3. **Data Preparation:** The main goal of this phase is to adjust the data for the modeling phase. Since the data has several sources, it is common that they are not prepared for the DM methods to be applied directly to them. Depending on the quality of this data, some actions are necessary, including: table, record, and attribute selection, as well as transformation and cleaning of data (SHEARER, 2000, OLSON and DELEN, 2008).

4. **Modeling:** It is at this stage that the mining algorithms will be applied. To create the mining model, the algorithm analyzes the data set and searches for patterns and trends. Thus, the algorithm uses the results of this analysis to define the mining parameters. These parameters are then applied to the complete set of data for extra and statistical patterns. In this phase, various modeling techniques can be selected and applied according to the proposed objective (MCCUE, 2007). For this, their parameters are calibrated to optimal values (SHEARER, 2000).

5. **Evaluation:** This is a review phase that aims to ensure the effectiveness of the model in accordance with business objectives. For this reason, this step is critical to determine if some important business issue has not been sufficiently considered (SHEARER, 2000). Several graphical tools are used for the visualization and analysis of the results. Tests and validations should be performed (cross validation, set of tests supplemented, use training set, percentage division) and indicators to help the results obtained (confusion matrix, correction index and inaccuracy) of mined instances, kappa statistics, absolute mean error, mean relative error, precision, F-measure and others) (OLSON and DELEN, 2008).

6. **Deployment:** In this final phase of the CRISP-DM process the results will be evaluated and subsequently used to determine a strategy for deployment. If a general procedure has been identified to create the relevant models, this procedure is documented here for later deployment. In this point, it is worth remembering that it makes sense to consider the ways and means of deployment during the business understanding phase as deployment is crucial to the success of the project. This is where predictive analytics really helps to improve the operation side of your business.

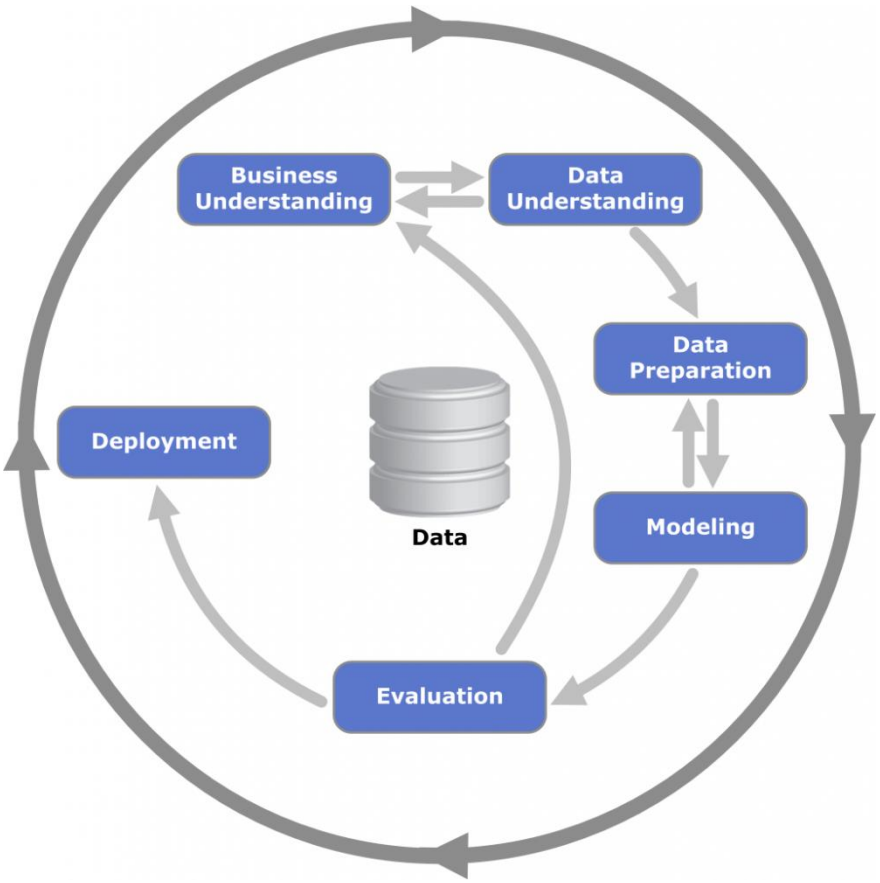

Figure 2 - Representation of the CRISP-DM process (Adapted by SHEARER, 2000).

Table 1 - Tasks and Outputs of the CRISP-DM Reference Model (Adapted by SHEARER, 2000).

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives • Background • Business Objectives • Business Success Criteria	**Collect Initial Data** • Initial Data Collection Report	**Data Set** • Data Set Description	**Select Modeling Technique** • Modeling Technique • Modeling Assumptions	**Evaluate Results** • Assessment of DM Results w.r.t. Business Success Criteria • Approved Models	**Plan Deployment** • Deployment Plan
Access Situation • Inventory of Resources • Requirements, Assumptions and Constraints • Risks and Contingencies • Terminology • Costs and Benefits	**Describe Data** • Data Description Report	**Select Data** • Rationale for Inclusion/Exclusion	**Generate Test Design** • Test Design	**Review Process** • Review of Process	**Plan Monitoring and Maintenance** • Monitoring and Maintenance Plan
Determine DM Goals • DM Goals • DM Success Criteria	**Explore Data** • Data Exploration Report	**Clean Data** • Data Cleaning Report	**Build Model** • Parameter Settings • Models • Model Description	**Determine Next Steps** • List of Possible Actions • Decision	**Produce Final Report** • Final Report • Final Presentation
Produce Project Plan • Project Plan • Initial Assessment of Tools and Techniques	**Verify Data Quality** • Data Quality Report	**Construct Data** • Derived Attributes • Generated Records	**Assess Model** • Model Assessment • Revised Parameter Settings		**Review Project** • Experience Documentation
		Integrate Data • Merged Data			
		Format Data • Reformatted Data			

3.1.2 Learning Techniques and Tasks

Traditionally DM can be divided into supervised (predictive) and unsupervised (descriptive) learning techniques as shown in Figure 2. The main difference between the two methods lies in the need to use a target attribute. While supervised techniques are based on a pre-categorization to classify new records, instances or rows, the unsupervised techniques use an algorithm capable of extracting the characteristics of the data provided by grouping them into specific classes (CIOS, 2007, FAYYAD, 1996, HAN, 2006). For this, we use similarity measures among the attributes (MCCUE, 2007).

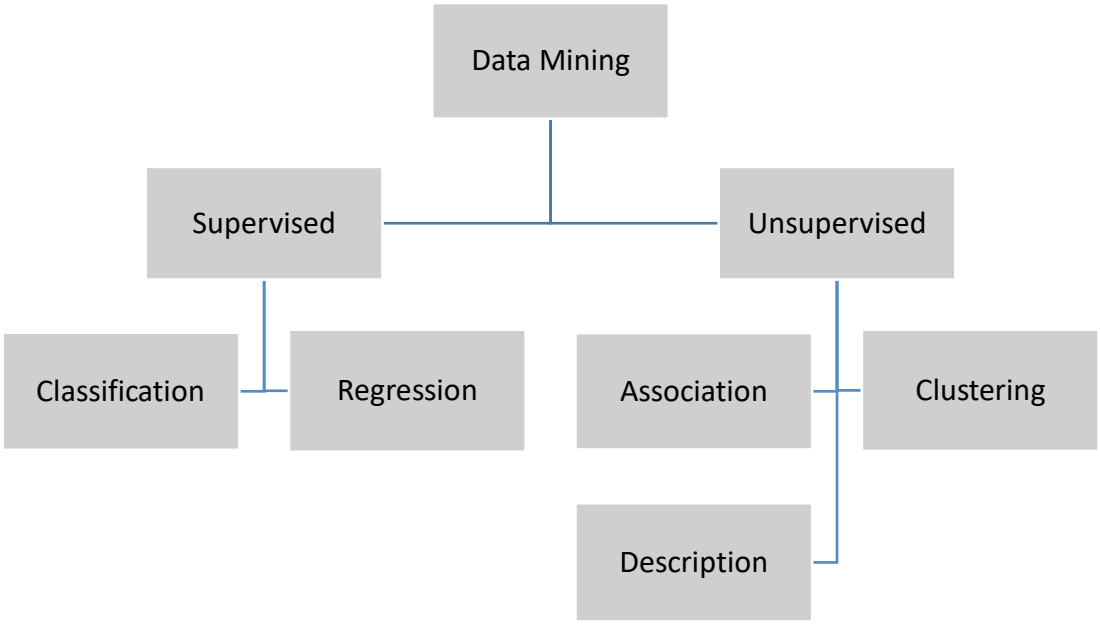

Figure 3 - Taxonomy of data mining tasks (Adapted by BECKMAN, 2017).

The DM can be classified, secondarily, according to their ability to perform certain tasks (Figure 3), corresponding to different objectives for the person who is analyzing the data (LAROSE, 2005). The categorization below captures the main types of data mining tasks and previews the major types of data mining algorithms we will describe later in this thesis.

- **Description Modeling:** The goal of this model is to describe all patterns and trends revealed the data (or the process that generates it). The description task is

widely used in conjunction with exploratory data analysis techniques to prove the influence of certain variables on the result obtained. Examples of such descriptions include: “models for the overall probability distribution of the data (density estimation), partitioning of the p-dimensional space into groups (cluster analysis and segmentation), and models describing the relationship between variables (dependency modeling)” (HAND, 2001, LAROSE, 2005). In HAND (2001), a comparative analysis is made on the various ways in which descriptive modeling has been used in recent years:

- “Segmentation has been extensively and successfully used in marketing to divide customers into homogeneous groups based on purchasing patterns and demographic data such as age, income, and so forth (WEDEL and KAMAKURA, 1998).

- Cluster analysis has been used widely in psychiatric research to construct taxonomies of psychiatric illness. For example, EVERITT, GOURLAY AND KENDELL (1971) applied such methods to samples of psychiatric inpatients; they reported (among other findings) that "all four analyses produced a cluster composed mainly of patients with psychotic depression."

- Clustering techniques have been used to analyze the long-term climate variability in the upper atmosphere of the Earth's Northern hemisphere. This variability is dominated by three recurring spatial pressure patterns (clusters) identified from data recorded daily since 1948”.

- **Classification Modeling:** The classification task aims to identify which class a given record belongs to. The classes are subset of features already known that can be used to categorize the records. The model is built based on a sample containing records from the given classes. The algorithm analyzes a set of records containing an indication of the subject class and then creates the classification rules that will be used later to predict the unknown record. Classification models can be applied for a number of purposes, such as: determining when a credit card transaction can be a fraud; identifying which classes are best for a student; diagnosing diseases; identifying when a person can be a security threat. In practice, we can cite as examples:

  - “The SKICAT system of FAYYAD, DJORGOVSKI, and WEIR (1996) used a tree-structured representation to learn a

classification tree that can perform as well as human experts in classifying stars and galaxies from a 40- dimensional feature vector. The system is in routine use for automatically cataloging millions of stars and galaxies from digital images of the sky.

- Researchers at AT&T developed a system that tracks the characteristics of all 350 million unique telephone numbers in the United States (CORTES & PREGIBON, 1998). Regression techniques are used to build models that estimate the probability that a telephone number is located at a business or a residence" (HAND, 2001).

- **Regression Modeling:** Regression modeling is similar to the classification previously described, however, there is not a categorical class to predict, but the record value (LAROSE, 2005). Thus, it estimates the value of a given record by analyzing the values of the others. One example is to use this approach to predict the value of the Dow Jones Index, instead of a positive surge class or any other identified classes.

- **Association Modeling (Discovering Patterns and Rules):** This type of task is related to the construction of models (HAND, 2001). For an association model, predictions typically are based on rules, and can be used to make recommendations, whereas queries on content typically explore the relationship among itemset. Generally, each association model has a simple structure containing only one parent node that represents the model and its metadata. Each parent node consists of a simple list of sets of items and rules, respectively ordered. Each rule contains a node that, in turn, includes the definition of the itemset, the number of cases that contain this itemset, and other information (MICROSOFT, 2018). Examples of data mining systems of pattern and rule discovery include the following:

    - "Professional basketball games in the United States are routinely annotated to provide a detailed log of every game, including time-stamped records of who took a particular type of shot, who scored, who passed to whom, and so on. The Advanced Scout system of BHANDARI *et al.* (1997) searches for rule-like patterns from these logs to uncover interesting pieces of information which might otherwise go unnoticed by professional coaches (e.g., "When Player X is on the floor, Player Y's shot accuracy decreases from 75% to 30%.") As of 1997 the system was in use by several professional U.S. basketball teams.

- Fraudulent use of cellular telephones is estimated to cost the telephone industry several hundred million dollars per year in the United States. FAWCETT and PROVOST (1997) described the application of rule-learning algorithms to discover characteristics of fraudulent behavior from a large database of customer transactions. The resulting system was reported to be more accurate than existing hand-crafted methods of fraud detection" (HAND, 2001).

- **Clustering:** The clustering task aims to identify and approximate similar records. A clustering consists of a collection of records similar to each other, but different from the other records in the other cluster. Unlike classification, this task does not require records to be previously categorized (unsupervised learning). In addition, it does not pretend to classify, estimate or predict the value of a variable, it only identifies similar data sets, as shown in figure 4 (LAROSE, 2005). Clustering applications can be found and many different areas as follow: image processing, market research and segmentation, geographic research, data analysis, pattern recognition, plant and animal taxonomy, Web document classification and detection of atypical behavior (OLIVEIRA, 2008).

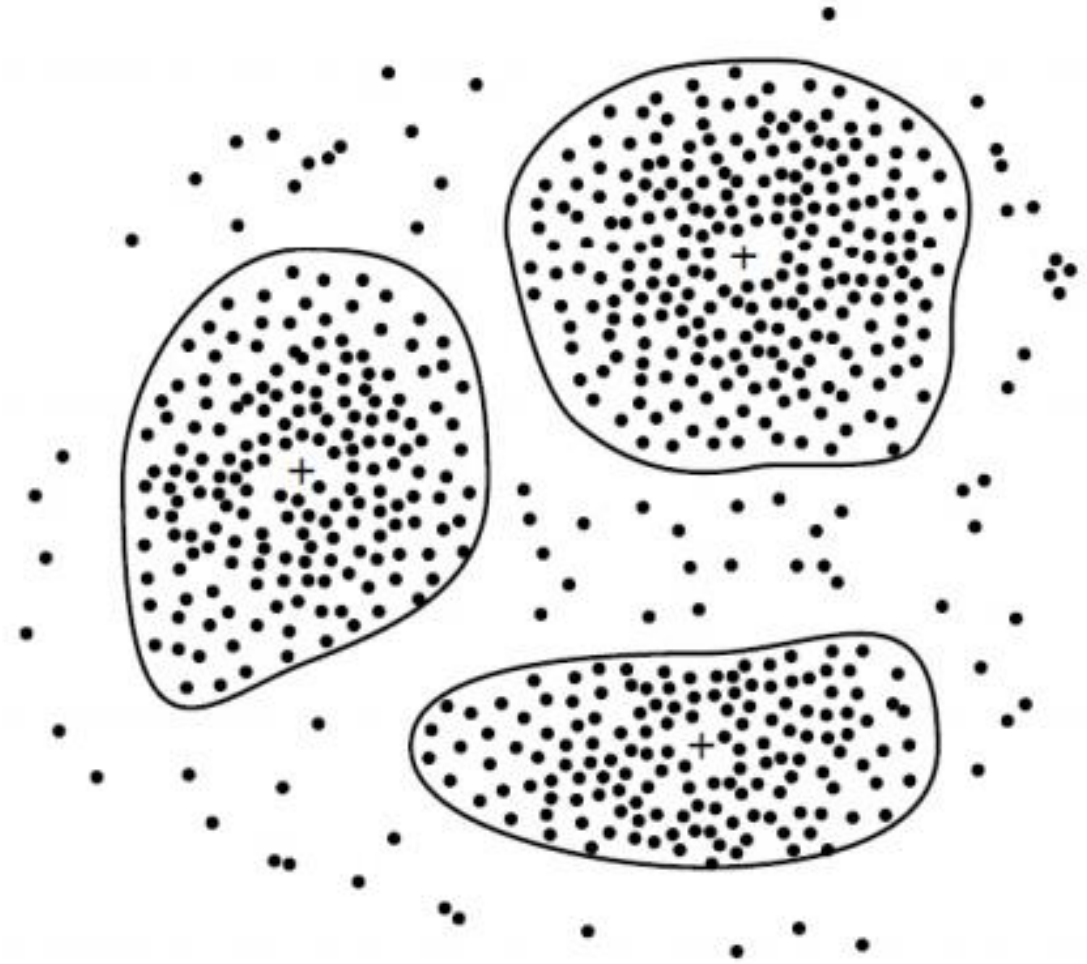

Figure 4 - Records grouped into three clusters (HAN, 2006).

- **Deep learning Modeling:** Deep Learning (also known as deep structured learning, hierarchical learning, or deep machine learning) is a machine learning branch that allows computational models composed of multiple layers of learning to learn data representations with multiple levels of abstraction (e.g. vector values of intensity per pixel, or a set of edges). This process has dramatically improved state-of-the-art speech recognition, object detection, recognition of visual objects, and even drug discovery and genomics. In general, deep learning discovers complex structures in large data sets using the backpropagation algorithm to indicate how a machine must change its internal parameters that are used to calculate the

> representation in each layer of the representation in the previous layer. (LECUN, *et al*., 2015).

Briefly, in unsupervised learning, the classes are unknown. It is based on exploratory tasks looking for unknown patterns, groups, and similarities (BECKMAN, 2017). This work focuses on supervised learning and classification tasks with emphasis in two algorithms: SVM (Support Vector Machines) and KNN (K Nearest Neighbors). In the next sections, the classification algorithms, evaluation measures, and text mining techniques will be described.

3.2 Classification Algorithms

As we saw in the previous section, classification is a modeling activity that uses machine learning algorithms, and it is considered a supervised learning task, because each tuple in a dataset must be labeled according to its features. The tuples are randomly sampled from the database under analysis and, in the context of classification, can be referred to as samples, examples, instances, data points, or objects.

Basically, classification can be divided into two steps/can be described as a two-step process, consisting of a learning step (where a classification algorithm is used on a dataset) and a testing or classification step (where the model is used to predict class labels for a given data) (HAN, 2006). For this task, there are multiple learning techniques, and according to CAMILO and SILVA (2009), they are mainly categorized as Decision Trees, Bayesian Classification, Rule-Based Classification, Neural Networks, Support Vector Machines (SVM), Lazy Algorithms (e.g., KNN), Genetic Algorithm, Fuzzy Set and Rought Set.

In the first step (Figure 5a), a classifier is built describing a dataset, entitled training set. A training set is made up of database tuples and their respective pre-defined class labels. Since the class label of each training tuple is pre-defined, this classification step is also known as a supervised learning activity (HAN, 2006, BECKMAN, 2017). The

training algorithm will generate a predictive model based on the relationship between the attribute values and the class the instance belongs to, that is, we can say that first step of the classification process can also be viewed as: “the learning of a mapping or function, y = f (X), that can predict the associated class label y of a given tuple X” (HAN, 2006).

In the second step (Figure 5b), the model is applied to the test data. A good performance is achieved when the capacity of the classification function matches the size of the training set (BOSER *et al.*., 1992). First, the predictive accuracy of the classifier is estimated using a test set, made up of test tuples and their associated class labels. They are independent of the training tuples, meaning that they were not used to construct the classifier. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier (HAN, 2006). If the accuracy is considered acceptable, new tuples of unknown class will be identified, using the predictive model generated in the training phase to decide which class the new tuple belongs to, thus completing the process of machine learning and classification (HAN, 2006).

At this stage, it is necessary to compute measures that can assert the quality of the predictive model obtained during the training phase (BECKMAN, 2017). These measures are known as classification measures and they will be described in section 3.3.

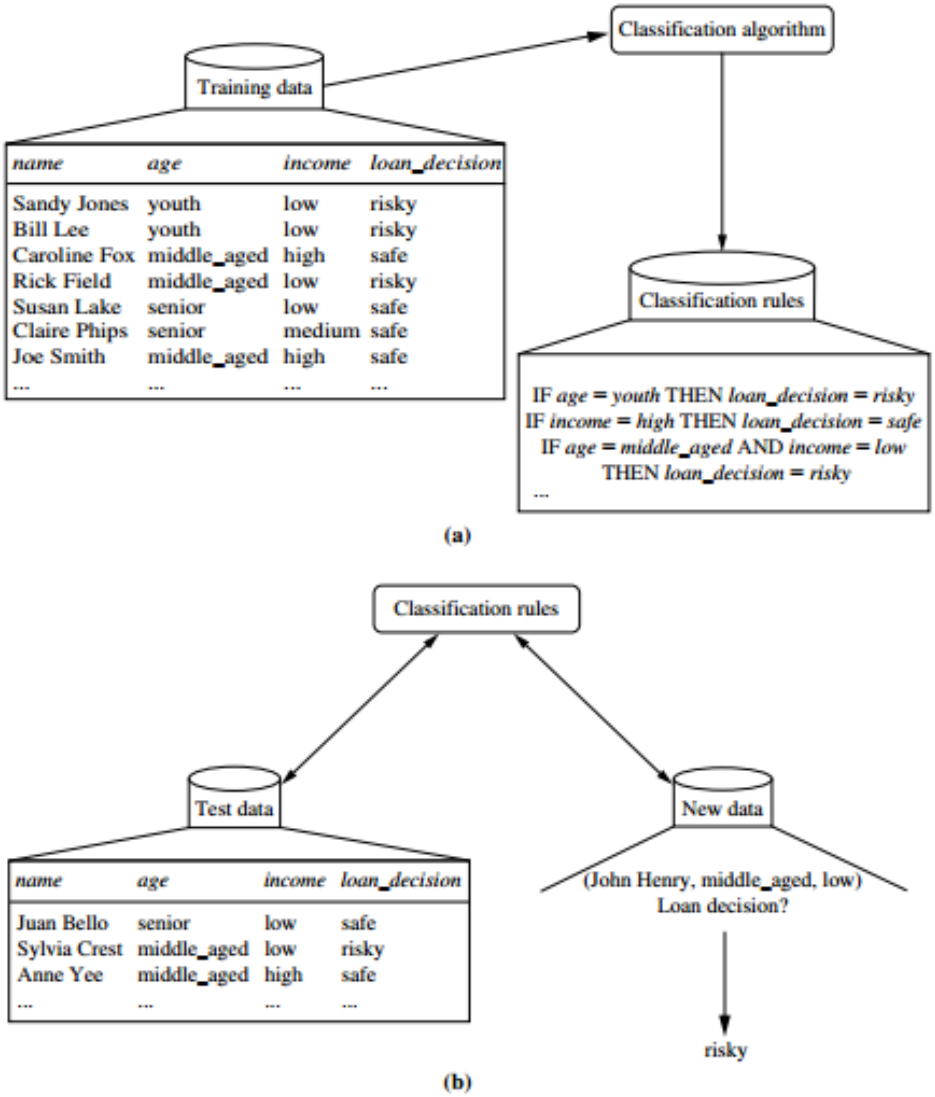

Figure 5 - Hypothetical model of the data classification process. (a)Learning: Training data are analyzed by a classification algorithm. Here, the class label attribute is loan decision, and the learned model or classifier is represented in the form of classification rules. (b)Testing: Test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples (HAN, 2006).

The training and testing process is also known as predictive model selection. The central concept of cross-validation techniques is the partitioning of the data set into mutually exclusive subsets. The simplest technique consists in the separation of a portion (normally 70%) of the dataset for training, and the remaining for test (BECKMAN, 2017).

(1)

During the testing process of DM, several techniques must be tested and combined in order to obtain the best result (MCCUE, 2007). Figure 6 shows an example of how the combination of these techniques occurs. Nowadays, there is a great number of classification methods and several variations of them. The objective of this section is to describe the classification methods used in this work.

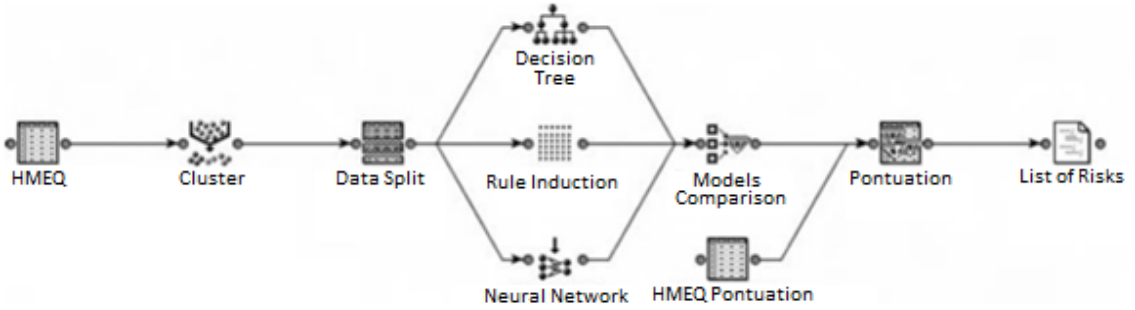

Figure 6 – Example of combination of data mining techniques described by MCCUE, 2007.

3.2.1 Support Vector Machine

The Support Vector Machine algorithm (SVM) is a supervised learning technique applicable for classification and regression tasks, whose aim is to find the ideal separation hyperplane which maximizes the margins of the training base and decision boundary. This technique was launched by VAPNIK & LERNER in the 1960s; however it was only in 1992 that a first article was presented by BOSER *et al*. Although it is a new technique, it has drawn a lot of attention for its results. According WU *et al*. (2007, p.10):

> "In today's machine learning applications, support vector machines (SVM) are considered a must try—it offers one of the most robust and accurate methods among all well-known algorithms. It has a sound theoretical foundation, requires only a dozen examples for training, and is insensitive to the number of dimensions. In addition, efficient methods for training SVM are also being developed at a fast pace".

Different from other methods, the SVM searches for the best classification function to distinguish between members of the two classes in the training data using a simple

approach that does not exceed a predetermined level of error, besides minimizing the structural risk (WU *et al.*., 2007

For a linearly separable dataset, a linear classification function is adopted that is a separating hyperplane *d* (x) (4) that passes through the middle of the two classes (Figure 7). This separation is given by a training set, where y (t) can be 1 or -1 (5). Once this function is determined, new data instance x_n can be classified by simply testing the sign of the function *d* (x_n);$x_n$belongs to the positive class if function *d* (x_n) > 0 (WU *et al.*., 2007).

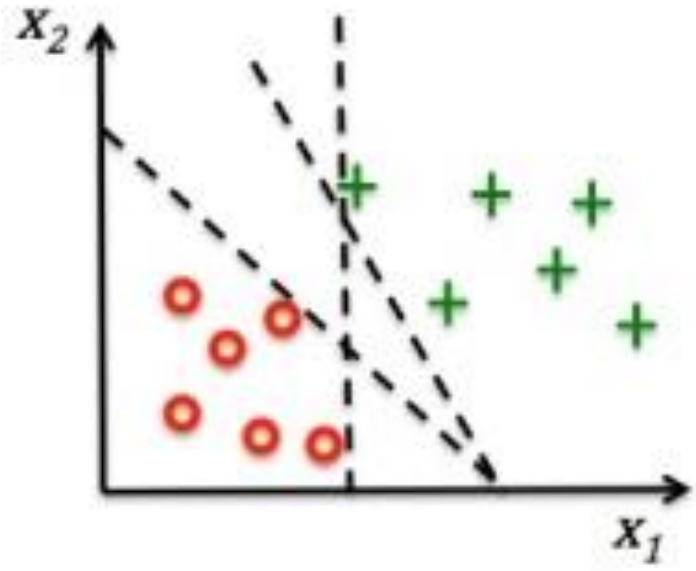

Figure 7 - Infinite separation surfaces in a binary classification problem (HEARTY *et al*., 2016).

In order to ensure that the best such function is found, the metric used to conceptualize the classification function is performed geometrically, allowing the hyperplane margins to be maximized. Intuitively, the margin can be defined as the amount of space or separation existing between the two classes. This corresponds, geometrically, to the shortest distance between the closest data points to a point on the hyperplane. Thus, although there are an infinite number of hyperplanes, only the few which stand as far away as possible from the points of each category will qualify as the solution for SVM (BECKMAN, 2017, WU *et al.*., 2007).

This selection allows not only the best classification performance (e.g., accuracy) on the training data, but also leaves much room for the correct classification of the future data; that is, it offers a better generalization capacity to the model (WU *et al.*., 2007).

The algorithm seen so far was developed for linear classification problems, this being an important limitation of the technique. To solve this limitation, BOSER *et al.*. (1992) suggested a way to make a nonlinear SVM classifier (Figure 10) using the kernel trick (AIZERMAN, *et al.*., 1964), thus allowing the modeling of complex nonlinear situations generating simple interpretation models.

By making use of a nonlinear Kernel function, it allows the algorithm to adjust the hyperplane with a maximized margin in a transformed space of infinite dimensionality (BECKMAN, 2017).

In a problem with no linear separability between the classes, as shown in the 2-dimension plot from Figure 8, maybe it is possible to separate green and red points with another plane (a hyperplane) in another dimension.

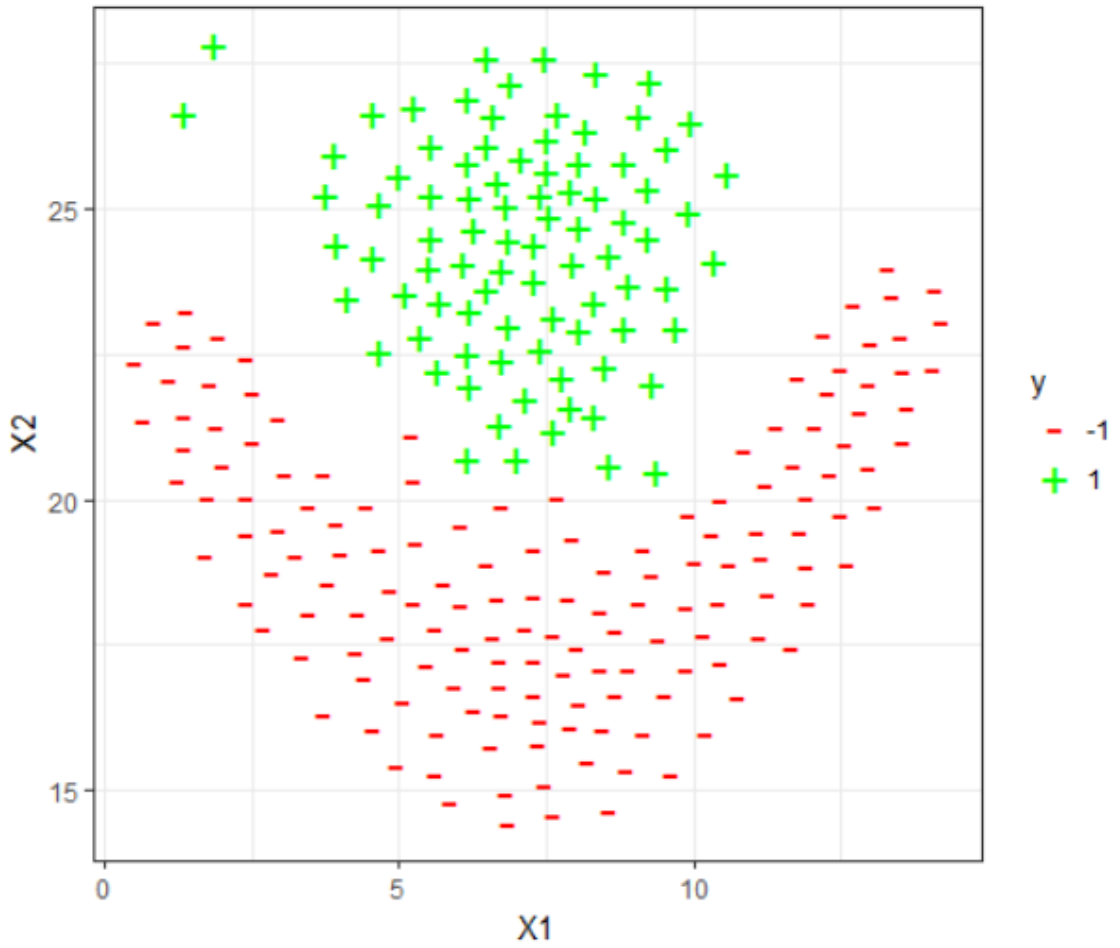

Figure 8 - A two-class dataset with non-linear separation (SANDIPAN, 2018).

Another problem that prevented the acceptance and immediate application of the SVM in the first decades, was the fact that the algorithm was initially designed to deal with completely separable classes. Thinking about it, years later, CORTES & VAPNIK, (1995) introduced to the technique a relaxing constraint variable (9), allowing hyperplanes with flexible margins and finally making the SVM a viable and successful algorithm.

The right choice of the SVM parameters is considered its weakness, since the wrong set of parameters makes the algorithm perform poorly. The solution for this problem can be an optimization procedure to find the best (I think) set of values according to the dataset under study. HSU *et al*. (2003) provides a practical guide to SVM and proposes the use of a grid search to find the best parameters. The grid search

consists of using exponentially growing sequences of each parameter. The settings which achieved the best accuracy after a cross-validation will be picked.

3.2.2 K Nearest Neighbors

Most classification techniques, such as SVM, use the set of training data to learn how to classify a new record. Thus, when they are submitted to a new record they are already ready, that is, they have already learned. There is, however, another category of methods, which only perform this learning when asked to classify a new record. In this case, learning is considered late. Although they require less training time, these methods demand more computational power, since they require techniques to store and retrieve training data. On the other hand, these methods allow incremental learning. An example of this classification model is the algorithm known as kNN (k - Nearest Neighbor) (FIX & HODGES, 1951).

The kNN algorithm was described in the 1950s by FIX & HODGES (1951) and, despite its simplicity, is considered one of the top 10 data mining algorithms (WU *et al.*, 2007). According BECKMAN (2017) the kNN algorithm "creates a decision surface that adapts to the shape of the data distribution, making possible to obtain good accuracy rates when the training set is large or representative". Basically, this algorithm stores the training data and, when a new object is submitted for classification, it identifies a group of k objects in the training set as close to test objects as possible and then assign the label to the closest class in this neighborhood (WU et al., 2007).

According WU *et al.* (2007) there are three key elements of this approach: "a set of labeled objects, e.g., a set of stored records, a distance or similarity metric to compute distance between objects, and the value of *k*, the number of nearest neighbors".

A small value of k means that noise will have a higher influence on the result. A large value makes it computationally expensive and defeats the basic philosophy behind KNN: points that are close might have similar densities or classes. Typically, in the literature odd values are found for *k*, normally with $k = 5$ or $k = 7$, and (DASARATHY,

1991) reports $k = 3$ allowing one to obtain a performance very close to the Bayesian classifier in large datasets.

According MEWADA & PATIL (2011), another challenge is the approach to combining the class label, and the simplest way to solve this problem is to take the majority vote. However, this can be a problem if the distance between the nearest neighbors varies widely and the more reliable neighbors are the class of the object. Thus, an approach that can be taken to solve this problem is to weight each object's vote by its distance. This approach is usually much less sensitive to the choice of k.

For MEWADA & PATIL (2011): "although various measures can be used to compute the distance between two points, the most desirable distance measure is one for which a smaller distance between two objects implies a greater likelihood of having the same class". Some distance measures can also be affected by the high dimensionality of the data. Although the Euclidean distance measure is considered to be the least discriminatory, the algorithm may use other distance metrics besides Euclidean (SIDOROV *et al.*, 2014b, ARGENTINI & BLANZIERI, 2010, BORIAH *et al.*, 2007, WILSON & MARTINES, 1997).

3.2.3 Word2Vec

The work of translating a text does not consist only of translating one word into another, but rather dealing with the morphology, syntax and semantics of two distinct languages to make a connection between them. Thus, although dating from the seventeenth century (HUTCHINS, 1995), the idea of having machines capable of automatically translating a text became a reality in the 20th century. However, although in force, automatic translators still can't do as good a job as a professional translator.

In general, automatic translation systems are developed in order to use information only at the phrase level, ignoring the context of the document, contrary to HARRIS (1954) contention hypothesis. Usually these systems use n-gram models to represent a word in the target language, however, other forms of representation can be used for this purpose, such as vectors. The conception of distributed representation is

not recent (RADOVANOVIĆ *et al*., 2010), and there are models that use it, e.g. word2vec.

Created by MIKOLOV et al. (2013a), word2vec is an open-source tool used to calculate word representations as vectors that offers two architectures as an option: continuous bag-of-words (CBOW) and skip-gram. In practice, skip-gram offers a better representation of words when the database is smaller. CBOW is faster and can be applied in large databases (MIKOLOV et al., 2013b). The tool works by taking a corpus as input and returning word vectors as output. In the process a vocabulary is constructed, and the learning of vocabulary representations as vectors occurs. Figure 9 presents a simplified model of how architectures work. While the CBOW predicts a context-based, skip-gram predicts the targets around a given word (MIKOLOV et al., 2013c).

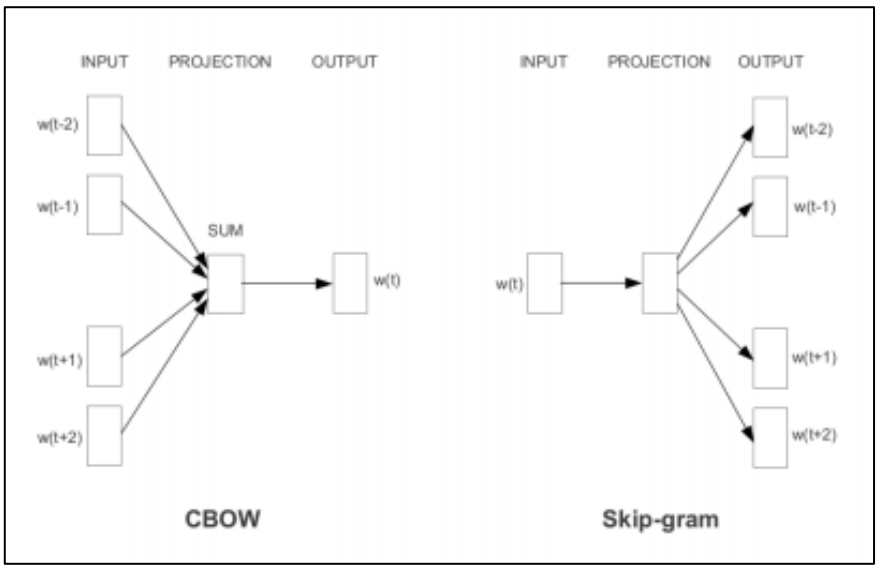

Figure 9 - Representation of CBOW and Skip-Gram architecture (MIKOLOV *et al*., 2013c).

The word2vec architecture proposed by MIKOLOV *et al*. (2013b) has a bilingual extension, whose base is composed of dictionaries and sentence tables. For this, MIKOLOV *et al*. (2013b) generated and used a bilingual dictionary for training and testing in order to solve the optimization problem.

(14)

In order to build this model, it is necessary to construct monolingual models using a large number of texts, then a bilingual dictionary is used to learn the projection between languages. Treating each language as a vector space, one can capture the relationship between them with/using a mapping. This method is based on the premise that words having the same meanings, in different languages, have a similar geometric arrangement, i.e., sharing similar concepts, e.g. the cat is a smaller animal than a dog and how high is the opposite from below.

3.3 Evaluation Measures

In all supervised learning, it is common to use some measure to evaluate the results obtained with a classifier algorithm in terms of the error rate. The classifier predicts the class of each instance; if it is correct, it is counted as "success", if not, it is an "error". The error rate corresponds to the frequency of errors made over the whole set of instances and it determines the overall performance of the classifier. One of the most common techniques used to measure this performance is the Confusion Matrix (OLSON and DELEN, 2008).

From the previous section, a binary classification model (e.g. SVM) classifies each instance into one of two classes: “true” or “false”. Four types of classification are generated, as following: true positive (a.k.a. *hits*), true negative (a.k.a. *correct rejection*), false positive (a.k.a. *false alarms*) and false negative (a.k.a. *misses*) (SOKOLOVA & LAPALME, 2009). These four classes are represented as the confusion matrix (also known contingency table) as shown in Figure 10.

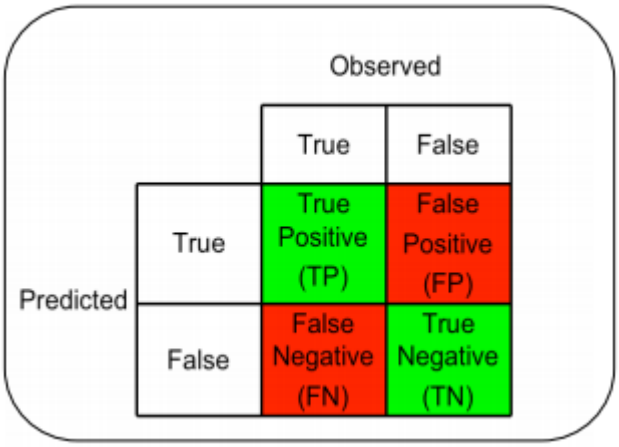

Figure 8 - Confusion matrix format. The true classifications, both positive and negative, that lie along the major diagonal (green one) are the correct classifications. The other fields (in red) state error in model. A perfect model is configured by only the true positive and true negative fields filled out, the other fields would be set to zero (OLSON and DELEN, 2008).

Basically, the confusion matrix juxtaposes the observed classifications for a phenomenon (columns) with the predicted classifications of a model (rows), providing not only the count of errors and hits, but also the necessary variables to calculate other measures. According to SOKOLOVA & LAPALME (2009), the evaluation metrics based on the values of the matrix of confusion commonly used in Text Classification are: precision, recall and F-Measure. All three have formulas that neglect the correct classification of negative examples, reflecting the importance of retrieval of positive examples in text classification:

- Precision: the number of correctly classified positive examples divided by the number of examples labeled by the system as positive.
- Recall: the number of correctly classified positive examples divided by the number of positive examples in the data.
- F-Measure: a combination of the above.

The following sessions will discuss each of the three metrics presented in this session, which will be used to demonstrate the results of the experiments in Chapter 6.

3.3.1 Precision

Precision can be translated as a measure of exactitude that denotes the percentage of occurrences related to all positive objects. This measure is sensitive to class distribution, as the divisor is a sum of positive and negative instances (BECKMAN, 2017). In a classification task, a perfect precision score (ratio = 1.0) for a criterion X, means that each item labeled as belonging to criterion X, in fact, belongs to criterion X. The precision of a classifier is estimated by dividing the number of examples classified as belonging to a class, which actually are of that class (true positives), by the sum between this number and the number of examples classified in this class, but belonging to others (false positives) (OLSON and DELEN, 2008):

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

Where:

- TP: True Positives
- FP: False Positives

3.3.2 Recall

Recall (a.k.a. sensitivity), is a completeness measure, and it denotes the percent of positive objects identified by the classifier. Analyzing the equation (15) and the confusion matrix (Figure 8) together, it is possible to notice a ratio between the true positives and the sum of the elements in the line “positive class”. Because Recall just computes positive instances in its formula, this measure is not sensitive to class distribution (BECKMAN, 2017). The recall of a classifier is estimated by dividing the number of examples classified as belonging to a class, which actually belong to that class, by the total number of examples belonging to this class, even if they are classified

in another class. In the binary case, true positives are divided by total positives (OLSON and DELEN, 2008):

$$\text{Re}\,call = \frac{TP}{TP + FN} \tag{15}$$

Where:

- TP: True Positives
- FN: False Negatives

3.3.3 Accuracy

Accuracy is the overall positive true and positive false classification rate. It computes the weighted arithmetic mean of true positives and true negatives samples as shown below.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

Where:

- TP: True Positives
- FP: False Positives
- TN: True Negatives
- FN: False Negatives

3.3.4 F-Measure

The F-Measure, also known as F-Score, F1-Score, or simply F1, is an evaluation measure that provides a way of combining recall and precision to get a single measure (harmonic mean) which falls between them (VAN RIJSBERGEN, 1979). Recall and precision can have relative weights in the calculation of the F-measure giving it the

flexibility to be used for different applications (CHINCHOR, 1992). The general formula involves a positive real β so that F-score measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision:

$$F - Measure = (1 + \beta^2) \cdot \frac{\mathrm{P.R}}{(\beta^2.\mathrm{P}) + \mathrm{R}} \quad (16)$$

Where:

- P is precision;
- R is recall;
- β is the relative importance given to recall over precision.
- If β= 1, F1 comes to be equivalent to the harmonic mean of P and R. If β > 1, F becomes more recall-oriented and if β < 1, it becomes more precision-oriented, e.g., F0 = P.

The harmonic mean is more intuitive than the arithmetic mean when computing a mean of ratios. A harmonic mean tends strongly towards the smallest elements of a population, having an inclination (if compared to the arithmetic mean) to mitigate the impact of large outliers and aggravate the impact of small ones; that is, the F-measure is higher if the values of recall and precision are more towards the center of the precision. So, for β = 1.0, a system which has recall of 50% and precision of 50% has a higher F-measure than a system which has recall of 20% and precision of 80% (CHINCHOR, 1992).

According to BECKMAN (2017), in terms of classification results, “F-Measure shows lower results, when compared with other measures, denoting that F-Measure tends to be a pessimistic measure". For CHINCHOR (1992) “this behaviour is exactly what we want from a single measure”. On the other hand, as mentioned, the F-Measure, like the Precision and Recall, assumes one class as positive. By default, to compute these measures for both (positive and negative) or more classes, most of the machine learning tools use an average weighted by the number of instances for each class. This can be used for class imbalanced problems to compensate for the disproportion of

instances, but it can result in an F-Measure that is not between Precision and Recall (BECKMAN, 2017).

Throughout the experiments in all this work, the pessimistic behavior of F-Measure proved useful in adjusting the user parameters passed through(?) the algorithms (also known as hyperparameters) along the modeling process. To properly represent the F-Measure for more than one class, the arithmetic mean was used instead of the weighted average.

3.4 Text Mining

Derived from data mining research initiated during the 80's, text mining is considered a set of methodologies for extracting useful information from text content. Currently a myriad of crucial textual information about the financial market is available in different sources on the Internet in unstructured form. However, approaches that deal with unstructured data are hardly used because of the difficulty in extracting the relevant information from them (MITTERMAYER, 2004).

However, with the advent of computer science and programming techniques, there is a growing interest in developing new models that could more reliably explain the behavior of financial series, especially in terms of conditional variance or volatility. From the correct processing and categorization of these data, it is possible to use this rich source of information so that the result is correlated with the financial market and the discovered trend (SCHÜNKE and DIAS, 2013).

According WÜTHRICH (1998), the unstructured data, such as textual statements, contains not only the event (ex: the fall of the Dow Jones Index), but also why it happens (ex: stocks falling due to the dollar weakening and hence a weakening of the treasury of titles). Thus, the exploitation of textual information, especially beyond the numerical time series, increases the quality of the data entry and, therefore, better predictions are expected of this type of input WÜTHRICH (1998). In addition, forecasting techniques that rely on structured information disregard the fact that traders'

expectations are built to some extent from unstructured information (MITTERMAYER, 2004).

Overall, most algorithms used in automatic text categorization (ATC) are derived from data mining applications, but these algorithms are not able to deal directly with unstructured data, as they need a structured format, normally in a matrix shape (WEISS, *et al.*., 2010). A text mining system normally has the architecture as described in Figure 11.

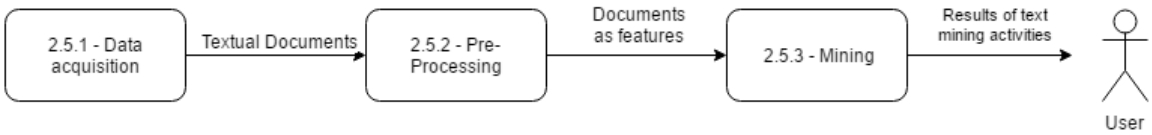

Figure 9 - Text mining system architecture.

3.4.1 Data Acquisition

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

3.4.2 Text Preprocessing

The data analyzed by data mining are numeric, which means they are already in the format required by the algorithms. These algorithms can be applied in ATC, but first it is necessary to convert the content of the documents to a numeric representation. This step is called text preprocessing, and it is often divided into the activities feature extraction, feature selection, and document representation (BRÜCHER *et al.*., 2002).

Feature extraction is the first step in text preprocessing and consists mainly in parsing the document collection. The goal is to generate a dictionary of words and phrases (i.e., features) that describes the document collection adequately. For this, the textual documents must be parsed into simple words, with the blank spaces and punctuation used to distinguish and separate the words. This process is also known as tokenization. A list with all existing words and the respective number of occurrences in the corpus can also be generated during this phase. After this, the words or terms are selected to form features. In this context, a feature can be understood as a value, and the feature name is the meaning of this value. Features can represent a word, a sequence of words or n-grams, which consists in a series of consecutive n words (SIDOROV *et al.*., 2014a), types of entities (e.g., company names, stock symbols), quantitative values (e.g., stock prices, date, time), syntactical structures like noun-phrases and part-of-speech, etc.

The feature candidates are first compared against a list of stop words, and the dictionary is then usually made free of "noise" (e.g., articles, prepositions, and numbers). Terms with occurrence per document lower than or above a specified threshold are also recommended for removal, because a number of words have no representation, and do not carry significant information in the document. The same applies to repeated and abundant words. The min/max thresholds must be adjusted according to the problem under study, but normally values lower than ~5%, or greater than ~90% are reported in the literature.

Furthermore, word stemming techniques can be applied so that features that differ only in the affix (suffix or prefix), i.e., words with the same stem, are treated as single features. Commonly applied word stemming techniques are affix removal, successor variety, n-grams, table lookup, peak & plateau, and Porter's algorithm (BOLLEN, 2003, BAEZA-YATES and RIBEIRO-NETO, 1999). The use of stemming requires caution and must be adjusted according to the problem under study, as it may remove important information existing in the original words.

The most common type of feature representation is the Bag of Words (BOW), first mentioned by (HARRIS, 1954) and still a predominant technique nowadays (MINER, *et al.*., 2014), (ZHAI & MASSUNG, 2016). A BOW is basically a matrix,

where each document is represented as a vector row, and the features (normally words) as the columns of this matrix. The columns of this matrix must contain not only the existing terms in the document, but also all the existing terms in the corpus. Not all the documents share the same terms, then the missing terms in a document are filled with zero or null, which can result in a sparse matrix, as demonstrated in Figure 36.

Feature extraction is followed by feature selection. The main objective of this phase is to eliminate those features that provide few or less important items of information. Indicators commonly used to determine feature importance are term frequency (TF), inverse document frequency (IDF), and their product (TF×IDF). When TF is used, it is assumed that important terms occur in the document collection more often than unimportant ones. The application of IDF presupposes that the rarest terms in the document collection have the highest explanatory power. With the combined procedure TF×IDF the two measures are aggregated into one variable. Whatever metric is used at the end of the feature selection process only the top *n* words with the highest scores are selected as features. While more sophisticated feature selection techniques, such as information gain, Chi-square, correlation coefficient, and relevance score, have been proposed, the above techniques (especially TF) have proved very efficient (SEBASTIANI, 1999).

Document representation is the final task in text preprocessing. At this stage the documents are represented in terms of the features to which the dictionary has been reduced in the preceding steps. Thus, the representation of a document is a feature vector of n elements, where n is the number of features remaining when the selection process is complete. The whole document collection can therefore be seen as an m×n-feature matrix F (with m as the number of documents), where the element fij represents the frequency of occurrence of feature j in document i. Typical frequency measures are, again, TF, IDF, and TF×IDF, but a difference from the previous task is that these frequencies are now measured per document. Sometimes the frequency measure is limited to the values {0, 1}, which indicate whether or not a certain feature appears at all in the document (binary representation). At the end the feature vectors are usually cosine normalized, since some of the ATC classifiers require feature vectors of length 1 (MADANI, 2003).

In recent years, various techniques have been developed to reduce the size of the feature matrix F, which is sometimes enormous. Examples of these techniques are term clustering and latent semantic indexing (DEERWESTER *et al.*., 1990). Major approaches for ATC classifiers involve the use of decision trees, decision rules, k-nearest neighbors, Bayesian approaches, neural networks, regression-based methods, and vector based methods (BRÜCHER *et al.*., 2002; YANG and LIU, 1999).

3.4.3 Mining

Once the data is prepared to be processed, two approaches can be taken to analyze the data: unsupervised learning and supervised learning. In unsupervised learning, techniques can be used to group similar documents, identify common rules, taxonomies and sentiments. These groups are then labeled and used in supervised learning (classification and regression) and recommendation systems (WEISS, *et al.*., 2010), (MINER, *et al.*., 2014), (ZHAI & MASSUNG, 2016).

The most common text mining processes and techniques were briefly described in this section and the previous ones as base of the work presented here. In Chapter 4 other techniques will be described along with a bibliographic review. The methodology applied in this work will be presented in Chapter 5.

CHAPTER 4 – Prediction Models Using TMFP

In this section, a bibliographic review about the developments in state-of-the-art Text Mining applied to Financial Market Prediction (TMFP) will be conducted. In general, the design of TMFP systems follows a structure like Figure 12. The criteria used to select research regarding this subject are that it must have some text mining or Neuro Linguistic Programming (NLP) methodology, and it must predict economic events or changes in some financial instrument. The bibliographic review in this current work aims to use important aspects from that survey.

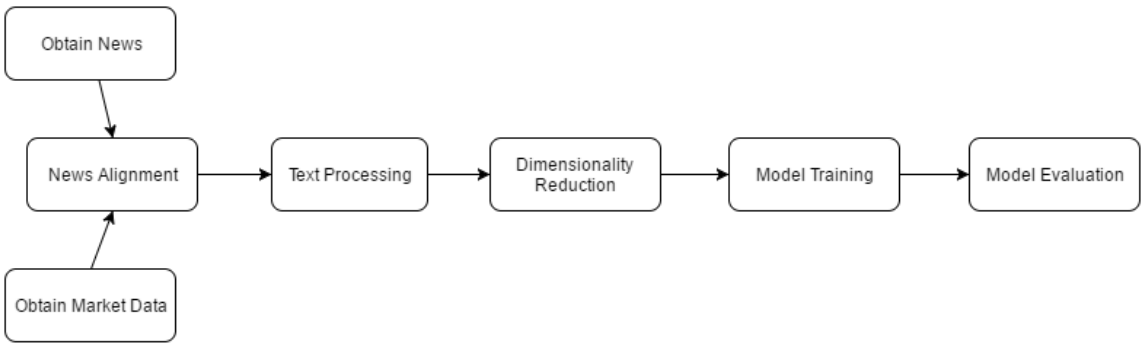

Figure 10 - General design of a TMFP process.

4.1 Model Developed by Thomas & Sycara

This model follows two approaches in making predictions about financial markets, using text data loaded from web bulletin boards. The first uses maximum entropy text classification to make predictions based on the whole body of text; the second uses a genetic algorithm to learn simple rules based solely on numerical data of trading volume, number of messages posted per day, and total number of words posted per day. For the text classification they used the rainbow package developed by McCallum, which provides a variety of potential classification methodologies. Their results proved to be strong with excess return statistically significant by integrating both techniques (Table 2).

Table 2 - Performance by Integrated Approach.

Stock:	Excess Returns	std. dev.
All stocks (22 stocks)	2.88%	4.75
>10K posts (12 stocks)	19.26%	8.84

4.2 Model Developed by Wüthrich *et al.*

The model proposed by WÜTHRICH (1998) (Figure 13) is one of the first research studies published about TMFP and seeks to predict the daily trend of five stock market indices, among them the DJIA from a base of articles published daily in web news portals. In particular, the lead articles appearing in The Wall Street Journal and the Reuters are taken as input. From these articles, the daily closing values of the major stock markets in Asia, Europe and America are predicted. Such a forecast is publicly available at 6:45 p.m. Eastern Time (ET), so that all forecasts generated by the model are available before major Asian markets such as Tokyo, Hong Kong and Singapore start their trading day.

The documents were labeled according to a model of three categories: up, steady, and down. A dictionary consisting of 392 keywords, each considered a typical buzzword capable of influencing the stock market in either direction, was defined by several experts. The Bayesian, Nearest Neighbor, and Neural Network classifiers were trained and categorized all newly published articles overnight.

Each morning, the system scans all pages containing information on financial analysis and movements involving the stock, currency and securities markets around the world. It stores them in five distinct data repositories, containing: day news prediction, stock closing values from the day before the prediction day, past day news, the last hundred stock close values, and hundreds of simple and / or compound keywords. All information obtained is then downloaded by a rule generator coupled to the five data repositories, able to elaborate them and apply them to the prediction. The operation of the prediction model is done as follows:

1. The number of occurrences of keywords in each day's news is counted. The counting of keyword records is case insensitive; stemming algorithms are applied and the system is not limited to exact matches. For example, if we have a keyword record "stock drop", and a web page contains a phrase "stocks have really dropped", the system does still count this as a match;

2. The occurrences of the keywords are then transformed into weights, so for each day, each keyword is given a different weight. For this, the approach used is based on three components (frequency of terms, document discrimination and normalization) that together, will generate a weight between 0 and 1 for each keyword;

3. Through the weights and values of closing of the actions, probabilistic rules are generated based on articles by the same author (WÜTHRICH, 1995, WÜTHRICH, 1997);

4. The generated rules are applied to the daily news. This predicts whether a specific index, such as the Dow Jones, will rise (change above + 0.5%), decline (below -0.5%) or remain stable (variations below 0.5% of its previous close value);

5. From the previous forecasts generated, the expected closing value is also predicted;

6. The final forecasts were then sent to www.cs.ust.hk/-beat/Predictat 7:45 local time in Hong Kong (6.45 pm ET).

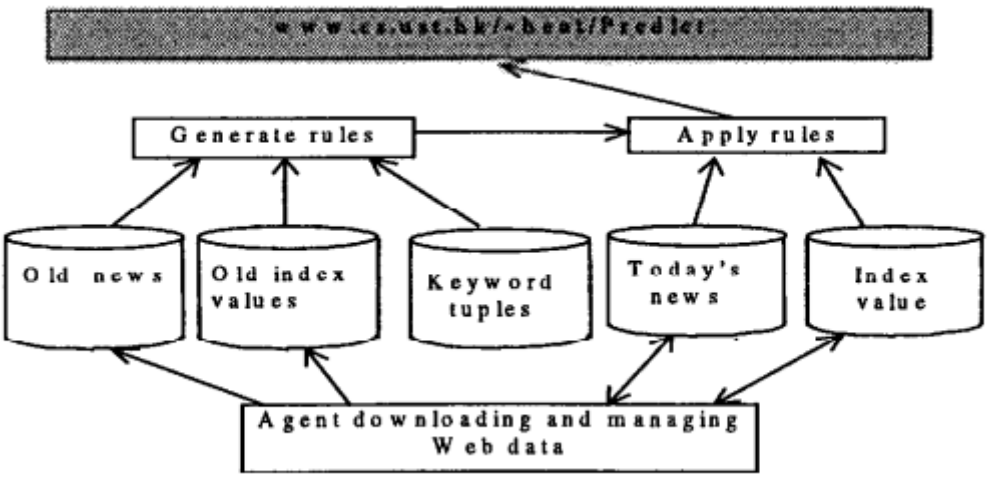

Figure 11 - Architecture of the prediction model of WÜTHRICH (1998).

According to its results, WÜTHRICH (1998) shows an average accuracy of 45% obtained by the model, as shown in Table 3 below. This percentage is significantly better than the precision of a random predictor, which would reach no more than 33% accuracy. These predictions were used for investment simulation, with 7.5% of cumulative return after three months, what can be considered a good result, if compared with the return of 5.1% from DJIA index in the same period.

Table 3 - Accuracy of the model proposed by WÜTHRICH *et al.*. (1998) according to the stock market index, the K-NN algorithm and the neural network.

	Model (%)	K-NN (%)	Neural Network (%)
Dow Jones Index	45	40	36.8
FTSE	46.7	42	35.4
Nikkei	41.7	47	34.1
Hang Seng	45	53	43.9
Strait Index	40	40	32.5

4.3 Model Developed by Lavrenko *et al.*

The model developed by LAVRENKO *et al.*. (2000) is able to predict the behavior of 127 stocks in the US financial market. It identifies the specific news that can influence the market, through the implementation of AEnalyst, a system capable of recommending interesting news stories which affect market behavior. According to Yahoo Finance, AEnalyst has a bank of approximately 38,469 news stories.

The flowchart of the model, illustrated in Figure 14, operates using textual information as input data, correlating the news content identified by the AEnalyst System with the time series trends regarding stock prices. These are identified from piecewise linear fitting and classified according to an automated categorization procedure; the information retrieval techniques are meanwhile used to extract the relevant documents from the news so as to correlate them simultaneously. Next, from the moment in which the correlation is established, language models are generated for each type of trend, capable of predicting the future trend of an action when new news is published.

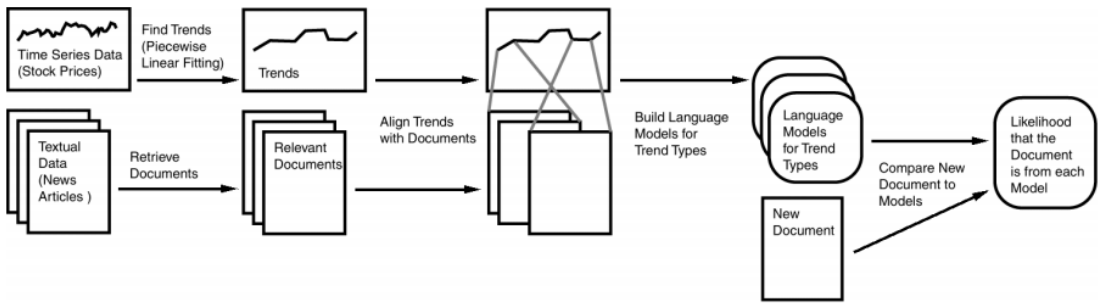

Figure 12 - Flowchart of the model developed by LAVRENKO *et al.* (2000).

LAVRENKO *et al.* (2000) identifies in its work three types of trends: positive, negative and without trend. For the author, a trend is defined as an interval of time in which there is a predominance of the increase, decrease or stability of the price of a stock. Thus, language models provide a framework for classifying texts, where words such as loss, fall and bankruptcy are associated with a downward trend in stock prices,

while words like acquisition and alliance, are associated with a bullish trend in stock prices.

The authors also define five categories of news, through the correlation established between types of trends and language models: Surge (when the news generated a rise greater than + 0.75% in the share price), slight + (a high between +0.5 % and + 0.75%), no recommendation (a volatility between + 0.5% and -0.5%), slight- (when it generated a decrease between -0.5% and -0.75%) and plunge (a decrease greater than -0.75% in share price). Thus, from a classifier trained using the Näyve Bayes approach, at the time a news item is published, the model generates a buy or sell recommendation for a particular action only if the news is classified in the categories “Surge” and “Slight + “Or” Slight- “and” Plunge, “respectively. Also, if the news is classified as “No Recommendation”, the template does not generate recommendations.

As a result, with each new news release on some of the 127 actions defined in the scope, LAVRENKO *et al*. prediction model (2000) generates a recommendation from the collected news bank and thus a share is bought or sold. In this context, AEnalyst is able to generate more than 12,000 buy or sell recommendations, with a profitability of 0.23% on average per recommendation.

However, unlike the model proposed by WÜTHRICH 1998, a priori domain knowledge was not considered here, and in addition, the authors measured the performance of their system by performing only market simulation. Thus, the strategy presented by LAVRENKO *et al*. (2000) led to a lower average trade profit than expected for its commercial policy, whose estimated margin was 1% or more immediately, or to wait for 60 minutes and take a loss if necessary.

The same data were later reused by GIFDOFALVI (2001) to determine, among other things, the improvement of the data maintenance period. According to the findings, the purchases or short sales should generally even out after 20 minutes. However, no market simulation was performed to confirm these results.

4.4 Model Developed by Mittermayer

Based on the assumption that the random walk of stock prices immediately after the publication of a press release is skewed, the model developed by MITTERMAYER (2004) seeks to predict stock market stock price trends from press releases using the NewsCATS (Notices Categorization and Trading System), which automatically analyzes and categorizes press releases and generates stock trading recommendations. According to the authors, the system developed by them differs from their predecessors, especially regarding the way in which the learning examples are chosen and in the determination of the best negotiation strategy.

According to MITTERMAYER (2004), press releases are a good source of information for traders, since they may reveal unexpected information and therefore have a high ability to move stock prices abruptly. Negative press releases, such as bad earnings reports, often cause traders to sell stocks, which translates into a decline in stock prices. By analogy, traders tend to buy stock after positive press releases, such as good earnings reports. This phenomenon, therefore, translates into buying pressure and increases the stock price. In addition, unlike web news, press releases are more reliable, since all communication generated by them follows strict rules of corporate governance and government oversight.

Figure 15 represents the constitution of the model NewsCATS, whose architecture consists basically of three components. The first component is responsible for pre-processing the text and serves as input to the second component which in turn uses the SVM light classifier engine to classify the releases into three categories: "GOOD NEWS", "BAD NEWS" or "NO MOVERS". Finally, the third component is responsible for receiving all information processed and signaling the trend of actions.

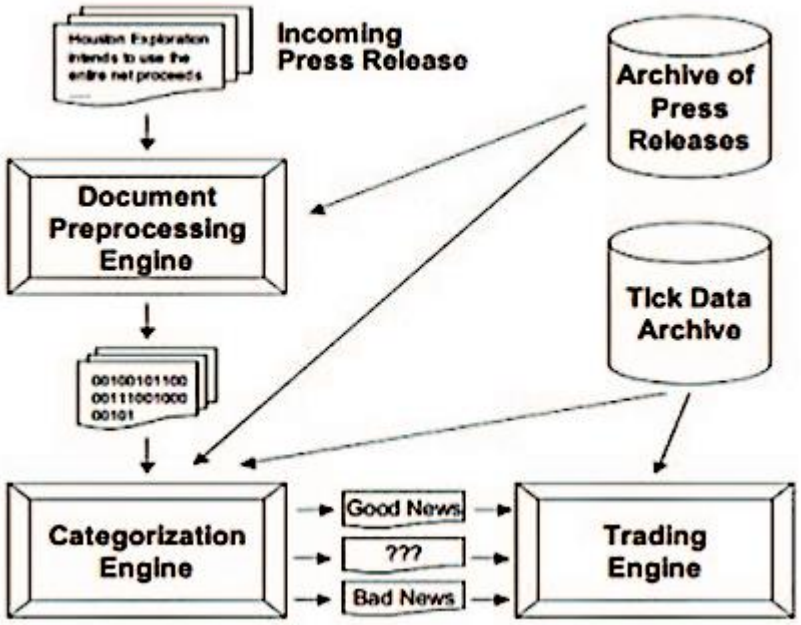

Figure 13 - Architecture of the model of MITTERMAYER (2004) and its components.

NewsCATS is connected to an archive of press releases and to an archive of intraday trades and quotes. With these archives NewsCATS is able to learn a set of categorization rules that allow the Categorization Engine to sort new press releases automatically into a defined number of categories. Each of these categories is associated with a specific impact on the stock prices, e.g., increase or decrease. Depending on the results yielded by the Categorization Engine (i.e., the category assigned to the new press release), the Trading Engine produces trading signals that can be executed via an online broker or other intermediaries.

In his article, MITTERMAYER (2004) presents the results of the operations that were bought with the results of a random system that bought or sold the same actions indicated by the model. Table 4 presents the results of the experiment, noting that on average the model generated more profit than the random system.

Table 4 - Relationship between executed businesses and profitability. L/N=Profit per business.

NewsCATS	Random System
Business Average L/N	Business Average L/N
2.6020.11%	2.5990.00%

The results indicate that NewsCATS can provide trading strategies that significantly outperform a trader randomly buying and shorting stocks immediately after the publication of press releases. However, the results also reveal that there is still much room for improvement. In particular, the output of the Categorization Engine needs to be enhanced. Since the selectivity of the "No Movers" category is good but the selectivity of the two other categories is fairly poor, learning could be improved by inserting a new first step to distinguish between "No Movers" and "Movers" only. In the second step, "Movers" could then be split into "Good News" and "Bad News."

Furthermore, the outcome of the categorization process depends heavily on the feature matrix created by the Document Preprocessing Engine. One possible way of improving the preprocessor is to apply a priori domain knowledge.

4.5 Model Developed by Schumaker & Chen.

The model developed by SCHUMAKER & CHEN (2009) shows how to predict the value of a given action right after the first week of the publication of a piece of news. For this, the authors used a supervised machine learning approach. In this model (Figure 16), three textual analysis techniques were used, responsible for the use of important information terms and their storage in a database. They are: BOW, noun phrases and leaders. In addition, a second database was added to the system for the purpose of storing the stock quote per minute. The model architecture proposed by SCHUMAKER

& CHEN (2009) also uses three distinct textual analysis techniques to perform the preprocessing of the news and an SVR approach to generate the prediction. Three different metrics are applied in the results generated by the model in order to evaluate it: measures of Closeness, Directional Accuracy and a Simulated Trading Engine.

Closeness measure is the comparison between the estimated value given by machine learning against the actual value in a Mean Squared Error (MSE) measure (PAI, 2005). Directional Accuracy compares the value from the direction of the predicted value with actual value (CHO, 1998). Simulated Trading is a simple trading engine that capitalizes on large predicted value differences (LAVRENKO, 2000).

Through this approach, the authors investigated 9,211 financial news articles and 10,259,042 stock quotes covering the S & P 500 stocks during a five weeks period.

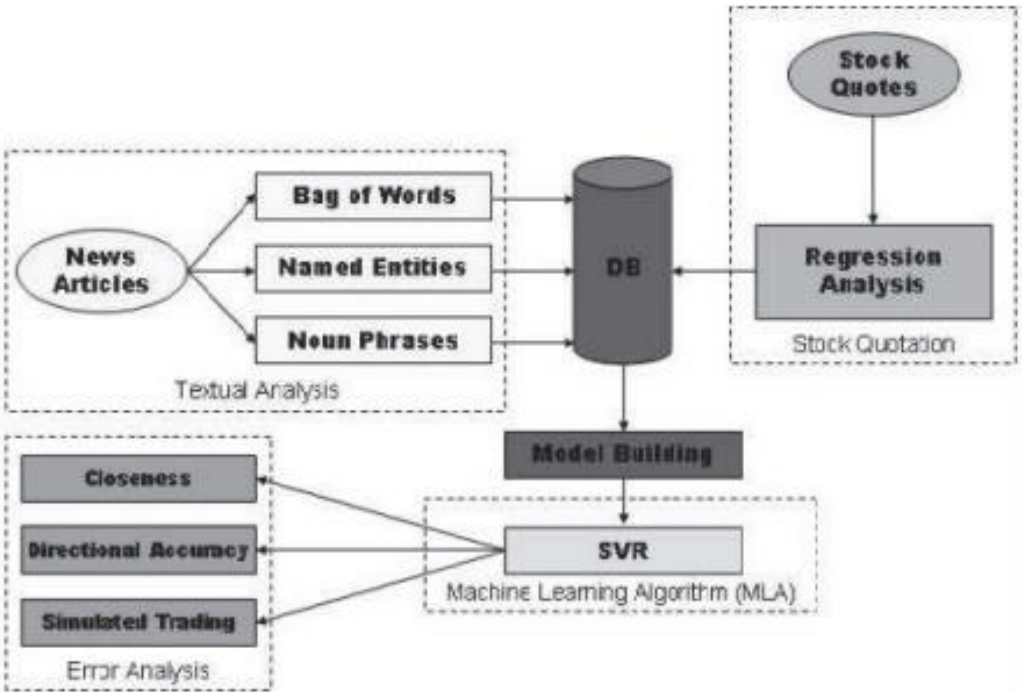

Figure 14 - Architecture of the model of SCHUMAKER & CHEN (2009) and its components.

According to the results generated, the nominal phrases technique obtained the best performance in Directional Accuracy and Simulated Trading Engine, while the technique Named Entities obtained a better performance in Closeness (Table 5).

Table 5 - Metrics used to evaluate the three types of textual analysis together with the respective results. DA = Directional Accuracy; STE = Simulated Trading Engine.

	Closeness	DA	STE
Bag of Words	0.044	55.3%	0.10%
Noun Phrases	0.048	56.5%	0.64%
Named Entities	0.034	55.6%	0.57%

4.6 Model Developed by Soyland

The model developed by SOYLAND (2015) investigates the power of prediction of online news in the face of the daily price variations of the 19 important banks and financial institutions that make up the MSCI World Index. The news date corresponds to news articles, stock exchange information and press releases, which were obtained by a web-crawler, which scanned around 6000 online sources for news and saved them in the database. This news is partitioned and labeled into two classes according to price change class or trade volume class; both were done with a binary document classification approach. An automated document classification model was created for the prediction analysis, and Python programming language was used exclusively for these experiments.

To measure the performance of the text document classification model, the random forest predicted labels for the test set, and precision, recall, accuracy, and F-measure were used. A randomized labeling algorithm was also generated in order to benchmark the model. The randomized model assigned high or low labels to the documents with equal probability. The results obtained in the one-day return experiment and one-day volume experiment are shown in Tables 6 and 7, respectively. The columns in the table correspond to the predicted labels of the test set, and the rows correspond to the true labels.

Table 6 - Confusion matrix for evaluating the performance of the one-day return classifier.

Evaluation metric	Value
Accuracy	51.1%
Precision (high)	59.4%
Precision (low)	40.8%
Recall (high)	55.9%
Recall (low)	44.3%
F-measure (high)	57.6%
F-measure (low)	42.5%

Table 7 - Confusion matrix for evaluating the performance of the one-day volume classifier.

Evaluation metric	Value
Accuracy	78.3%
Precision (high)	78.8%
Precision (low)	77.6%
Recall (high)	82.5%
Recall (low)	73.1%
F-measure (high)	80.6%
F-measure (low)	75.3%

According to the results presented by this model, this one fails in predicting the one-day stock price changes; however, the percentage of correctly labeled documents in the one-day trade volume experiment was 78.3%; i.e. a classification accuracy of 78.3% was achieved, suggesting that online news contains some valuable predictive information.

4.7 Comparative Analysis of Models

From the models presented above, we can readily see that several algorithms were used (Table 8) and that almost all instances commonly classify predicted stock movements in a set of classification categories, not a discrete price forecast. The first technique presented is the Genetic Algorithm. In this study, discussion boards were used as a source of independently generated financial news. In their approach, THOMAS & SYCARA (2002) attempted to classify stock prices using the number of postings and number of words posted about an article on a daily basis. It was found that positive share price movement was correlated to stocks with more than 10,000 posts. However, discussion board postings are quite susceptible to bias and noise.

Table 8 - Prior algorithmic research.

Algorithm	Classification	Source Material	Bibliography
Genetic Algorithm	2 categories	Undisclosed number of chatroom postings	Thomas & Sycara, 2002
Naïve Bayesian	3 categories	Undisclosed number of daily news	Wüthrich *et al.*., 1998
	5 categories	38,469 articles	Lavrenko *et al.*., 2000
	3 categories	Over 5,000 articles borrowed from Lavrenko *et al.*.	Gidofalvi *et al.*., 2001
	3 categories	6,602 articles	Mittermayer, 2004
SVM	3 categories	9,211 articles	Schumaker & Chen, 2009
	2 categories	6,000 online sources for news	Soyland, 2015

Then we come across another machine learning technique, Naïve Bayesian. This technique represents each article as a weighted vector of keywords (WÜTHRICH *et al*. 1998, LAVRENKO *et al*.., 2000, GIFDOFALVI, 2001). Phrase co-occurrence and price directionality is learned from the articles which lead to a trained classification system. One such problem with this style of machine learning is from a company mentioned in passing. An article may focus its attention on some other event and superficially reference a particular security. These types of problems can obfuscate the results of training by unintentionally attaching weight to a casually-mentioned security.

Finally, a third method, known as Support Vector Machines (SVM), is presented in the works of MITTERMAYER (2004), SCHUMAKER & CHEN (2009) and SOYLAND (2015), whose results were more interesting than the first two. MITTERMAYER (2004) used SVM in his research to find an optimal profit trading engine. While relying on a three tier classification system, his research was based on empirically establishing trading limits and, as a result, he found that profits can be maximized by buying or shorting stocks and taking profit on them at 1% up movement or 3% down movement. This method slightly overcame random trading by yielding 0.11% average return.

Guided by MITTERMAYER (2004) and SCHUMAKER & CHEN (2009), the Minimal Sequential Optimization (SMO) method was implemented to solve the problems of scalability of the use of large training sets. Applying these regression based methods and textual representation techniques to a supervised machine learning algorithm such as SVM can lead to a trained system with discrete numeric output. Moreover, unlike what was being done, the evaluation of SCHUMAKER & CHEN (2009) focused on three different metrics that, according affirms to JOACHIMS *et al*.. (1998), avoid the problem of unmanageably large feature spaces. From there, the author showed that the model containing both article terms and stock price at the time of article release had the best performance in closeness to the actual future stock price, the same direction of price movement as the future price (57.1% directional accuracy) and the highest return using a simulated trading engine (2.06% return).

Years later, SOYLAND (2015) intended to use machine-learning based on the predictive power of text data, which is usually more unstructured and fuzzy than the

numerical data. However, the model fails in predicting both experiments created. There are some limitations that may have affected the results. First of all, the web-based news data used in this study is extremely voluminous. Secondly, this model only used statistical and machine-learning tools to filter out the irrelevant or noisy data. Lastly, another fact that could affect the above result is the aggregation of companies, since news events that give rise to a stock price change for one company may not necessarily give rise to a stock price change for another company.

The models presented in this article deal with different economic times, due to the fact that each period corresponds to a unique database. This makes it difficult to compare which model can achieve a better performance in relation to the others, since a model analyzed in a predominantly positive or negative economic period can more easily predict the financial market trend than another model, even though it may perform better in the same period. In addition, according to POON and GRANGER (2003), in a study of 93 articles and working papers published between 1976 and 2002, the diversity of results generated by the prediction models does not yet allow consensus to be reached on their accuracy, forming a broad field for research and debate.

However, as we can see, SVM is one of the algorithms for categorization of texts that have been used in recent years and, therefore, has shown to be one of the most efficient in this area (YANG & LIU, 1999, DUCKER, WU & VAPNIK, 1999, JOACHIMS, 1998, DUMAIS et al., 1998):

- YANG & LIU (1999) made comparisons between 5 algorithms for text categorization: SVM, Nearest Neighbor, Naive Bayes, Neural Network and LLSF. In all the tests performed, using F1 for comparison, we see that SVM Nearest Neighbor has a better efficiency when compared to the other algorithms.

- DUCKER *et al*. (1999) made comparisons between 4 algorithms for categorizing texts for the purpose of identifying spam emails. The algorithms were Ripper, Rocchio, Decision Tree – Boosting and SVM. According to their results, SVM and Decision Tree – Boosting are the best candidates for better efficiency, and for Ripper there is a slightly less error rate, but error dispersion is better for SVM.

- JOACHIMS (1998) did an experiment similar to the previous one, comparing 5 categorization algorithms: Naive Bayes, Rocchio, Nearest Neighbor, C4.5 and SVM. Using all features the best performance was achieved with SVM.

- DUMAIS *et al.* (1998) made comparisons between 5 algorithms: Rocchio, Decision Tree, Naive Bayes, Bayes Networks and SVM. In this study precision and coverage measures were used, and SVM was the best algorithm presented for both.

This chapter presented a summary of the characteristics of the main text mining models present in the literature, for predicting the financial market trend. However, other papers about TMFP have been published over the years (Figure 17, Table 26). There are just a few publications each year, likely due to the difficulty of dealing with large amounts of data and other discouraging results in this area, so any research that can show improvement is an important milestone for future studies. A corpus containing 36 papers published to date will be used as a basis for the next sections (Table 26). The similarities and differences between the presented models were synthesized in order to meet the objectives of the present work.

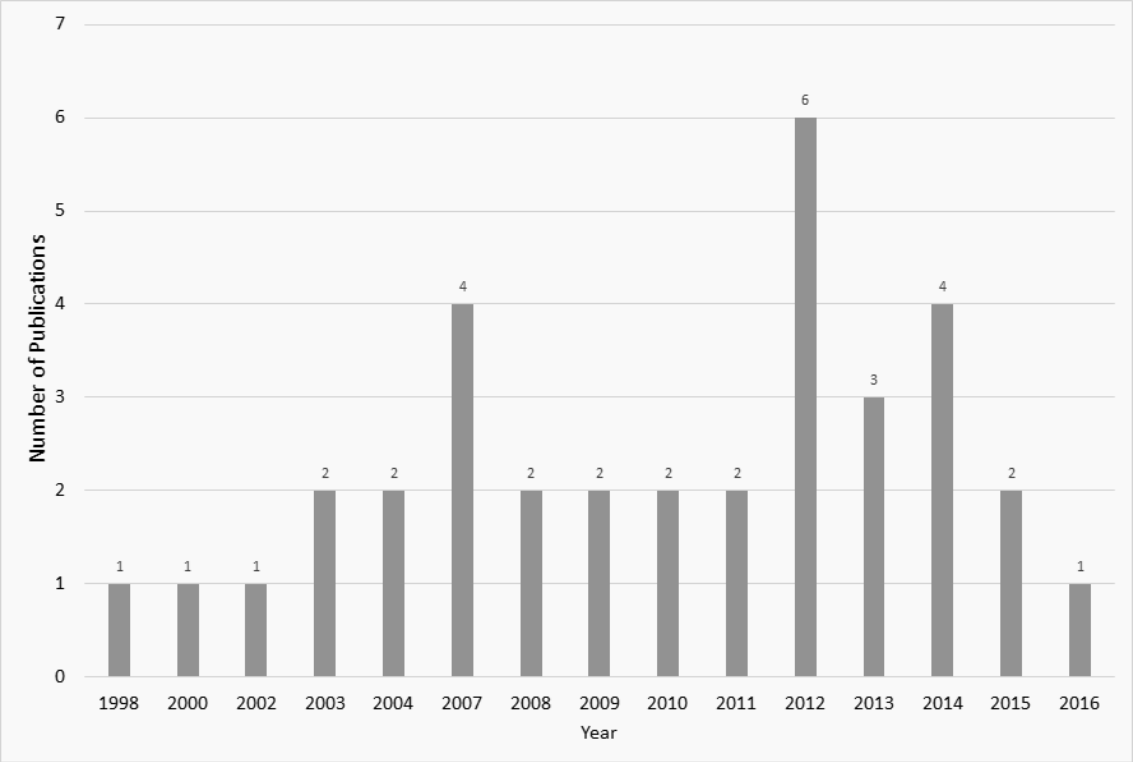

Figure 15 - Number of publications related to TMFP grouped by year.

4.7.1 Meta-analysis of Related Literature

Generally, most models predict the trend of financial market stocks. Typically, textual information is categorized as negative, positive, or without impact on the trend. In particular, the models differed according to:

- **Number of items to be processed:** among the reviewed works, the most common number of news articles is between 10k and 1M, being associated with the time period and news source. Volumes range from 216 (ZHAI et al., 2007) in the Australian Financial Review to 30M items (MAKREHCHI et al., 2013).

- **The sources of textual information:** Bloomberg, Down Jones, German Society for Ad Hoc Publicity (DGAP), Financial Times (FT), Forbes, Reuters, Wall Street Journal (WSJ), Yahoo!Finance, Google Finance and social media (blogs, forums and Twitter).

- **The time periods used for extracting data used to predict the trend:** there is a discrepancy among the reviewed works with the majority of data having been gathered in less than six months or more than 24 months.

- **The terms of market:** the great majority of the reviewed works are devoted to predicting the movements of stocks and foreign exchange (ForEx).

- **The terms of index:** most studied indexes are DJIA and S&P 500, followed by the local indexes according to author's country.

- **The terms of exchange:** just as with index terms, most studies involving exchanges focus on NYSE, NASDAQ, and other exchanges depending on the author's country.

- **The time-frame:** most of the reviewed works aim at predicting the market movements on a daily basis (21 papers), followed by the intraday time-frame (12 papers) and yearly basis (2 papers). In addition, one study was conducted to predict the effect of news on the stocks before and after the elections in India (VAKEEL & SHUBHAMOY**,** 2014).

- **The number of classes and target prediction:** most of the reviewed works focus on classification and the majority of publications use two classes for prediction (15 papers).

- **Feature Selection:** about 2/3 of the reviewed works use bag of word (BOW). In terms of feature representation, 1/4 of the reviewed works use TF-IDF; another 1/4 use term frequency (TF), TF-CDF, or binary representation; and most of these feature representations occur together with BOW.

- **Dimensionality Reduction:** the majority of reviewed works use some type of statistical measurement such as Language Models, Information Gain, Chi-Square, Minimum Occurrence Per Document to define the most valuable features given a threshold (FORMAN, 2003), and normally this is combined with BOW, stemming, and stop words removal. Another common approach is the use of pre-defined dictionaries, where the non-existing words will be removed.

- **Learning Algorithms:** the most common machine learning algorithms applied to TMFP are the SVM (CORTES & VAPNIK, 1995), two or more algorithms like k-NN (FIX & HODGES, 1951), Decision Trees (KOHAVI & QUINLAN, 2002), and others (WU et al., 2007) in the same study.

- **Number of publications grouped by the percentage of test and training size:** unfortunately, almost 1/3 (10 papers) of the reviewed works did not provide this information. The data splitting between 20% and 50% is the second most frequent group, and ratios between this range are a common practice in machine learning.

- **Sliding Window:** only 22% of the reviewed works applied this technique: WÜTHRICH et al. (1998), PERAMUNETILLEKE & WONG (2002), TETLOCK et al. (2008), BUTLER & KEŠELJ (2009), VU et al. (2012), JIN et al., (2013), VAKEEL & SHUBHAMOY (2014), NASSIRTOUSSI et al. (2015), BECKMANN, 2017 and this current work.

- **Semantics:** more than half of the reviewed works used some semantics approach, but in general it was only applied to discover word relationships like synonyms and hypernyms, aiming at the word weighting and dimensionality reduction by weighing or replacing related words using a thesaurus or WordNet (MILLER, 1995).

- **Data Balancing:** only seven studies paid attention to this subject (PERAMUNETILLEKE & WONG, 2002, MITTERMAYER, 2004, SONI et al., 2007, DE FARIA et al., 2012, MAKREHCHI et al., 2013, BECKMANN, 2017).

CHAPTER 5 – Methodology

This model was based on text mining techniques applied to the news from companies that compose the DJIA index. The construction of prediction models based on price changes with text mining requires a long and automated processing that encompasses the collection of news and stock prices, treatment of this textual data into a bag of words vector, training, test and simulation. Therefore, to accomplish the objective, a complete process of data mining and text mining was developed to predict the DJIA index movements along the day. The entire TMFP process was developed with the RapidMiner platform and its respective extensions (MIERSWA, *et al*.., 2006).

The process uses existing text mining processes and a new alignment technique that is briefly made by picking up the news published by YahooFinance[1] corresponded to the 5 stocks with highest trading volume in each minute. The quality of the model is measured by Precision, Recall and F-Measure indices. The main processing flow can be seen in Figure 18. Each sub-process from Figure 18 will be explained in the next subsections.

Some open source software components and RapidMiner experiments previously developed in BECKMANN (2017) were used in this methodology. In his work, an extensive review of the literature related to TMFP was conducted, and problems like the use of Accuracy as classification measure, lack of information about the model evaluation, and incorrect use of cross validation were identified.

[1] http://finance.yahoo.com

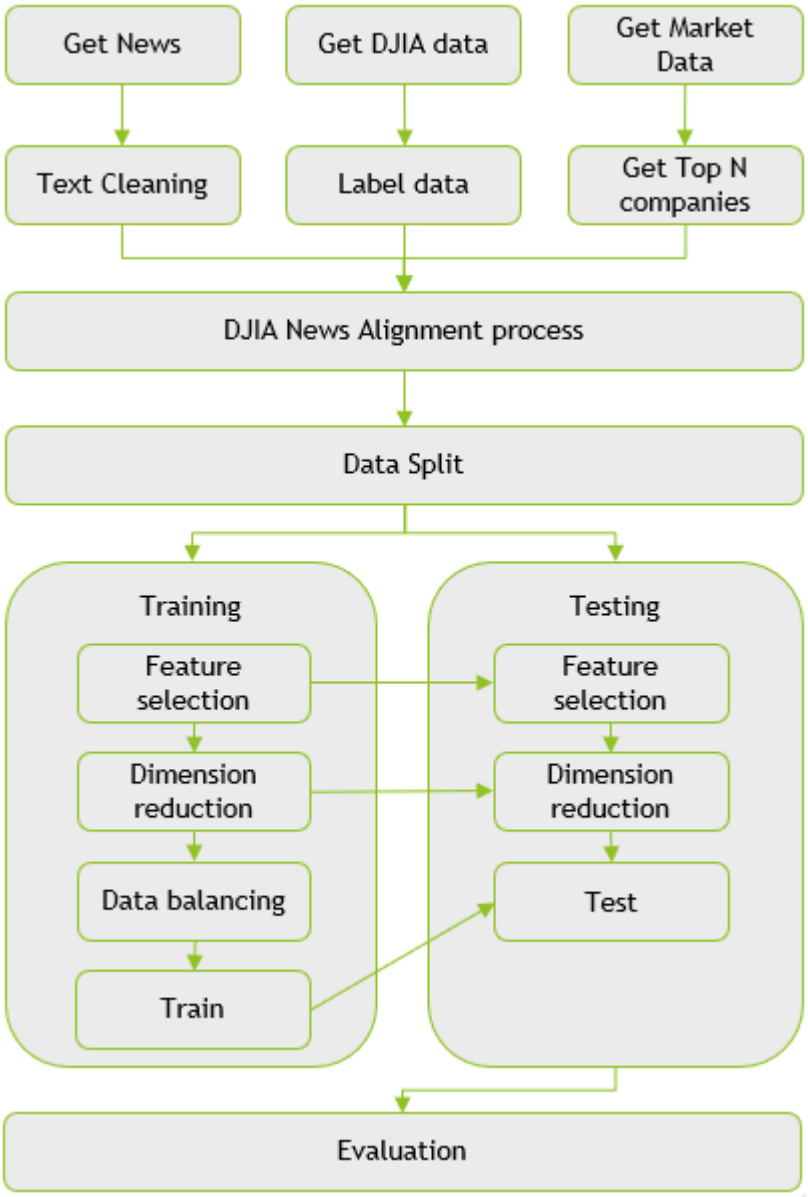

Figure 18 - Main process workflow.

5.1 Data Gathering

5.1.1 Obtain News

The first step in the TMFP process is data collection. The research component of data collection is common to all fields of study and this fact allows us to respond to our hypotheses and results. Following this purpose, the news articles and stock prices were

collected from the internet to build the experiment's dataset. To make this possible, a web crawler (DHAKA, *et al.*., 2013) was developed using the RapidMiner's Web Mining extension that collected the news related to the 30 companies listed in DJIA.

The source of news came from YahooFinance, a media property which provides financial news associated with the company's stock symbol. The information retrieved is composed of news in English, the stock symbol, and the published date and time in the time zone where the stock is traded. For news articles released when the markets are closed, the publication date and time to be considered are the exchange opening time in the next available trade date.

5.1.2 Text Cleaning

Text cleaning is an essential step before the data is ready for analysis since data obtained from the web is usually highly unstructured and noisy. To achieve strong insights and build enhanced algorithms, it is necessary to play with clean data. This step removes escaping HTML characters which get embedded in the original data of the text contents and it also removes records of empty content or system messages such as "page not found" messages.

5.1.3 Obtain Stock Labeling

In order to obtain stock prices associated with the companies under study, a Java languages services client of the already developed tools for this purpose in BECKMAN (2017) was used. The market data were collected by a free web service[2] that provides minute-by-minute stock prices and other quantitative values from different companies traded on the NASDAQ and NYSE stock exchanges. At the same time, DJIA historical indexes were first retrieved from EODData[3]. Each market data record collected includes the stock symbol, the traded volume, the previous day's closing price, the opening price, the last price traded at that minute, and the date and time this trade occurred (both in

[2] http://restfulwebservices.net
[3] http://eoddata.com

EST and UTC times). Such records are rigorously stored and processed in their chronological order.

5.1.4 Stock Price Labeling

As soon as market data are available, it is necessary to identify the highest and lowest observed prices as well as to assign a label for each record. Thus, in this work, the market data records that were linked exclusively to the DJIA index were labeled as SURGE and PLUNGE, whereas the individual records were labeled NOT RECOMMENDED. These labels are respectively identified as 2, -2, and 0 in the database.

The labeling method uses slopes to measure price changes, where, for positive slope, that is, those that have obtained an increase greater than or equal to 75% of the maximum rise observed during the period, the label of SURGE is assigned. Equivalently, negative slopes are assigned the label of PLUNGE. Therefore, in all other cases the records are labeled as NOT RECOMENDED. Later, the labels PLUNGE and NOT RECOMENDED were merged in order to simplify the model, minimize the classification error and improve performance. At the end, the labels defined later become:

0 = NOT RECOMENDED

2 = SURGE

In Figure 19 it shows the slopes identified in DJIA index by minute in a day.

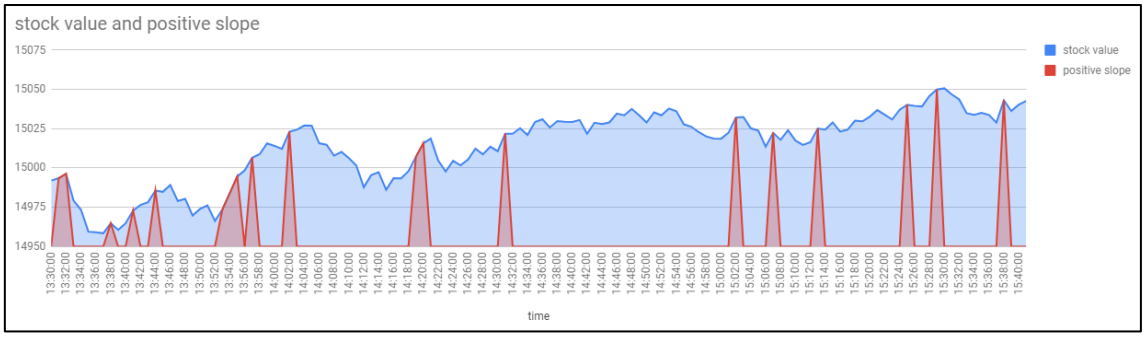

Figure 19 - Identified positive slopes by minute in DJIA index

There are just few surges identified per day. As observed in collected data, this behavior is rare, which leads to unbalanced data with the majority being set as NOT RECOMMENDED.

Table 9 shows a sample of the quantity of NOT RECOMMENDED and RECOMMENDED data per day. The average of surges per day in an observed year is 1.25% of the total data in that year.

Table 9 - Total number of identified classes per day.

Date	Recommended	Not Recommended	Class 2 %
2013-01-02	3	388	0.7673
2013-01-03	1	389	0.2564
2013-01-04	1	390	0.2558
2013-01-07	15	375	3.8462
2013-01-08	3	387	0.7692
2013-01-09	2	388	0.5128
2013-01-10	2	388	0.5128
2013-01-11	2	389	0.5115
2013-01-14	3	388	0.7673
2013-01-15	1	387	0.2577
2013-01-16	9	382	2.3018
2013-01-17	2	389	0.5115
2013-01-18	3	388	0.7673
2013-01-22	2	389	0.5115

2013-01-23	1	389	0.2564
2013-01-24	2	389	0.5115
2013-01-25	3	388	0.7673
2013-01-28	2	389	0.5115
2013-01-29	9	381	2.3077
2013-01-30	3	388	0.7673
2013-01-31	8	383	2.046
2013-02-01	2	389	0.5115
2013-02-04	4	387	1.023
2013-02-05	2	389	0.5115
2013-02-06	7	384	1.7903
2013-02-07	15	376	3.8363
2013-02-08	6	385	1.5345
2013-02-11	7	384	1.7903
2013-02-12	3	387	0.7692
2013-02-13	9	368	2.3873

5.2 Top n Stocks Negotiated

This step is the key process of this model along with the alignment process. Processing news minute by minute from 30 companies is insane and can take weeks to yield any result. It is also important to mention that news from 30 companies can include not only important information that can lead one stock to be highly negotiated, but also many others irrelevant that don't lead to anything. This irrelevant information only brings noise to the model and unnecessary processing.

Figure 20 shows how the relevant information can be suppressed by the irrelevant ones. It shows the quantity of news in a day and the quantity of news from the top 5 trading companies in each minute of a day.

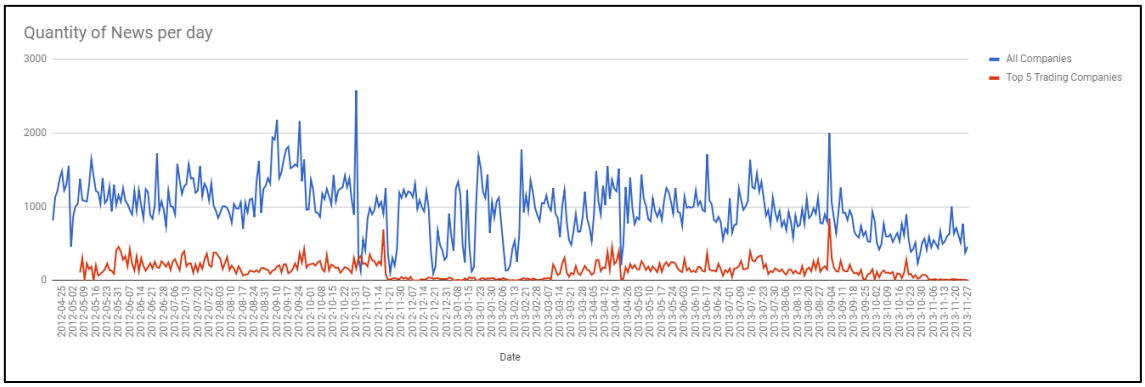

Figure 16 - Comparison of all News in a day from 30 companies composed by DJIA index against the top 5 trading companies.

All experiments running under this scenario required approximately one week of processing and resulted in a very poor classification. The new method proposed gets the top N stocks negotiated in each minute. The text alignment process will deal only with the news from the top N companies given by this process in each minute. Irrelevant news is not processed, and noise is drastically reduced using the TopN method.

Figure 21 shows the number of different companies that produced any news on a given day. News were generated about almost all companies each day, not just from the top 5 most traded companies in a minute. Figure 21 also shows that on a given day the number of companies that dominate the trade is around 10.

Experiments were done with different values of N, but the best tradeoff was achieved with N from 1 to 5. Due to the complexity and amount of data, in order to improve performance, this step and the next one (text alignment) were done using complex SQL scripts and the results were later used as input in the RapidMiner process.

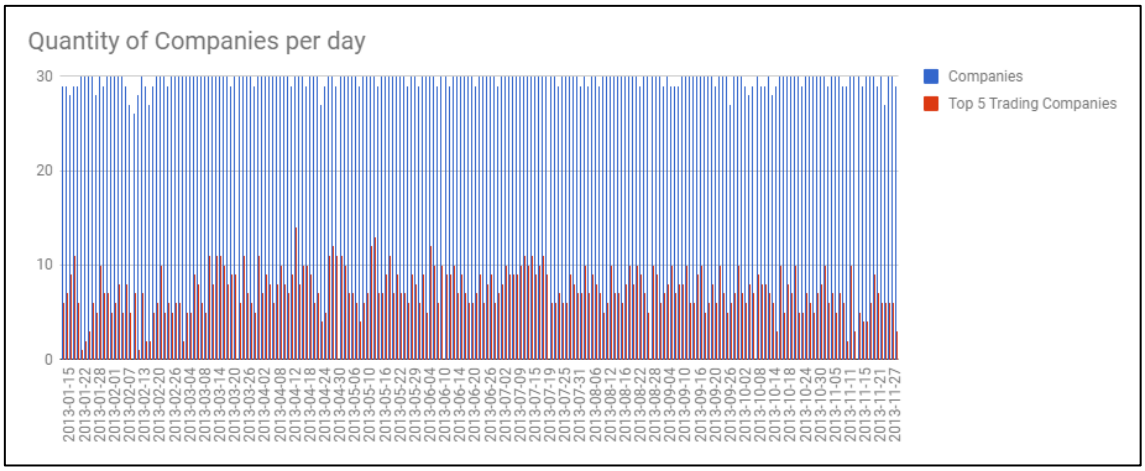

Figure 17 - Comparison of all companies composed by DJIA index in a day that produced any news against the top 5 trading companies.

Table 10 shows a sample of the top 5 trading companies per minute listing the trade time, the company symbol, the volume negotiated, the percentage negotiated among the top 5 and the top 5 rank in that minute.

Table 10 - Sample of the top 5 trading companies by minute

utc_time_	symbol	volume	symbol_percent	symbol_rank
14:14:00	BAC	33195	68.67%	1
14:14:00	GE	6094	12.61%	2
14:14:00	CAT	4050	8.38%	3
14:14:00	JPM	3700	7.65%	4
14:14:00	AA	1300	2.69%	5
14:15:00	BAC	29059	66.25%	1
14:15:00	AA	8084	18.43%	2
14:15:00	GE	3100	7.07%	3
14:15:00	KO	1852	4.22%	4
14:15:00	CAT	1768	4.03%	5
14:16:00	BAC	475791	98.99%	1
14:16:00	GE	1450	0.30%	2
14:16:00	HPQ	1316	0.27%	3
14:16:00	JPM	1050	0.22%	4
14:16:00	MRK	1050	0.22%	5
14:17:00	BAC	321093	97.11%	1

14:17:00	GE	5600	1.69%	2
14:17:00	JPM	2468	0.75%	3
14:17:00	AA	900	0.27%	4
14:17:00	PG	600	0.18%	5

In this sample, the volume traded per minute is dominated by one company. In some cases, it represents over 90% of the volume traded among the top 5. Still considering the top5, the volume traded of over 40% of the whole data in a minute is dominated by just one company, while over 60% is dominated by two companies. Three companies represent almost 80% of the total samples. Although the volume traded per minute is dominated by one company in most of the cases, four other companies, although not so representative, are part of the companies in play. The other 25 companies are much less representative or show no trade at all and the news related to them only add noise and processing time to the model.

Each minute is dominated by different sets of companies and the purpose here is to deal only with news related to these companies. According to what was observed regarding to the dominated companies, dealing with just the most traded company per minute may still produce good results and dramatically decrease processing time in the training phase.

5.3 DJIA Text Alignment Process

This process is responsible for aligning news from the top N stocks negotiated each minute with the DJIA index minute by minute as shown in the Figure 22. As said in the previous section, the alignment was done using SQL scripts and added later in RapidMiner process.

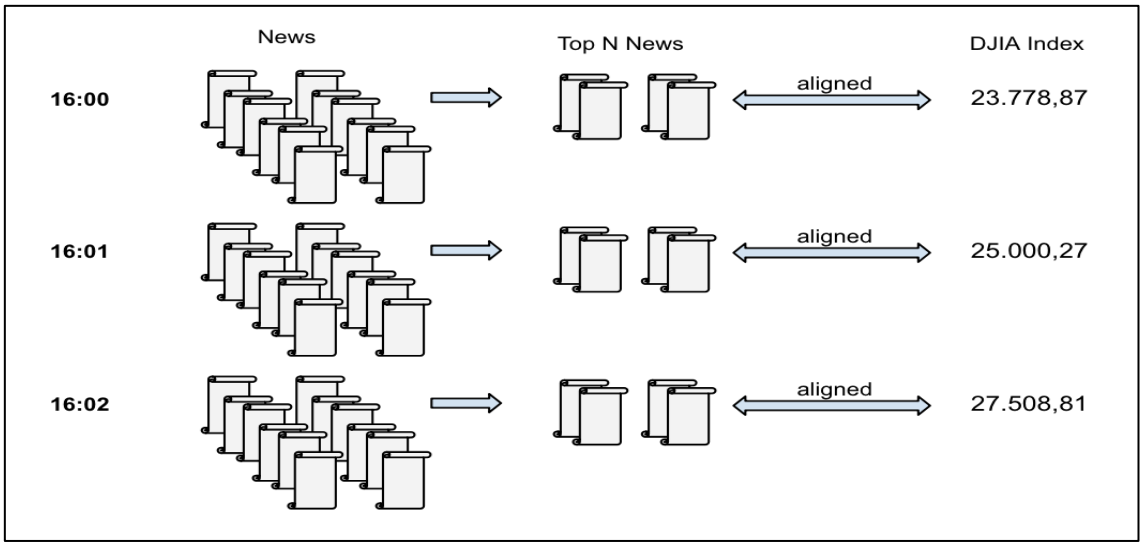

Figure 18 - DJIA Text alignment process.

The alignment process is a procedure in MySql that works as following:

- Defines the N companies in TopN process by minute to use in alignment
- Defines the window in which the news will affect the DJIA index. It is a range starting with the published date.

  ie. Window of 3 minutes defines a range in which the news will affect the DJIA index starting with the published date (see Figure 23).
- Align the news and the DJIA index according to the definition from the previous steps.

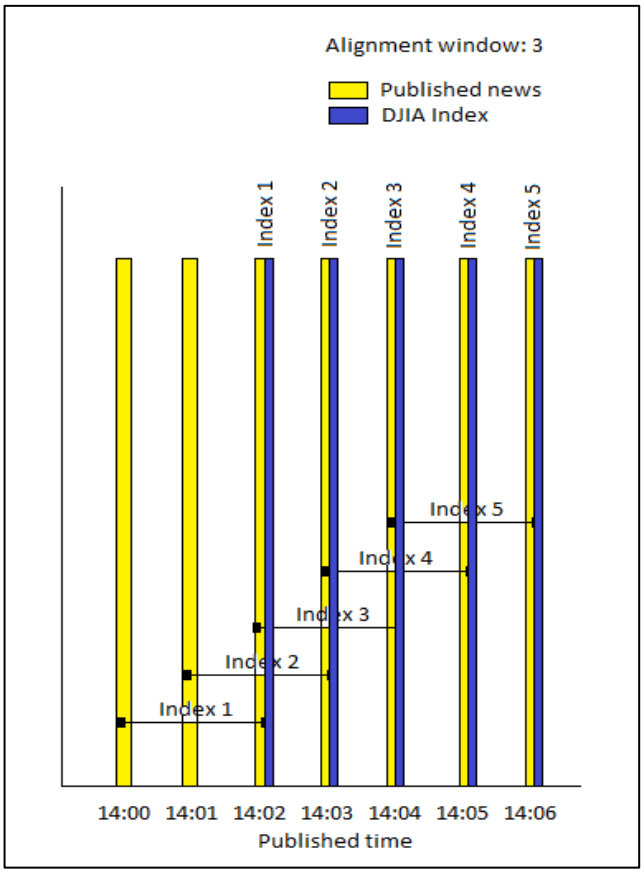

Figure 23 - Demonstration of the DJIA alignment window of 3 minutes. 3 minutes range starting by the published date.

The total samples per index aligned depends on:

- N companies selected for the TopN: more companies, more news to align
- Alignment window: small window, less news to align

Increasing the alignment window, increases the number of news to align. Although it suggests more time to process, it was not observed during experiments.

Table 11 shows the total number of news aligned using different TopN and alignment window configurations.

Table 11 - Comparison of total news aligned per index using different TopN and alignment window configurations.

	Alignment Window			
TopN	1	3	5	10
3	79554	159120	238659	437302
5	111942	224060	336136	616052

5.4 DJIA Text Processing

The raw text existing in the news articles is transformed into vectors, then a predictive model and word lists are built in order to construct the bag of words (BOW). Word weights are generated by feature selection and the balanced records are generated by KNN-Undersampling (BECKMAN, 2015). In general, this process is responsible for preparing the data for training and testing phases as described in the following sections.

5.4.1 Feature Selection

Once labeled and stored, news articles must be transformed into a structured format to be processed by statistical methods and machine learning. This process is performed using the text mining extension of RapidMiner (MIERSWA, *et al.*., 2006), a lightning-fast data platform that unites data preparation, machine learning, and predictive model deployment. This software provides several operators for text processing such as stemming, tokenisation, n-grams, stop words and integrated dictionaries. In addition to that, the entire set of documents, also known as corpus, is transposed into a matrix that is called the Bag of Words (BOW). In this model, each document is represented as the bag (multiset) of its words and terms which in turn are represented by a matrix (Figure 24).

All documents and their lists of words are converted to lowercase; words that do not contain information, such as stop words and words with less than two characters, are

removed by the template. Words with frequency lower than 2% and greater than 95% are also removed.

Stemming, described by LOVINS (1968) and PORTER (1980) was also used, but it didn't improve the results and therefore was removed in order to avoid the loss of any important information.

complete...	complex	compliance	complianc...	complianc...	comply	comply_g...	comply_g...	component	compone...	compone...	compone...	composite	comprehe...	comprehe...	comprehe...
0	0	0	0	0	0	0	0	0	0	0	0	0	0.023	0.028	0.028
0	0	0	0	0	0	0	0	0	0	0	0	0	0.023	0.028	0.028
0	0	0	0	0	0	0	0	0.024	0	0	0	0.032	0	0	0
0	0	0	0	0	0	0	0	0.024	0	0	0	0.032	0	0	0
0	0	0	0	0	0	0	0	0.024	0	0	0	0.032	0	0	0
0	0	0	0	0	0	0	0	0.024	0	0	0	0.032	0	0	0
0	0	0.035	0	0	0.046	0.052	0.052	0	0	0	0	0	0	0	0
0	0	0.035	0	0	0.046	0.052	0.052	0	0	0	0	0	0	0	0
0	0	0.035	0	0	0.046	0.052	0.052	0	0	0	0	0	0	0	0
0	0	0.035	0	0	0.046	0.052	0.052	0	0	0	0	0	0	0	0
0	0	0.037	0.040	0.040	0	0	0	0	0	0	0	0	0	0	0
0	0	0.037	0.040	0.040	0	0	0	0	0	0	0	0	0	0	0
0	0	0.037	0.040	0.040	0	0	0	0	0	0	0	0	0	0	0
0	0	0.037	0.040	0.040	0	0	0	0	0	0	0	0	0	0	0
0.027	0	0	0	0	0	0	0	0	0	0	0.021	0	0	0	0
0.027	0	0	0	0	0	0	0	0	0	0	0.021	0	0	0	0
0.027	0	0	0	0	0	0	0	0	0	0	0.021	0	0	0	0
0.027	0	0	0	0	0	0	0	0	0	0	0.021	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 24 - Bag of words sample matrix with TF-IDF representation from articles of DJIA's companies

The concept of probabilistic language model n-grams (SIDOROV *et al.*., 2014a), which consists in a contiguous sequence of n words from a given sample of text, was used in this work. This model can store more contexts with a well-understood space time tradeoff, enabling small experiments to scale up efficiently and reduce the dimensionality. As a last step, the words are represented as a TF-IDF measurement (17). The BOW matrix contains columns that represent the selected words and n-grams. The resulting word list is composed ofx the BOW matrix along with the corresponding word frequency. The documents are converted into this word list during training and test steps.

$$TFIDF(t, d, D) = TF(t, d).log\frac{|D|}{n_t} \quad (17)$$

Where:

- *D* is the corpus that contains the document *d;*
- $|D|$ is the number of documents existing in the corpus;
- n_t is the number of documents where the term *t* appears.

5.4.2 Dimensionality Reduction

To make a more representative set of variables and strippers, Pearson's Chi-Square statistic was also applied.

5.4.3 Data Balancing

When working with supervised learning, one of the main problems that affects classification activities refers to the treatment of data sets whose classes have a minority of instance. This phenomenon leads to an unbalanced dataset and depending on how unbalanced it is, it can lead to an algorithm failure (WEISS & PROVOST, 2001). To correct this error, the KNN-Undersampling method presented by BECKMANN (2015) was applied. KNN-Und acts by removing instances from the majority classes and at the same time cleaning the decision surface, reducing the class overlapping. Also, distinct from other methods, KNN-Und has a deterministic behavior, which leads to stable results with standard deviations equals to 0.

5.5 Data Splitting

In order not to contaminate the training data with information from the future, a duly adjusted recommendation model was used for a time series in which the data set was kept chronologically adjusted and divided into training and testing. In addition, with the aim of reducing noise throughout the data, a new training was conducted each week with the aim of adjusting the model to the new reality and maximizing the efficiency of the classifier. This technique is known in the literature as a sliding window (DIETTERICH, 2002).

5.6 Training

The training dataset contains 7 months of data while the test dataset contains 1 week in a sliding window of 5 weeks in order to evaluate the model. As processing proceeds to a new week, the training data set integrates the previous week and discards the first week. Thus, according to how the data is split, it is necessary to rebuild and test a new mode. This work applied the SVM (CORTES & VAPNIK, 1995) as the machine learning algorithm, with LIBSVM implementation (CHANG & LIN, 2011) and Radial Basis Function (RBF) as kernel. The parameters C and Gamma required by SVM are adjusted through a grid search, using the training dataset with a 10-fold cross validation to discover the best value of F-Measure obtained with the SVM classifier, given a pair for C and Gamma parameters, as described in HSU, *et al*. (2003).

5.7 Test

Once the training phase is complete, it generates the model which is then tested against the test data. The result of this step is accuracy, precision, recall and f-measure for records classified as PURGE and records classified as NOT RECOMMENDED. These were the metrics used to evaluate the quality of the model.

5.8 Evaluation

Evaluation is the last step of the process, and it checks the quality and applicability of the predictive models constructed during the previous phases. It is a manual step in which the metrics produced by the previous phase are evaluated and the process and parameters are updated until the best result are achieved.

5.9 Implementation

The methodology described in the previous sections were implemented as following:

- Data gathering: process to collect news and DJIA indexes throughout the day based on open source software components and RapidMiner experiments developed in BECKMANN, 2017.
- Labeling process: label DJIA indexes as 2 (SURGE) or 0 (NOT RECOMMENDED) using database procedure.
- TopN process: data base procedure that generates data with the top N most traded companies each minute.
- DJIA News Alignment process: database procedure that aligns news and the DJIA index by minute according to the top N process and alignment window definitions.
- Classification model process: supervised learning process implemented in RapidMiner.

The first four processes were already explained in the previous sections. The Classification model process in RapidMiner will be addressed in the following section.

5.9.1 Classification Model Process

This process was implemented in RapidMiner and it assumes that the first three processes are complete. The main process is designed as shown in the next figure.

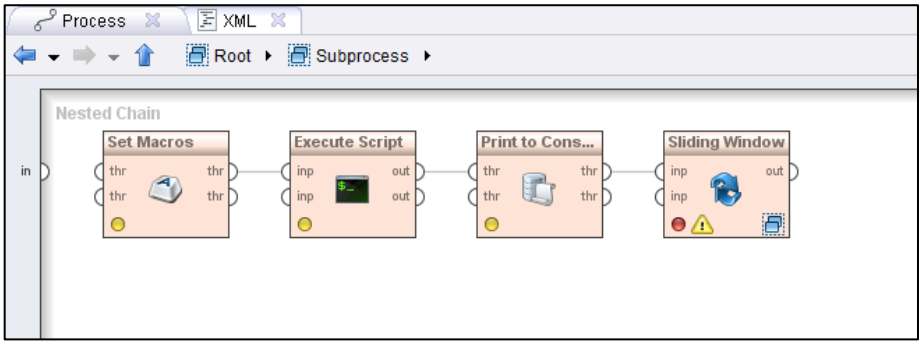

Figure 25 - DJIA RapidMiner process - main

The main process is divided into 4 sub-processes as shown here:

- Set Macros: set up the parameters used in the process.

Table 12 - DJIA RapidMiner setup parameters

Macro	Value
symbol_	DJIA
experiment_description_	Experiment of 7 months training and 1 week test. Sliding window for 5 weeks.
ticket_	20180605_DJIA_3
db_host	localhost
k_	21
att_weight_	0.10
outputdir_	/home/mvale/trademiner/data
algo_	libsvm
outputtable_	experiment_result_auc4
labels_	0,2
min_performance_	0.50
base_date_	2017/06/10
sliding_window_unit_	WEEK
sliding_count_	5

training_interval_	7
training_interval_unit_	MONTH
testing_interval_	1
testing_interval_unit_	WEEK
experiment_id	1

- Execute Script: script in groove responsible for creating the directory structure in local machine to determine where the output files generated during the process will be placed and creation of information file describing the experiment.

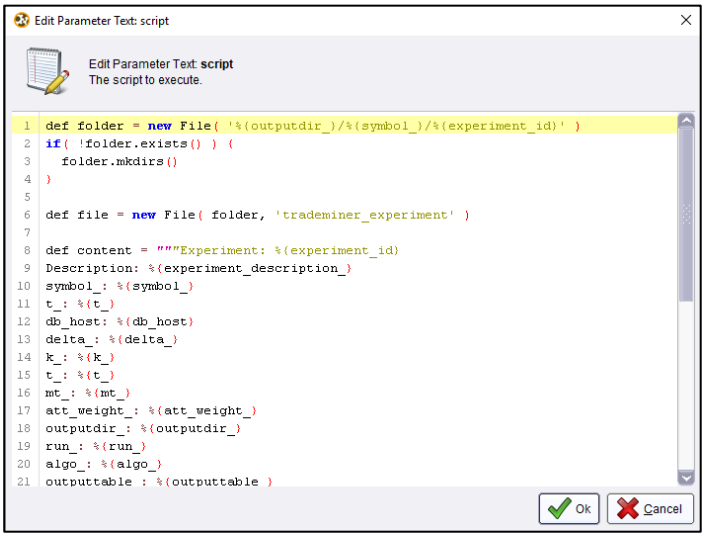

Figure 26 - DJIA RapidMiner process - run log script

- Print to console: prints the parameters to console for debug purpose.
- Sliding Window: the training and testing process is repeated for each window as defined in Macros. The window can move by day, week or month which means that the dates defined for training and testing data will move according to it. The subprocesses are shown below.

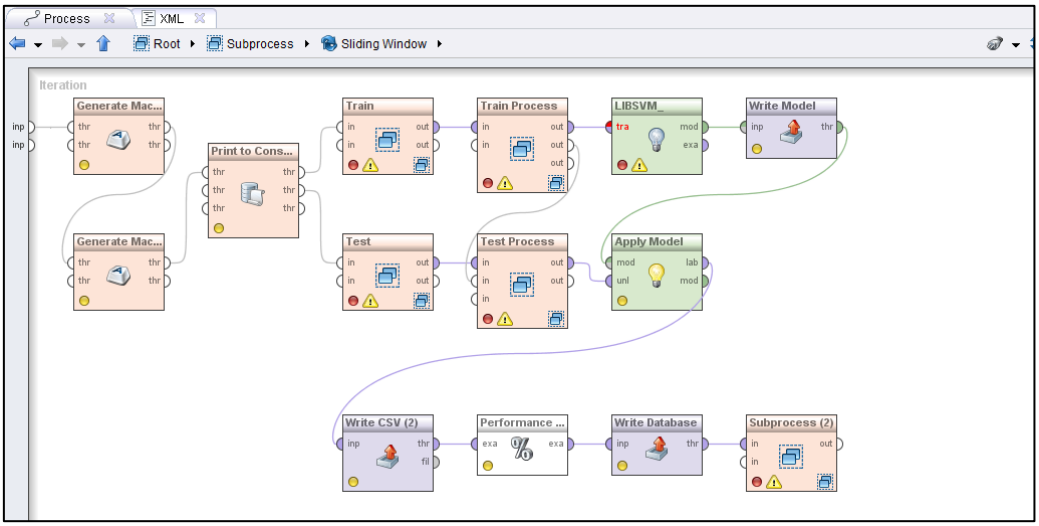

Figure 27 - DJIA RapidMiner process - Main text mining process run in each sliding window iteration.

The first two processes will generate the start and end dates for training and testing data according to the defined sliding window. The training dates never overlap with the testing dates, which guarantees that the training data is not contaminated by future data used in testing.

The first Train and Test processes are responsible for getting the news along with the alignment labels from database related to the periods defined in previous steps.

The next Train process is responsible for preparing the data for the SVM operator. All the sub-processes defined in this step were already addressed in previous sections.

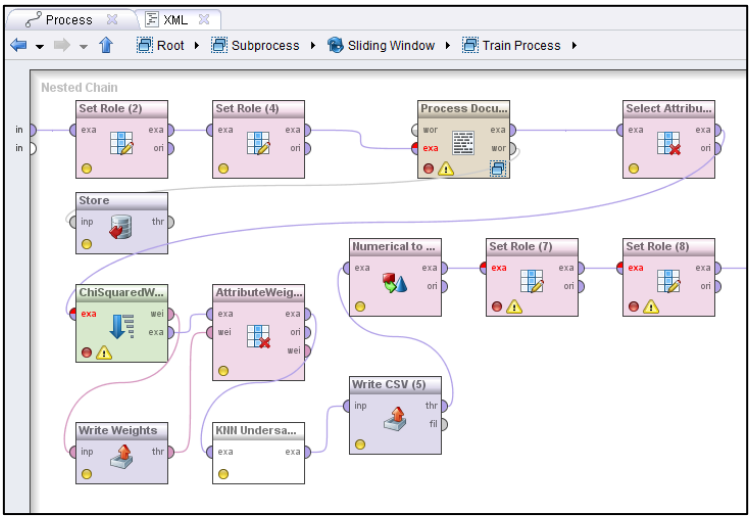

Figure 28 - DJIA RapidMiner training data preparation process

Once the training preparation data is complete, the output is used in the SVM operator to start the classification learning process.

The next Test process is responsible for preparing the data for testing. The BOW and weights generated in Train process are used as input in this step.

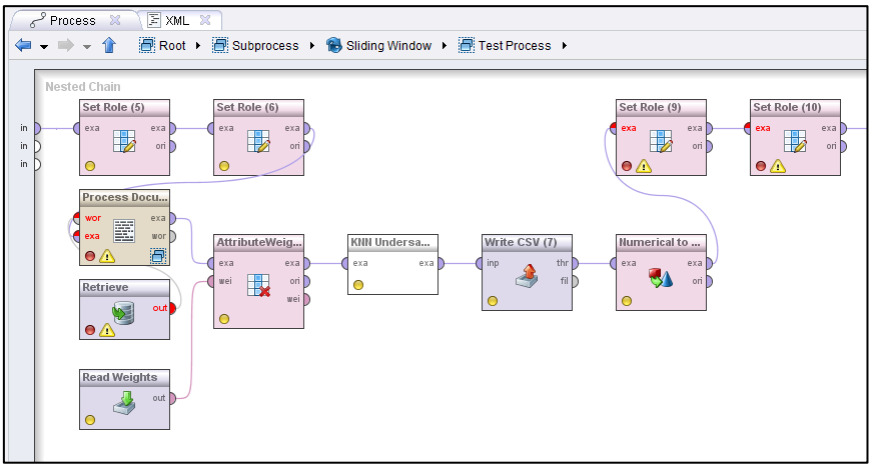

Figure 29 - DJIA RapidMiner testing data preparation process

Finally, the model generated by SVM operator is tested with the output from the Train process and the metrics are generated and saved in a database during the next processes.

CHAPTER 6 – Experiments

The experiments were executed respecting chronological order, since news from a given period corresponds to the indexes for that same period, so it is important to not shuffle and split the data. Therefore, training and testing data must be in chronological order during split.

All experiments described in this work used the default values for KNN Und and SVM from trademiner described by BECKMANN, 2017.

The computer specification used was a Core i7 Quad-core with 16 GB RAM DDR4 running a Linux distribution as the operational system. The total data collected was over 15 GB of data.

The number of samples tested will differ when using different training configuration. The word list and chi-square weights generated from the training phase are used as input for the testing data preparation, which filters the attributes and samples to be tested. So, it is expected to have some variation between the number of training samples in each experience.

The experiments were done first without the Top N Stocks method previously described, using all news available from companies and the results were poor. Afterward, by applying the new method, the results improved significantly and showed that previous results were distorted by news that was only adding noise to the model as shown in Table 13.

Table 13 - Experiment running 1 year training, 1 month testing and no sliding window. Comparison between all news against TopN method.

Experiment	Precision	Recall	F-measure
Top 3	88.83%	89.81%	89.19%
Top 5	87.15%	88.03%	87.51%
All News	45.29%	38.85%	36.45%

The next experiment aimed to discover the behavior of different alignment windows. This experiment was set up with 7 months training, 1 week testing and sliding

window per week for 5 weeks as shown in Table 14.

Table 14 - Experiment: Comparison of results using different alignment window running 7 months training, 1 week testing through a sliding window of 5 weeks.

Experiment id	Week	Alignment Window	Precision (Surge)	Accuracy	Recall	F-measure
2	5	1 minute	0.1739	0.0324	0.0384	0.0154
2	4	1 minute	0.7143	0.0938	0.1513	0.1310
2	3	1 minute	0.9771	0.8253	0.8870	0.8786
2	2	1 minute	0.1818	0.0486	0.0523	0.0129
2	1	1 minute	0.7538	0.1119	0.1809	0.1600
1	5	3 minutes	0.9387	0.8948	0.8939	0.8944
1	4	3 minutes	0.9761	0.9075	0.9317	0.9172
1	3	3 minutes	0.9689	0.5733	0.7133	0.6800
1	2	3 minutes	0.9402	0.8500	0.8693	0.8693
1	1	3 minutes	0.9393	0.7942	0.8446	0.8165

The results using an alignment window of 1 minute were very poor, except for week 3. The overall results using 3 minutes are clearly better which demonstrate that news affects the index beyond the time of publishing. The Figure 30 shows the comparison of results following these two approaches.

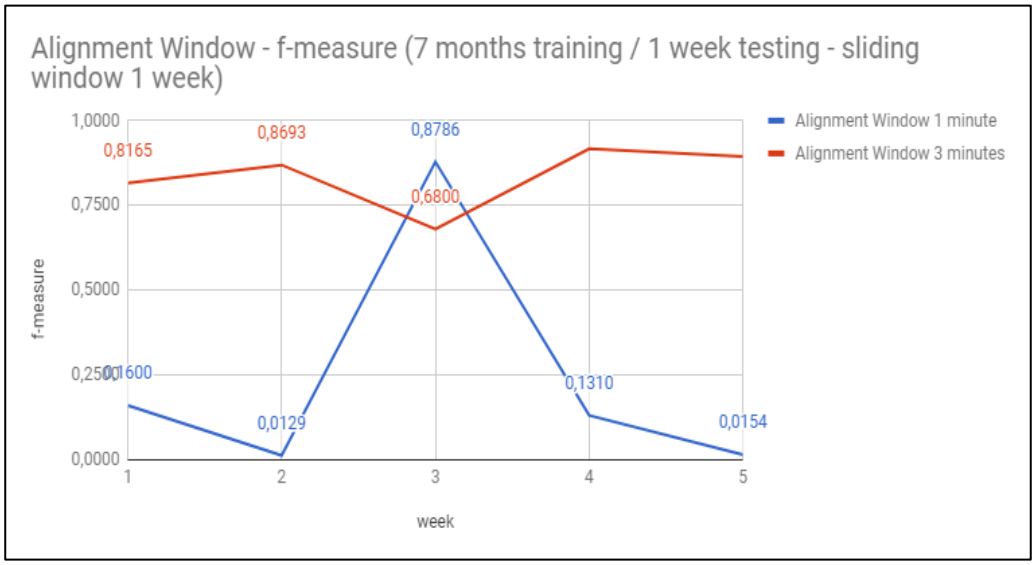

Figure 30 - Comparison chart from running experiment with alignment window of 1 minute and 3 minutes.

It is important to mention that indexes without any published news are not tested, so an alignment window of 1 minute will reduce the number of indexes evaluated. The Table 15 shows the number of samples evaluated in each week and in Table 16 and Table 17 shows the samples of these two experiments generated from a confusion matrix.

Table 15 - Comparison of samples evaluated per week for alignment window of 3 minutes and 1 minute.

	Week					Total
Alignment Window	1	2	3	4	5	
3 minutes	933	1240	1057	724	1606	5560
1 minute	554	1029	578	352	1233	3746
Difference	59.38%	82.98%	54.68%	48.62%	76.77%	67.37%

Table 16 – Samples evaluated running experiment of 7 months training, 1 week testing, sliding window of 5 weeks and alignment window of 1 minute.

Week	False Positive	True Negative	False Negative	True Positive
5	32	1155	38	8
4	8	309	10	25
3	7	90	11	470
2	46	961	18	4
1	13	476	16	49

Table 17 - Samples evaluated running experiment of 7 months training, 1 week testing, sliding window of 5 weeks and alignment window of 3 minutes.

Week	False Positive	True Negative	False Negative	True Positive
5	120	83	86	1317
4	44	52	15	613
3	46	433	18	560
2	0	119	67	1054

1	91	150	42	650

Although experiment 2 outperformed experiment 1 in week 3, the samples evaluated represent 54.68% of the samples in experiment 1, so experiment 1 has obtained f-measure of 68% evaluating almost the double of samples.

The next experiments compare the results from using an alignment window of 3 minutes and 5 minutes in a long period of testing. The experiments were set up with 7 months training, 3 months testing and no sliding window. The results are demonstrated in Table 18.

Table 18 - Experiment: Comparison results for different alignment window running 7 months training, 3 months testing and no sliding window.

Alignment window	Precision (Surge)	Accuracy	Recall	F-measure
5	0.9513	0.7757	0.8418	0.8084
3	0.9557	0.8287	0.8733	0.8552

Table 19 shows the confusion matrix and the samples difference between the experiments.

Table 19 - Comparison of results based on confusion matrix from experiments running 7 months training, 3 months testing and no sliding window.

Alignment window	False Positive	True Negative	False Negative	True Positive	Total
5	1809	3969	741	14480	20999
3	786	2143	573	12357	15859
				Difference	75.52%

Using an alignment window of 5 minutes increased the number of samples evaluated but decreased the quality of the model. Increasing the alignment window also increases the processing time since there is more news to process for each affected index. The Table 20 shows the total amount of news to process for a different alignment

window configuration.

Table 20 - Comparison of the total news to process for different TopN and alignment window configuration.

	Alignment Window			
TopN	1	3	5	10
3	79554	159120	238659	437302
5	111942	224060	336136	616052

An alignment window of 10 minutes, as compared to a window of 5 minutes, requires processing about 83% more news. The training process time for 4 months takes over 2 days and the process hangs most of the time.

According to these experiences, the best trade off was an alignment window of 3 minutes which didn't compromise the model in terms of quality and processing time.

The next experiment compares the results between choosing different N for the TopN method. This experiment was set up to run for 7 months training, a long period of testing of 3 months, an alignment window of 3 minutes and no sliding window. This experiment compares the training using the top 5, top 3 and top 1 most traded companies by minute.

Table 21 - Experiments results running 7 months training, 3 months testing, alignment window of 3 minutes.

Top N	False Positive	True Negative	False Negative	True Positive	Precision (Surge)	Accuracy	Recall	F-measure
1	885	1935	437	12493	0.9662	0.8494	0.8928	0.8725
3	988	1804	554	12376	0,9572	0.8500	0.8863	0.8682
5	804	1890	512	12418	0,9604	0.8463	0.8865	0.8688

This experiment showed that training with just the top trading company is enough to achieve good results, although using Top 3 or Top 5 also worked very well, and perhaps could provide a broader outlook compared with Top 1. One good point to observe here is that the training process time using Top1 is much faster than the others,

so even though the others can be more generalist in other periods, the processing time is a big issue and must be considered.

All these experiments were done using the SVM classification method. The next experiment tested the Naïve Bayes method by running 7 months training, 3 months testing and no sliding window (see Table 22).

Table 22 – Experiments results running Naïve Bayes classification method, 7 months training, 3 months testing, alignment window of 3 minutes.

Top N	False Positive	True Negative	False Negative	True Positive	Precision (Surge)	Accuracy	Recall	F-measure
1	1733	1087	2394	10536	0.8148	0.7790	0.7434	0.7641
3	2312	480	5709	7221	0.5585	0.6063	0.4899	0.5611

The results using the Naïve Bayes classification method were fair, but not comparable to the results achieved by the SVM.

As mentioned earlier, processing time is crucial in this kind of problem; therefore it is an important item to point out here. Table 23 shows the processing time using different configurations for Top N and alignment window.

Table 23 – Processing time of experiments running 6 months training, 1 week testing and different alignment window and Top N settings.

Top N	Alignment window	Duration
3	1	1:03:06
5	1	1:47:20
5	2	2:39:35
5	5	6:29:13
5	8	14:57:09
5	12	27:56:53

The processing time is directly affected by the number of companies and the alignment window size. Experiments have shown that increasing any of these two settings much leads to either bad results or no results at all due to extremely high processing time.

The next experiments aimed to identify the shortest training period that still produces reliable results. The minimum period of training can be a constraint when the processing time is as critical as the results. See table 24.

Table 24 – Experiment results for different training periods against the same testing data.

Training Period (months)	F-measure	Duration (H:M:S)
1	0.1850	0:12:32
2	0.1958	0:19:01
3	0.2014	0:37:52
4	0.9034	1:06:58
5	0.8844	1:22:42
6	0.8776	1:25:04

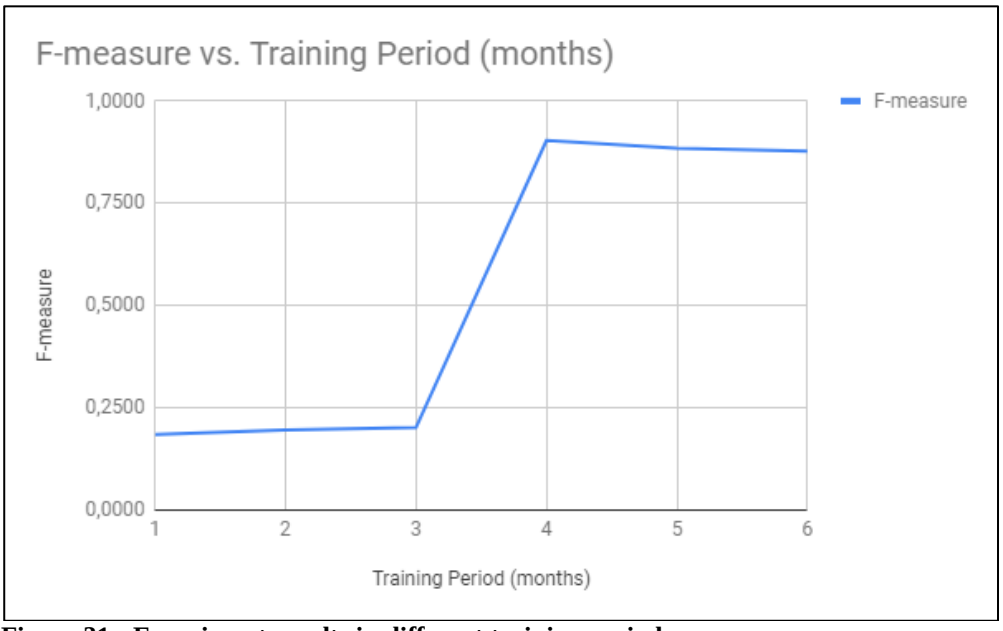

Figure 31 - Experiment results in different training periods.

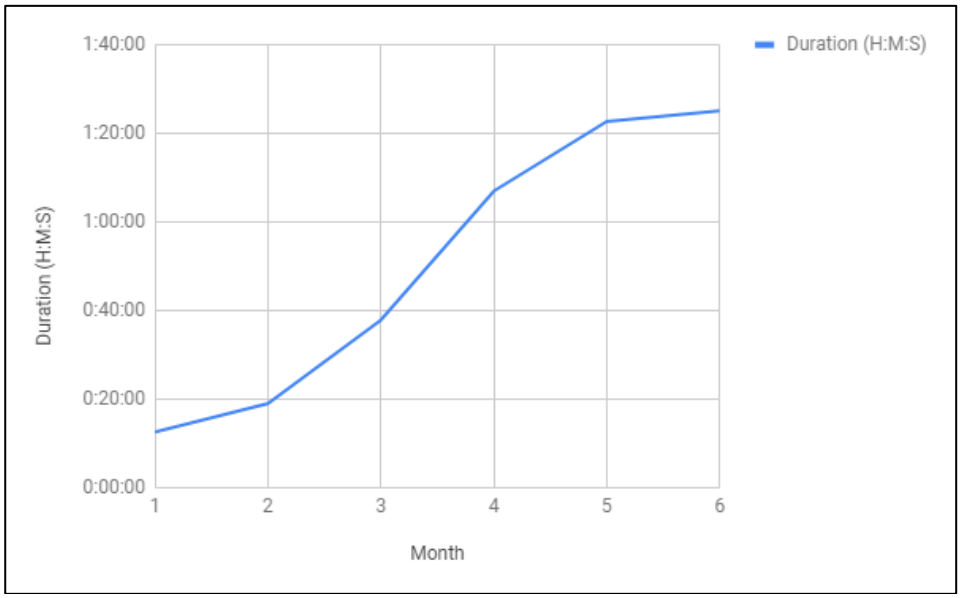

Figure 32 - Experiment training duration results in different training periods.

Using a training set with less than 4 months of data produced bad results. Based on this experiment it is possible to say that the constraint of this model for the period tested is 4 months.

The model proposed by WÜTHRICH (1998) is the one of the first researches published about the use of a priori domain knowledge. In this work, the author sought to predict the daily trend of five stock market indices, among them the DJIA index, from a base of articles published daily in web news portals. According to the results, his model shows an average accuracy of 46% based to three categories: the stock market index, the K-NN algorithm and the neural network, as shown in above. This percentage is significantly better than the precision of a random predictor, which would reach no more than 33% accuracy. With these significant results he provides evidence against the Efficient Market Hypothesis (FAMA, 1965) which states that new information is usually incorporated into stock prices within a very short time, forming a broad field for research and debate.

More recently, SOYLAND (2015) intended to use machine-learning based on the predictive power of text data for investigating the power of prediction of online news in the face of the daily price variations of the 19 important banks and financial institutions that make up the MSCI World Index. According to his results, the model they developed presented an accuracy of 59.4%; however, it failed in predicting one-day

stock price changes.

There are some limitations that may have affected his results. First of all, the web-based news data used in this study is extremely voluminous. Second, this model only used statistical and machine-learning tools to filter out the irrelevant or noisy data. Lastly, another fact that could affect the above result is the aggregation of companies, once news events, that give rise to a stock price change for one company, may not necessarily give rise to a stock price change for another company.

The models differed in particular from the financial markets they used as the data base, the sources of textual information and the time periods employed for extracting data, which makes it difficult to effectively select a superior model. However, from the comparative analysis presented above, we can readily see that the proposed method in this work outperformed the results of other studies and indicates that it can be used as a base for efficient knowledge discovery in data as shown in Table 25.

Table 25 - Table comparing some metrics from the model described in this work.

	Gidofalvi et al.. (2001)	Mittermayer (2004)	De Faria et al.. (2012)	Siering (2012)	Soyland (2015)	New Method
Precision	~55%	-	66.57%	68.45%	59.4%	88.83%
Recall	~60%	60%	65.37%	64.48%	55.9%	89.81%
F-measure	-	-	65,17%	64.4%	57.6%	89.19%

CHAPTER 7 – Conclusions

The DJIA is one of the most important indexes in the financial market and it is a parameter for the global economy, as mentioned earlier in this work. Given the significance of the DJIA, the purpose of this work is mainly to predict surges of this index, i.e. moments when the share prices of the companies that compose the index are about to rise, based on news from internet. One big concern in dealing with the DJIA index is processing the extraordinary amount of data generated by 30 companies. To make matters more difficult, the purpose of this work is to identify a minute-by-minute solution, so the amount of data is even larger. Without the proper approach, it will be nearly impossible to identify an appropriate model. Even with a good model, the processing time requires special attention; otherwise, the model may remain theoretical. As of the conclusion of this research, I have not found any better results in literature.

This research integrates existing algorithms with a new alignment technique of Dow Jones index and news throughout the day, minute by minute. In order to make this model viable for practical use, it is important to first reduce the amount of data to process without losing any important information. In this research, it was achieved not only by reducing the amount of data but also drastically reducing the noise. The experiments showed that the system simply hangs or requires days to process the amount of data related to 30 companies and it also does not produce any reliable results. In this work, the first challenge was achieved by training with the news from the top most traded companies each minute along with an alignment technique that aligns the news from a certain period to each index, so the news not only affects the index at the published time, but also the indexes in the subsequent period. Some experiments were done in order to discover the best alignment window. Using an alignment window of 1 minute meant that the index was aligned to the news at its time of publication only. Although doing this improved the processing time since there was much less news aligned, the results were poor, which demonstrates that news requires some time to be processed by the analysts. Other experiments were done by increasing the alignment window, the period during which news affects the index. Experiments showed that using an alignment window of 3 and 5 minutes produces better results without increasing the process time too greatly. Above that range, the results get worse and

drastically affect the processing time. These experiments revealed that analysts are taking over 3 minutes to react after news is published. With this research, reactions can be delivered at the moment of news publication.

Ensuing experiments were aimed at identifying the optimal number of top traded companies by minute. Good results were achieved by using the top 5 and top 3 traded companies and, surprisingly, the top 1 were slightly better than others; this is a great discovery since it indicates reduced processing time. The processing time does not increase linearly with data size, so this finding was really important as it opens the chance for an online training system in future works. Some other experiments were done in order to find out the minimum training period in which the results remain accurate. The experiments showed that the proposed model requires 4 months as a minimum training period in order to produce good results; the results were not reliable using shorter periods. In other experiments the model showed to be very robust, by achieved best results in long periods of testing. All experiments were done using the SVM method. Later, other experiences were done using the Naïve Bayes method and the results were poor.

As mentioned earlier, this work can be improved by designing an online training system. Future works can also include new approaches to identify Surge, making it more sensitive to index rise. Other research can be made by improving BOW generated by the model. MIKOLOV et al. (2013a) developed the word2vec, and although it was not used in this work, it can be used in future works and it looks very promising. This work identifies two classes, the main one when the index is about to rise. Other research can be done in this area trying to identify more classes, or identify the Plunge, the moment at which the index is about to decline. The approach of this work can also be applied in previous research studies such as the ones mentioned in this work, as well as other important indexes to the economy such as NASDAQ which is composed of over 2700 companies. It is relevant to mention that the solution proposed in this work demands a large amount of data to be processed minute by minute, and today there are some technologies specifically designed to deal with it using distributed computing. This feature can be found in RapidMiner Radoop[4] and may be a good approach to take in future works. Other classification models can also be proposed such as neural

[4] https://rapidminer.com/products/radoop/

network and deep learning which also looks very promising. As we can see, there is room for improvement in this area that can reveal important achievements in future.

The results obtained surpass existing techniques today and in the researched literature on this topic. The model proved to be robust and efficient after adjusting the parameters throughout many experiments and demonstrated that the use of text mining techniques along with the correct strategy applied in the financial market is an alternative to be considered and contributes to the state of the art in this area of research.

References

ADRIÃO, M. C., 2009. *Um estudo de caso de previsão de tendência em uma série temporal financeira utilizando análise técnica*. MSc., dissertation, COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

AIZERMAN, A., BRAVERMAN, E. & ROZONER, L., 1964. *Theoretical foundations of the potential function method in pattern recognition learning*. Automation and Remote Control.

ARGENTINI, A. & BLANZIERI, E., 2010. *About Neighborhood Counting Measure Metric and Minimum Risk Metric*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 4(32), pp. 763-765.

BAEZA-YATES, R. and RIBEIRO-NETO, B., 1999. *Modern Information Retrieval*. Addison-Wesley Longman, Boston, MA.

BEATTIE, A., 2017. *When was the Dow Jones Industrial Average first calculated?* Disponível em: <https://www.investopedia.com/ask/answers/08/dow-jones-industrial-average-history.asp> Acesso em: 12/05/2018.

BECKMAN, M., 2015. *A KNN Undersampling Approach for Data Balancin.* Journal of Intelligent Learning Systems and Applications. v.7, pp. 104-116.

BECKMAN, M., 2017, PhD. Thesis: *Stock Price Change Prediction Using News Text Mining*. COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

BELLI, M. M., 2002. *Características patrimoniais e de resultados das empresas que compõe o índice Nasdaq-100*. MSc., dissertation, USP, São Paulo, SP, Brasil.

BENGIO, Y., 2009. *Learning Deep Architectures for AI*. Foundations and Trends in Machine Learning. 2.

BHANDARI, I., COLET, E., PARKER, J., *et al..*, 1997. *Advanced Scout: data mining and knowledge discovery in NBA data*. Data Mining and Knowledge Discovery, 1(1), pp. 121–125.

BOLLEN, J., 2003. *Text Operations*. Available at: http://www.cs.odu.edu/~jbollen/spring03_IR/cs695_lect6.pdf. Acessed in June 2018.

BOLLEN, J. & HUINA, M., 2011. Twitter mood as a stock market predictor. *Computer, 44*, p. 91–94.

BORIAH, S., CHANDOLA, V. & KUMAR, V., 2007. *Similarity Measures for Categorical Data: A Comparative Evaluation.* Minneapolis, s.n., pp. 243-254.

BOSER, B. E; GUYON, I. M and VAPNIK, V. N., 1992. *A training algorithm for optimal margin classifiers*. In: Proceedings Of The 5th Annual Acm Workshop On Computational Learning Theory, p. 144–152. ACM Press.

BROWN, S. J., GOETZMANN, W. N., KUMAR, A., 1998. The Dow Theory: William Peter Hamilton's Record Reconsidered. The Journal of Finance; v. LIII, nº4.

BRÜCHER, H., KNOLMAYER, G., and MITTERMAYER, M.-A., 2002. *Document Classification Methods for Organizing Explicit Knowledge*. Proceedings of the 3rd European Conference on Organizational Knowledge, Learning, and Capabilities, ALBA, Athens.

BUTLER, M. & KEŠELJ, V., 2009. *Financial forecasting using character n-gram analysis and readability scores of annual reports*. Advances in artificial intelligence, p. pp. 39–51.

CABENA, P; HADJINNIAN, P; STADLER, R. *et al..*, 1998. *Discovering Data Mining: From Concept to Implementation*. Prentice Hall. Inc. Upper Saddle River, NJ, USA.

CAMBRIA, E., SCHULLER, B., XIA, Y. & HAVASI, C., 2013. *New avenues in opinion mining and sentiment analysis*. IEEE Intelligent Systems. 28 (2), pp. 15-21.

CAMILO, C. O. and SILVA, J. C., 2009. *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas*. MSc., dissertation, Instituto de Informática da UFG, Goiás, GO, Brasil.

CARVALHO, J. L., 1975. *Uma nota sobre números-índices. R.* bras. Econ., Rio de Janeiro, 29 (1) :60-88 jan ./mar.

CHANG, C. & LIN, C., 2011. *LIBSVM : a library for support vector machines*. ACM Transactions on Intelligent Systems and Technology, pp. 2:27:1-27:27.

CHAPMAN, P; CLINTON, J; KERBER, R., *et al..*, 2000. *CRISP-DM 1.0. Step by Step Data Mining Guide*. CRISP-DM consortium.

CHATRATH, A., MIAO, H., RAMCHANDER, S. & VILLUPURAM, S., 2014. *Currency jumps, cojumps and the role of macro news*. Journal of International Money and Finance, 40, p. 42–62.

CHEN, M.S., HANAND, J. and Yu, P. S., 1996. *Data Mining: An Overview From a Database Perspective*. IEEE Transaction on Knowledge and Data Engijeering. v.8, nº 6.

CHINCHOR, N., 1992. *MUC-4 Evaluation Metrics*. Science Applications International Corporation 10260 Campus Point Drive, M/S A2-F San Diego, CA.

CHO, V., WUTHRICH, B., and ZHANG, J., 1998. *Text Processing for Classification. Journal of Computational Intelligence in Finance*. 26.

CIOS, K and KURGAN, L, 2005. *Trends in data mining and knowledge discovery*. In Pal, N and Jain, L (eds) Advanced Techniques in Knowledge Discovery and Data Mining. Springer, pp. 1–26.

CIOS, K. J; PEDRYCZ, W; SWINIARSKI, R. W; *et al..*, 2007. *Data Mining - A Knowledge Discovery Approach*. Springer.

CORTES, C., & PREGIBON, D., 1998. *Giga-mining*. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, R. Agrawal and P. Stolorz (eds.), Menlo Park, CA: AAAI Press, pp. 174–178.

CORTES, C. & VAPNIK, V., 1995. *Support-Vector Networks*. Machine Learning, 09, 20(3), pp. 273-297.

CRONE, S. & KOEPPEL, C., 2014. *Predicting Exchange Rates with Sentiment Indicators: An Empirical Evaluation using Text Mining and Multilayer Perceptrons.* s.l., s.n., pp. 114-121.

DAS, S. R. & CHEN, M. Y., 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science, 53*, p. 1375–1388.

DASARATHY, B., 1991. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques.* s.l.: IEEE Computer Society Press.

DEERWESTER, S., DUMAIS, S. T., FURMAS, G. W., *et al..*, 1990. *Indexing by latent semantic indexing*. Journal of the American Society for Information Science 41. v.6, John Wiley & Sons, Hoboken, NJ, pp. 391-407.

DE FARIA, E., EBECKEN, N. & ALBUQUERQUE, M., 2012. *A Methodology for Bovespa Index Forecasting Using Text Mining,* Rio de Janeiro: Federal University of Rio de Janeiro.

DHAKA, V., KAUSAR, M. & SINGH, S., 2013. *Web Crawler: A Review*. International Journal of Computer Applications. v. 63, nº. 2, pp. 31-36.

DIEWERT, W.E., 1987. *Index Numbers*. In J. Eatwell, M. Milgate, and P. Neuman (eds.) The New Palgrave: A Dictionrary of Economics (v.2, pp. 767-780). London: Macmillan Press.

DIEWERT, W.E., 1993. *The Early History of Price Index Research*. Essays in Index Number Theory, v.1, pp. 33-65. W.E. Diewert and A.O. Nakamura (eds.), Amsterdam: North-Holland.

DIETTERICH, T., 2002. *Machine Learning for Sequential Data: A Review.* London, Springer-Verlag, pp. 15-30.

DONOHO, D., 2000. *High-dimensional data analysis: The curses and blessings of dimensionality.* s.l., s.n.

DUDA, R., HART, P. & STORK, D., 2001. *Pattern Classification*. 2nd ed. New York: John Wiley & Sons Ltd., pp. 202-220.

EVERITT, B.S., GOURLAY, A.J., and KENDELL, R.E., 1971. *An attempt at validation of traditional psychiatric syndromes by cluster analysis*. British Journal of Psychiatry, 138, pp. 336–339.

FAMA, E. F., 1965. *The Behavior of Stock-Market Prices*. The Journal of Business 38, The University of Chicago Press, Chicago, IL, pp. 34-105.

FARIAS, T. A. and SANTOS, D. L., 2016. *Índices De Bolsas De Valores: Uma Revisão Teórico Quantitativa Das Metodologias De Construção De Índices Do Mercado Acionário.* Revista de Desenvolvimento Econômico – RDE - Ano XVIII – V. 2 - N. 34 – August - Salvador, BA – p. 481 – 522.

FAYYAD, U.M., DJORGOVSKI S.G., AND WEIR N. (1996). *Automating the analysis and cataloging of sky surveys*. *In Advances in Knowledge Discovery and Data Mining* U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), Menlo Park, CA: AAAI Press, pp. 471–493.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G. and SMYTH, P., *et al..*, 1996. *Advances in Knowledge Discovevy and Data Mining*. AAAI/MIT Press.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G. and SMYTH, P., 1996b, *Data mining to knowledge discovery: an overview*. In Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthurusamy, R (eds) Advances in Knowledge Discovery and Data Mining. AAAI Press, pp. 1–34.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G. and SMYTH, P., 1996c, *The KDD process for extracting useful knowledge from volumes of data*. Communications of the ACM 39(11), 27–34.

FAWCETT, T., and PROVOST, F., 1997. Adaptive fraud detection. *Data Mining and Knowledge Discovery*. 1(3), pp. 291–316.

FEHRER, R. & FEUERRIEGEL, S., 2016. *Improving Decision Analytics with Deep Learning: The Case of Financial Disclosures.* Istanbul, Turkey, s.n.

FIX, E. & HODGES, J., 1951. *Discriminatory analysis, nonparametric discrimination: Consistency properties, Technical Report 4*. Randolph Field, Texas: USAF School of Aviation Medicine.

FORMAN, G., 2003. *An extensive empirical study of feature selection metrics for text classification*. Journal of Machine Learning Research 3, pp. 1289-1305.

FONSECA, J. S.; MARTINS, G. A.; TOLEDO, G. L., 1985. *Construção e Uso de Números-Índices*. Estatística Aplicada. 2. ed. São Paulo: Atlas. Cap. 5. p. 157-201.

FORTUNA, E., 1997. *Mercado Financeiro: Produtos e Serviços*. 10 ed. Rio de Janeiro: Qualitymark.

FRAWLEY, W, PIATESKY-SHAPIRO, G and MATHEUS, C, 1991. *Knowledge discovery in databases: an overview*. AAAI/MIT Press,pp. 1–27.

GIDOFALVI, G., 2001. *Using News Articles to Predict Stock Price Movements*. University of California, San Diego: Department of Computer Science and Engineering.

GIMPEL, K., SCHNEIDER, N., O'CONNOR, B., *et al.*., 2011. *Part-of-speech tagging for twitter: annotation, features, and experiments.* Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 42-47.

GO, A., BHAYANI, R. & HUANG, L., 2009. *Twitter sentiment classification using distant supervision*, s.l.: Stanford University.

GROTH, S. & MUNTERMANN, J., 2011. *An intraday market risk management approach based on textual analysis*. Decision Support Systems, pp. 680-691.

HAGENAU, M., M., L., HEDWIG, M. & NEUMANN, D., 2012. *Automated news reading: Stock Price Prediction based on Financial News Using Context-Specific Features.* Maui, Hawaii, s.n., pp. 1040 - 1049.

HAN, J; KAMBER, M., 2006. *Data Mining: Concepts and Techniques*. 2nd ed. Elsevier.

HAND, D; MANNILA, H and SMYTH, P., 2001. *Principles of Data Mining*. MIT Press.

HARRIS, Z., 1954. *Distributional Structure*. Word, 10, p. 146–162.

HASTIE, T., TIBSHIRANI, R. & FRIEDMAN, J., 2003. *Model Assessment and Selection.* The Elements of Statistical Learning, Data Mining, Inference and Prediction. New York: Springer Series in Statistics, pp. 245-247.

HSU, C. W., CHANG, C. C. & J., L. C., 2003. *A practical guide to support vector classification,* Taipei, Taiwan: National Taiwan University.

HUANG, C., LIAO, J., YANG, D., *et al.*., 2010. *Realization of a news dissemination agent based on weighted association rules and text mining techniques*. Expert Systems with Applications, 37, p. 6409–6413.

HUTCHINS, W. J., 1995. *Machine translation: A brief history*. Concise history of the language sciences: from the Sumerians to the cognitivists, pp. 431-445.

JOACHIMS, T., 1998. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. Proceedings of the 10th European Conference on Machine Learning. Springer-Verlag. 137-142.

JOHNSON, B., 2010. *Algorithmic Trading & DMA: An Introduction to Direct Access Trading Strategies.* s.l.:4Myeloma Press.

KIM, Y., JEONG, S. & GHANI, I., 2014. Text Opinion Mining to Analyze News for Stock Market Prediction. *Int. J. Advance. Soft Comput. Appl., Vol. 6, No. 1.*

KLOSGEN, W. and ZYTKOW, J., 1996. *Knowledge discovery in databases terminology*. In Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthurusamy, R (eds) Advances in Knowledge Discovery and Data Mining. AAAI. Press, pp. 573–592.

KOHAVI, R. & QUINLAN, J., 2002. *Decision Tree Discovery*, chapter 16.1.3. Handbook of Data Mining and Knowledge Discovery. s.l.:Oxford University Press, pp. 267-276.

LAROSE, D. T., 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley and Sons, Inc.

LAVRENKO, V., SCHMILL, M., LAWRIE, D., *et al.*., 2000. *Language models for financial news recommendation*. In Proceedings of the ninth international conference on Information and knowledge management, pages 389-396. ACM.

LECUN, Y., BENGIO, Y. & HINTON, G., 2015. Deep Learning. Nature. vol 521, pp.436-444.

LEITE, H.P. and SANVICENTE, A. Z., 1994. *Índice Bovespa: um padrão para os investimentos brasileiros*. São Paulo: Atlas.

LENG, G., MCGINNITY, T. & PRASAD, G., 2005. *An approach for on-line extraction of fuzzy rules using a self-organising fuzzy neural network*. Fuzzy sets and systems 150 (2), pp. 211-243.

LI, F., 2010. *The information content of forward-looking statements in corporate filings - a naïve Bayesian machine learning approach*. Journal of Accounting Research, 48, p. 1049–1102.

LIOU, C., CHENG, C., LOU, J. & LIOU, D., 2014. *Autoencoder for Words*. Neurocomputing, Volume 139, p. 84–96.

LOPES, D. C, 2006. *Análise quantitativa da volatilidade entre os índices Dow Jones, Ibovespa e S&P500*. MSc., dissertation, Programa de Pós-Graduação em Economia da Faculdade de Ciências Econômicas da UFRS, Rio Grande do Sul, RS, Brasil.

LOVINS, J., 1968. *Development of a Stemming Algorithm*. Mechanical Translation and Computational Linguistics, Vol. 11, pp. 22-31.

LUGMAYR, A. & GOSSEN, G., 2012. *Evaluation of methods and techniques for language based sentiment analysis for DAX 30 stock exchange – a first concept of a 'LUGO' sentiment indicator.* s.l., s.n.

MADANI, O., 2003. *ABCs of Text Categorization*. Disponível em: http://classes.seattleu.edu/computer_science/css e470/Madani/ABCs.html. Acesso em: 10/06/2018.

MAHAJAN, A., DEY, L. & HAQUE, S. M., 2008. *Mining financial news for major events and their impacts on the market.* s.l., s.n., p. 423–426.

MCCUE, C., 2007. *Data Mining and Predictive Analysis - Intelligence Gathering and Crime Analysis*. Elsevier.

MELLAGI, A. and ISHIKAWA, S., 2007. *Mercado Financeiro e de Capitais*. 2nd ed. São Paulo: Atlas.

MERRILL, W. C., FOX, K. A., 1977. *Estatística Econômica: uma introdução*. Tradução de Alfredo Alves de Faria. 1 ed. São Paulo: Atlas.

MICROSOFT, 2017. Mining Model Content for Association Models (Analysis Services -Data Mining). Disponível em: <https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/mining-model-content-for-association-models-analysis-services-data-mining?view=sql-analysis-services-2017>. Accessed in May 2018.

MIERSWA, I., WURST, M., KLINKENBERG, R., *et al.*., 2006. *YALE: Rapid Prototyping for Complex Data Mining Tasks.* Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining; pp. 935-940.

MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. AND DEAN, J., 2013a. *Distributed representations of words and phrases and their compositionality*. Advances in neural information processing systems, pages 3111–3119.

MIKOLOV, T., LE, Q.V. AND SUTSKEVER, I., 2013b. *Exploiting similarities among languages for machine translation*. arXiv preprint arXiv:1309.4168.

MIKOLOV, T., CHEN, K, CORRADO, G. AND DEAN, J., 2013c. *Efficient estimation of word representations in vector space*. In arXiv preprint arXiv:1301.3781.

MILLER, G. A., 1995. WordNet: *A lexical database for English*. Communications of the ACM, 38, pp. 39-41.

MINER, G., ELDER, J., HILL, T., *et al.*., 2014. *Pratical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* s.l.: Academic Press / Elsevier.

MITTERMAYER , M., 2004. *Forecasting intraday stock price trends with text mining techniques*. In System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on, pages 10pp. IEEE.

NETO, A. A., 2010. *Mercado Financeiro*. 9. ed. São Paulo: Atlas.

NG, A. *et al.*., 2010-2012. *Machine Learning.* Available at: http://openclassroom.stanford.edu/MainFolder/DocumentPage.php?course=MachineLearning&doc=exercises/ex8/ex8.html. Accessed at: 20/05/2018.

OLIVEIRA, R. R; CARVALHO, C. L., 2008. *Algoritmos de agrupamento e suas aplicações*. Technical report. UFG, Goiás, GO, Brasil.

OLSON, D. L. and DELEN, D., 2008. *Advanced Data Mining Techniques*. Springer, 1st ed.

PAI, P.-F. and C.-S. LIN, 2005. *A hybrid ARIMA and support vector machines model in stock price forecasting*. Omega. 33(6): 497-505.

PANG, B., LEE, L. & VAITHYANATHAN, S., 2002. *Thumbs up? Sentiment classification using machine learning techniques.* s.l., s.n., pp. 79--86.

PERAMUNETILLEKE, D. & WONG, R. K., 2002. *Currency exchange rate forecasting from news headlines*. Australian Computer Science Communications, 24, p. 131–139.

PINHEIRO, J. L., 2009. *Mercado de Capitais: Fundamentos e Técnicas*. 5 ed. São Paulo: Atlas.

POON, S.; GRANGER, C. W. J., 2003. *Forecasting Volatility in Financial Markets: a review*. Journal of Economic Literature, v. 41, n. 2, p. 478-539.

PORTER, M., 1980. *An Algorithm for Suffix Stripping.* Program, Vol. 14, No. 3, pp. 130-137.

PRASS, F. S., 2004. *KKD: Processo de descoberta de conhecimento em bancos de dados*. Grupo de Interesse Em Engenharia de Software, Florianópolis, v. 1, p. 10-14.

RADOVANOVIĆ, M., NANOPOULOS, A. AND IVANOVIĆ, M, 2010. *On the existence of obstinate results in vector space models*. Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pages 186–193. ACM.

REINARTZ, T, 2002. *Stages of the discovery process*. In Klosgen, W and Zytkow, J (eds) Handbook of Data Mining and Knowledge Discovery. Oxford University Press, pp. 185–192.

REZENDE, S. O., 2005. *Mineração de Dados*. XXV Congresso da Sociedade Brasileira de Computação.

SCHUMAKER, R. P. & CHEN, H., 2009. *Textual analysis of stock market prediction using breaking financial news: The AZFin text system*. ACM Transactions in Information System (TOIS), 27(2), pp. 1-19.

SCHUMAKER, R. P., ZHANG, Y., HUANG, C. & CHEN, H., 2012. *Evaluating sentiment in financial news articles*. Decision Support Systems.

SCHÜNKE, M. A. and DIAS, L.T., 2013. *Analise De Modelos De Predição Baseado Em Informações*. ISBN: 978-972-8939-95-3 © 2013 IADIS.

SEBASTIANI, F., 1999. *A Tutorial on Automated Text Categorisation*. A. Amandi and A. Zunino (eds.), Proceedings of the 1st Argentinean Symposium on Artificial Intelligence, ASAI, Buenos Aires, pp. 7- 35.

SHEARER C., 2000. *The CRISP-DM model: the new blueprint for data mining*. J Data Warehousing; 5:13—22.

SIDOROV, G., VELASQUEZ, F., STAMATATOS, E. *et al.*., 2014a. *Syntactic N-grams as machine learning features for natural language processing*. Expert Systems with Applications, Vol. 41, Issue 3, pp. 853-860.

SIDOROV, G., GELBUKH, A., GOMEZ-ADORNO *et al.*., 2014b. *Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model*. Computación y Sistemas, 18(3), p. 491–504.

SIERING, M., 2012. *"Boom" or "Ruin" – Does it Make a Difference? Using Text Mining and Sentiment Analysis to Support Intraday Investment Decisions*. s.l., s.n., pp. 1051-1059.

SILBERSCHATZ, A., STONEBRAKER, M. and ULLMAN, J.D., 1995. *Database Research: Achievements and Opportunities into the 21st Century*. Report NSF Workshop Futuve of Database Systems Research.

SOKOLOVA, M. & LAPALME, G., 2009. *A systematic analysis of performance measures for classification tasks.* Information Processing and Management, 45, pp. 427-437.

SONI, A., VAN ECK, N. J. & KAYMAK, U., 2007. *Prediction of stock price movements based on concept map information.* s.l., s.n., pp. 205-211.

SOYLAND, C., 2015. *Interday news-based prediction of stock prices and trading volume*. Msc., dissertation, Department of Applied Mechanics of Chalmers University of Technology. Göteborg, Sweden.

TETLOCK, P. C., SAAR-TSECHANSKY, M. & MACSKASSY, S., 2008. *More than words: Quantifying language to measure firms' fundamentals*. The Journal of Finance, 63, p. 1437–1467.

THOMAS, J. D. and SYCARA, K., 2002. *Integrating Genetic Algorithms and Text Learning for Financial Prediction*. Genetic and Evolutionary Computation Conference (GECCO). Las Vegas, NV.

VAKEEL, K. & SHUBHAMOY, D., 2014. *Impact of News Articles on Stock Prices: An Analysis Using Machine Learning.* New York, ACM, pp. 1-4.

VAN RIJSBERGEN, C., 1979. *Information Retrieval.* 2nd ed. Massachusets: Butterworths.

VAPNIK, V. & LERNER, A., 1963. *Pattern recognition using generalized portrait method*. Automation and Remote Control, Volume 24, p. 774–780.

VARTANIAN , P. R., 2012. *Effects of the Dow Jones index, commodities and exchange rate on Ibovespa: an analysis of contagion effects*. Rev. adm. contemp. vol.16 no.4 Curitiba July/Aug.

VIANA, O, 2009. *História da Análise Técnica*. Think Finance. Available at: http://www.monitorinvestimentos.com.br/aprendizado.php?id_aprendizado=67. Accessed in December 2017.

VU, T., CHANG, S., HA, Q. & COLLIER, N., 2012. *An experiment in integrating sentiment features for tech stock prediction in twitter.* Mumbai, India, The COLING Organizing Committee, pp. 23-38.

WANG, J, 2005. *Encyclopedia of Data Warehousing and Mining*. Idea Group Reference.

WEDEL, M., and KAMAKURA, W.A., 1998. *Market Segmentation: Conceptual and Methodological Foundations.* Boston, MA: Kluwer.

WEISS, G. & PROVOST, F., 2001. *The Effect of Class Distribution on Classifier Learning: An Empirical Study*. s.l.: Technical Report MLTR-43, Dept. of Computer Science, Rutgers University.

WEISS, R., 2000. *Mercado Acionário Brasileiro: proposta de novos índices para ampliar a abrangência e a capacidade de diagnóstico*. Revista Bndes, Rio de Janeiro, v.7, n. 14, p.29-54.

WEISS, S. M., INDURKHYA, N. & ZHANG, T., 2010. *Fundamentals of Predictive Text Mining.* s.l.:Springer Publishing Company, Incorporated.

WHITE, T., 2009. *Hadoop: The Definitive Guide, 1st Ed.*. s.l.:O'Reilly.

WIEDERHOLD, G, 1996. *Foreword: on the barriers and future of knowledge discovery*. In Fayyad, U, Piatetsky-Shapiro, G, Smyth, P and Uthurusamy, R (eds) Advances in Knowledge Discovery and Data Mining. AAAI Press.

WILSON, D. & MARTINEZ, T., 1997. *Improved Heterogeneous Distance Functions*. Journal of Artificial Intelligence Research, Volume 6, pp. 1-34.

WILSON, T. & HOFFMANN, P., 2005. *OpinionFinder: a system for subjectivity analysis.* Vancouver, Canada, s.n.

WONG, F., LIU, Z. & CHIANG, M., 2014. *Stock Market Prediction from WSJ: Text Mining via Sparse Matrix Factorization.* s.l., s.n.

WRIGHT, S., SMITHERS, A., WARBURTON, P., *et al.*., 2011. *Practica History of Financial Markets*. Edinburgh Business School, United Kingdom.

WU, X., KUMAR, V., QUINLAN, J. R. *et al..*, 2008. *Top 10 algorithms in data mining*. Knowl Inf Syst. 14:1–37.

WÜTHRICH, B., 1995. *Probabilistic Knowledge Bases*. IEEE Transactions of Knowledge and Data Engineering 7(5):691-698.

WÜTHRICH, B., 1997. *Probabilistic Knowledge Bases.ht*. Journal of Intelligent Systems in Accounting Finance and Management 6:269-277.

WÜTHRICH, B., PERMUNETILLEKE, D., LEUNG, S., *et al..*, 1998. *Daily prediction of major stock indices from textual www data*. In Proceedings of the 4th international conference on knowledge discovery and data mining, KDD98.

YANG, S. Y., SONG, Q., MO, S. Y. K., *et al..*, 2015. *The Impact of Abnormal News Sentiment on Financial Markets.* Journal of Business and Economics, Vol. 6, No. 10 , pp. 1682-1694.

YANG, Y. and LIU, X., 1999. *A Re-Examination of Text Categorization Methods*. Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, NY, pp. 42-49.

YU, Y., DUAN, W. & CAO, Q., 2013. The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*.

ZHAI, C. & MASSUNG, S., 2016. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining.* s.l.:ACM.

ZHAI, Y., HSU, A. & HALGAMUGE, S., 2007. *Combining news and technical indicators in daily stock price trends prediction.* Nanjing, China, Springer-Verlag, p. 1087–1096.

ZOTTE, J. L., COTRIN, A. M., SANTOS, A. L., 2012. *A evolução da contabilidade e o mercado de trabalho para o contabilista*. Revista Conteúdo, Capivari, v.2, n.1, jan./jul.– ISSN 1807-9539.

Appendix A

Table 26 - Main aspects of TMFP methodology over the years.

Reference / Year	Source of news	No. of items	Market/ Index/ Exchange	Time-Frame / Alignment offset	Period of news collection	Number of months	Number of Classes / Target prediction	Feature Selection / Representation	Dimensionality Reduction	Learning Algorithm	Training vs. testing	Sliding window	Sentiment Analysis	Semantics	Syntax	Balanced Data
(Hastie *et al.*., 2003)	WSJ, FT, Reuters, Down Jones, Bloomberg	-	Stocks from DJIA, Nikkei,FTSE, HSI, STI	Daily	6/Dec/1997-6/Mar/1998	3	3	BOW/ Binary	Expert dictionary	k-NN, ANN, Naïve Bayes, Rule-based	100 days vs. 1 day	Y	N	Y	N	-
(Lavrenko, Schmill, et al., Language Models for Financial News Recommendation 2000)	Yahoo! Finance	38000	Stocks from NYSE, NASDAQ	Intraday/ 1 hour	15/Oct/1999-10/Feb/2000	4	5	BOW	Bayesian Language Models	Bayesian Language Models	3 months vs. 40 days	N	N	N	N	N
(Peramunetilleke & Wong, 2002)	HFDF93 via www.olsen.ch	960	ForEx USD-DEM, USD-JYP	Intraday/ 3 hours	22/Sep/1993-27/Sep/1993	0	3	BOW/ Boolean, TF-IDF, TF-CDF	Set of keywords	Decision tree and rules	22/Sep 12:00-27/Sep 09:00 vs. 9:00-10:00 on 27 Sep.	Y	N	Y	N	Y
(Fung, et al., 2003)	Reuters Market 3000	600000	33 stocks from HIS	Daily	1/Oct/2002–30/Apr/2003	7	2	BOW/ TF-IDF	Stemming, stop words	SVM	6 months vs. last month	N	N	N	N	-
(Gidófalvi & Elkan, 2003)	Yahoo! Finance	6300	12 stocks from NASDAQ	Intraday/ -20 to 20 minutes	14/Nov/1999-11/Feb/2000	3	3	BOW, Wittenbell smoothing method	Stemming, stop words, highest 1000 words with mutual	Naïve Bayes	4650 vs. 1650	N	N	N	N	-

									information							
(Mittermayer, 2004)	PRSNews-Wire	7002	Stocksfrom NYSE and NASDAQ	Daily	Year 2002	12	3	BOW/ TF-IDF	Selected 1000 terms	SVM	400 vs. 6602 examples	N	N	N	N	Y
(Werner & Murray, 2004)	Yahoo! Finance, WSJ Raging Bull	1500000	DJIA stocks	Daily	Year 2000	12	3	BOW/ Binary	Minimum information criterion (top 1000 words)	Naïve Bayes, SVM	1000 messages vs. the rest	N	N	N	N	N
(Das & Chen, 2007)	Message boards	145110	Stocks of 24 tech-sectors fromMSH	Daily	Jul/2001-Aug/2001	2	Aggregate Sentiment index	BOW/ Triplets, discrete values for each classifier	Predefined dictionaries	Combinatory algorithms	1000 messages vs. the rest	N	Y	Y	Y	-
(Rachlin, et al., 2007)	forbes.com, reuters.com	-	5 stocks from NASDAQ	Daily	7/Feb/2006-7/May/2006	3	5	BOW/ common financial values, TF, Boolean, Extractor SW output	Automatic extraction of most influential keywords	C4.5 decision tree	-	N	N	N	N	-
(Soni, et al., 2007)	Financial Times Intelligence	3493	Stocks of 11 oil and gas companies	Daily	1/Jan/1995-15/May/2006	136	2	Visual coordinates	Thesaurus using term extraction tool	SVM w/ linear kernel	80% vs. 20%	N	N	Y	N	Y
(Zhai, et al., 2007)	Australian Financial Review	216	BHP Billiton Ltd. from ASX	Daily	1/Mar/2005-31/May/2006	14	2	BOW/ Binary, TF-IDF	Top 30 higher level concepts using WordNet	SVM w/ RBF and polynomial kernel	12 months vs. 2 months	N	N	Y	N	-
(Mahajan, et al., 2008)	-	700	Stocks from SENSEX	Daily	Aug/2005-Apr/2008	33	Categorical	LDA/ Binary	Extraction of twenty-five topics	Stacked classifier	Aug/2005-Dec/2007 vs. Jan/2008-Apr/2008	N	N	Y	N	-
(Tetlock, et al., 2008)	WSJ, Down Jones news from Factiva service	350000	Firms future cash flows from S&P 500	Daily	1980-2004	300	Regression	BOW for negative words/ Frequency divided by total words	Harvard-IV-4 psychosocial dictionary	OLS regression	33 trading days prior to an earnings announcement	Y	Y	Y	N	NA
(Butler & Kešelj, 2009)	Reports from companies'	-	1 Year market drift of stocks	Yearly	2003-2008	72	2	BOW / Character n-grams, n-gram	Minimum occurrence per document	CNG distance, SVM	x-1 and x-2 and all vectors vs. testing	Y	N	N	Y	-

	websites							frequency			year					
(Schumaker & Chen, 2009)	Yahoo! Finance	2800	S&P 500 stocks	Intraday/ 20 minutes	26/Oct/2005-28/Nov/2005	1	Categorical discrete numeric	BOW / noun phrases, named entities/ Binary	Minimum occurrence per document	SVM	-	N	N	Y	Y	-
(Huang, et al., 2010)	Leading electronic newspapers in Taiwan	12830	TAIEX stocks	Daily	Jun/2005-Nov/2005	6	Significant degree assignment	Simultaneous terms, ordered pairs / Weighted on the index fall/rise	Synonyms replacement	Weighted association rules	Jun/2005-Oct/2005 vs. Nov/2005	N	N	Y	Y	-
(Li, 2010)	Management discussion and Analysis section from SEC Edgar website	140000	(1) Index (2) Quarterly earnings and cash flows (3) Stock returns	Yearly	1994-2007	168	4	BOW, Tone and content / Binary, Dictionary value	Pre-defined dictionaries	Naïve Bayes and dictionary-based	30000 randomly vs. itself and the rest	N	N	N	N	N
(Bollen & Huina, 2011)	Twitter	9853498	DJIA stocks	Daily	28/Feb/2008-19/Dec/2008	10	Regression	Opinion finder	Opinion finder	Self organizing fuzzy NN	28/Feb-28/Nov vs. 1/Dec-19/Dec	N	Y	NA	NA	-
(Groth & Muntermann, 2011)	Adhoc corporate disclosures	423	Stock Market Risk	Intraday/ 15 minutes	1/Aug/2003-31/Jul/2005	24	2	BOW/ TF-IDF	Information Gain and Chi-Squared	Naïve Bayes, k-NN, ANN, SVM	Stratified cross validation	N	N	N	N	N
(De Faria, et al., 2012)	Macro-economic, financial news, social media	174993	Blue chips stocks from BOVESPA	Daily	23/Feb/2010-30/Jun/2011	17	3	BOW/ TF, TF-IDF, TF-CDF	Keep titles only, stop words, stemming, small dictionary	SVM, MLP, RBF, Naïve Bayes	Cross validation	N	N	Y	N	Y
(Hagenau, et al., 2012)	DGAP, EuroAdhoc	14348	Company specific stock	Daily	1997-2011	180	2	BOW/ noun phrases, n-grams / TF-IDF	Chi-Squared + Bi-normal separation for exogenous-feedback.	SVM linear, SVR	-	N	N	Y	Y	N

(Lugmayr & Gossen, 2012)	Broker newsletter	-	Stocks from DAX 30	Intraday/ open, midday, close	-	0	3	BOW/ Sentiment value	Stemming	SVM	-	N	Y	Y	N	-
(Schumaker, et al., 2012)	Yahoo! Finance	2802	Stocks from S&P 500	Intraday/ 20 minutes	26/Oct/2005-28/Nov/2005	1	Regression	OpinionFinder overall tone and polarity / Binary	Minimum occurrence per document	SVR	-	N	Y	Y	N	-
(Siering, 2012)	Down Jones News	11518	DAX blue chips stocks	Intraday/ 15 minutes	06/Apr/2006-08/Apr/2008	24	3	BOW / TF-IDF	Porter stemming, stop words, Info Gain	SVM w/ linear kernel	Cross validation	N	Y	N	N	-
(Vu, et al., 2012)	Twitter	5001460	NASDAQ Stocks AAPL, GOOG, MSFT, AMZN	Daily	1/Apr/2011-31/May/2011 online test: 8/Sep/2012-26/Sep/2012	12	2	Daily number of pos/neg on TST+ emoticon lexicon + PMI / Real number of pos/neg and bullish/bearish anchor words	Pre-defined company related keywords, Named Entity Recognition.	C4.5 decision tree	Previous day vs. current day	Y	Y	Y	Y	-
(Jin, et al., 2013)	General news from Bloomberg	361782	ForEx	Daily	2012	12	Regression	LDA / Each article's topic distribution	Manual top identification by manually aligning news articles with currency fluctuations.	Linear regression model	Previous day vs. a given day	Y	Y	Y	N	-
(Makrehchi, et al., 2013)	Twitter	30M	S&P 500 index	Daily	27/Mar/2012-13/Jul/2012	2.5	2	BOW / Binary	Mood word list	Rocchio	Cross validation	N	Y	Y	N	Y
(Yu, et al., 2013)	Blogs, forums, news, micro blogs (e.g., Twitter)	52746	AR and CAR from stocks of 824 firms	Daily	1/Jul/2011-30/Sep/2011	3	2	BOW / Binary	-	Naïve Bayes	-	N	Y	Y	N	-

(Crone & Koeppel, 2014)	Reuters MarketPsych	783	ForEx AUD-USD	Daily	4/Sep/2009-4/Sep/2012	36	2	14 built-in sentiment indicators from Reuters	NA	MLP	Cross validation	N	Y	N	N	-
(Kim, et al., 2014)	Naver.com	78216	KOSPI, stocks of 2 media firms	Daily	2011	12	2	BOW / TF	Stop words, automated sentiment dictionary	-	01/01/2011-31/Jul/2011 vs. 1/Aug/2011-31/Dez/2011	N	Y	Y	N	-
(Vakeel & Shubhamoy, 2014)	Times of India, Economic Times	3253	SENSEX Stocks	Pre/Post Election	17/Feb/2014-13/June/2014	4	2	BOW / TF-IDF, n-grams	Information Gain	SVM	80% vs. 20%	Y	N	N	Y	-
(Wong, et al., 2014)	WSJ	-	Stocks from DJIA, S&P 500, NASDAQ	Daily	1/Jan/2008-30/Sep/2013	69	2	Sparse Matrix Factorization + ADMM	Sparse Matrix Factorization + ADMM	Sparse Matrix Factorization + ADMM	2008-2011 vs. validation: 2012 vs. test: 2013	N	N	N	N	-
(Nassirtoussi, et al., 2015)	MarketWatch .com & others	6096	ForEx EUR/USD	Intraday/ 1 hour	2008-2011	48	2	BOW/ TF-IDF, SumScore weighting	Synchronous Target Feature-Reduction, WordNet	SVM, k-NN, Naïve Bayes	Several tests with training data proportion >=0.99	Y	Y	Y	Y	-
(Yang, et al., 2015)	Northern Light business news	678378	S&P 500 index	Daily	13/Jul/2012-16/Oct/2014	27	Regression	BOW/ Daily sentiment score from dictionary	Stemming, stop words	Regression with abnormal sentiment scores	Training = Test	N	Y	Y	N	-
(Fehrer & Feuerriegel, 2016)	Adhoc reports from DGAP	8359	Stocks from German firms	Daily	Jan/2004-Jun/2011	90	3	Neural Networks - Recursive auto encoders	Neural Networks - Recursive auto encoders	Neural Networks - Recursive auto encoders	80% vs. 20%	N	Y	N	Y	-
This current work	**Yahoo! Finance, Google Finance**	**393983**	**DJIA stocks**	**Intraday, 1 minute**	**03/Jan/2012-29/Nov/2013**	**9**	**2**	**BOW/ TF-IDF, n-grams**	**Chi Square, stop words, min/max occurrence per document**	**LIBSVM w/ RBF kernel**	**Last 7 months vs. 1 week, then repeating for 3 months**	**Y**	**N**	**N**	**Y**	**Y**

Appendix B

Table 27 - Glossary of terms and acronyms.

Term	Description
ADMM	Alternating Direction Method of Multipliers is an optimization algorithm suitable for non-convex problems.
AMH	Adaptive Market Hypothesis
AR	Abnormal return, a return of investment above the average and expectations.
Asset	An asset is a resource with economic value that an individual, corporation or country owns or controls with the expectation that it will provide future benefit.
ATS	Automated Trading System
Backtesting	The process of testing a trading strategy or algorithm on historical data to ensure its viability before to apply it in a real investment scenario.
Bearish	Represents a wish or trend for a fall in the price of an asset or market.
Blue Chips	A stock from a reputed and stable company.
BOW	Bag of Words
BOVESPA	São Paulo Exchange
Bullish	Represents a wish or trend for a rise in the price of an asset or market.
CAPM	Capital Asset Price Model
CAR	Cumulative abnormal return
CATS	Cascading Aggregation for Time Series
CR	Cumulative return
Daily	Trading operations with one day of duration.
DAX	Index with the 30 major companies from Germany.
DGAP	German Society for Ad Hoc Publicity
DJIA	Down Jones Industrial Average, is a stock market index that represents 30 large publicly owned companies based in the United States.
ENET	Elastic-net logistic regression
Equity Market	Same as Stock Market
Exchange	A highly-organized market where tradable securities, commodities, foreign exchange, futures, and options contracts are sold and bought.
Financial Instrument	Financial instruments are assets that can be traded.
FA	Fundamental Analysis
ForEx	Foreign Exchange, it is the financial market where currencies are traded.
FT	Financial Times
FTSE	Financial Times Stock Exchange 100 is an index calculated from 100 companies listed on the London Stock Exchange (LSE).

Future Market	A market where the long-term contracts are traded. The parties agreed in the present, a buy and sell price of an asset to be traded in the future.
GA	Genetic Algorithm
Hyperplane	In geometry, it is a representation of *n-1* dimension, being *n* the current number of available dimensions. For example, a 1-dimensional line is the hyperplane in 2 dimension spaces, a 2-dimensional plane is the hyperplane in 3 dimension spaces, and so on.
Hyperparameter	A parameter provided by the user to be applied by the pre-processing and machine learning algorithms.
HKEx	Hong Kong Stock Exchange
HSI	Hang Seng Index, is constituted with the 50 companies from HKEx.
Intraday	Trading operations with less than one day of duration.
KOSPI	Korea Composite Stock Price Index
LDA	Latent Dirichl*et al.*location
LSE	London Stock Exchange
MLP	Multi-Layer Perceptron neural network
NASDAQ	National Association of Securities Dealers Automated Quotations, is an American stock exchange, and concentrates the trading of the most important technology companies in the world.
Nikkei 225	The main stock index from Tokyo Stock Exchange (TSE).
NYSE	New York Stock Exchange, is the largest exchange in the world by volume and market capitalization.
Order, Order execution	The command to buy or sell financial instruments sent to an exchange.
PMI	Pointwise Mutual Information
Portfolio	A collection of investments held by an investment company or individual.
POS	Part of Speech, it is used to capture a sentence's syntactic aspects.
RBF	Radial Basis Function
Roundtrip	The entire operation of to buy and sell a share or other security.
S&P 500	Standard & Poor Index, which aggregates 500 American companies.
SENSEX 30	The stock index from the Bombay Stock Exchange (BSE).
Share	Share is a portion of a stock.
Security	Same as financial instrument, but its legal definition varies according the jurisdiction.
STI	FTSE Straight Times Index is constituted from the top 30 companies from Singapore Exchange (SGX).
Stock	A stock is a type of financial instrument that grants ownership in a corporation and gives the right to claim for part of the corporation's assets and earnings.

Stock Market	A group of buyers and sellers with the common interest to trade shares of stock.
SVM	Support Vector Machine
SVR	Support Vector Machine for Regression
TA	Technical Analysis
TF	Term frequency, number of occurrence of a term in a document, divided by the total number of terms in a document.
TF-IDF	Term frequency-inverse document frequency
TMFP	Text mining applied to financial market prediction.
TSE	Tokio Stock Exchange
TST	Twiter Sentiment Tool
UTC	Coordinated Universal Time
WSJ	Wall Street Journal