*Department*
*of Business and Management*

*Bachelor's Degree in Management and Computer Science*
*Course of Artificial Intelligence and Machine Learning*

# *Exploring Sentiment Analysis and its Applications:*

## *Predicting Stock Trends by Analysing Emotions and Opinions of Financial News*

*Supervisor:*
*Prof. Giuseppe*
*Francesco Italiano*

*Candidate:*
*Jona Cardamone*
*ID No: 249121*

# *Index*

## List of Abbreviations

| Abbreviation | Definition |
|---|---|
| SA | Sentiment Analysis |
| NLP | Natural Language Processing |
| ML | Machine Learning |
| DL | Deep Learning |
| AI | Artificial Intelligence |
| NER | Named Entity Recognition |
| SVM | Support Vector Machine |
| CNN | Convolutional Neural Network |
| RNN | Recurrent Neural Network |
| ABSA | Aspect-based Sentiment Analysis |
| IBSA | Intent-based Sentiment Analysis |
| EDSA | Emotion detection Sentiment Analysis |
| LSTM | Long-Short Term Memory |
| GRU | Gated Recurrent Units |
| POS | Part-of-Speech |

# *Acknowledgements*

I would like to take this opportunity to express my deepest gratitude to all those who have contributed to the successful completion of my work.

First and foremost, I would like to express my sincere appreciation to my supervisor, Professor Giuseppe Francesco Italiano, for his guidance, encouragement, and for the passion he managed to convey to me for this subject.

I would also like to express my heartfelt thanks to my family, who have been my constant source of inspiration and support. My mother, Sanja, and my father, Angelo, have always believed in me and provided me support and encouragement, even when things were tough. Their love, guidance, and unwavering commitment have been the foundation of my success. I would like to extend my gratitude to my brother Luca for his constant encouragement, humour, and moral support. His friendship and guidance have been invaluable in shaping my outlook on life and enabling me to achieve my goals.

Lastly, I would like to express my appreciation to my friends Alessandro, Simone and Francesca for their consistent support and encouragement throughout this journey. Their insightful advices and dedicated efforts have been critical in enabling me to stay focused and motivated throughout the project.

Thank you very much.

# *Introduction*

The advent of Web 2.0 brought several innovations from the aspect of online interactions. With social networks, blogs and forums, users started to share their opinions, complaints about products they bought, or simply express their points of view on any topic. When users interact on the web, they create data. There are about 5 billion active internet users worldwide (fig. 1) (62% of the world's population) (Zippia 2023), and in addition to that, social networks are expanding fast, resulting in an exponential data growth. Data, as explained in the next chapters, is of vital importance for many tools such as Sentiment Analysis (SA).

Opinions are very important for all human activities, and they can influence our behaviour and actions. For example, if a person wants to buy a new phone but is not sure which brand is more suitable for them, they can go online, find relevant information by reading other clients' reviews and make an informed decision that takes into account various opinions and different points of view. Opinions and the other concepts like sentiments and emotions are the focal subject of study of SA, also known as opinion mining. The aim of this thesis is to provide an in-depth analysis of this incredible tool and to go over its real-world applications with particular emphasis on its utilization in finance.

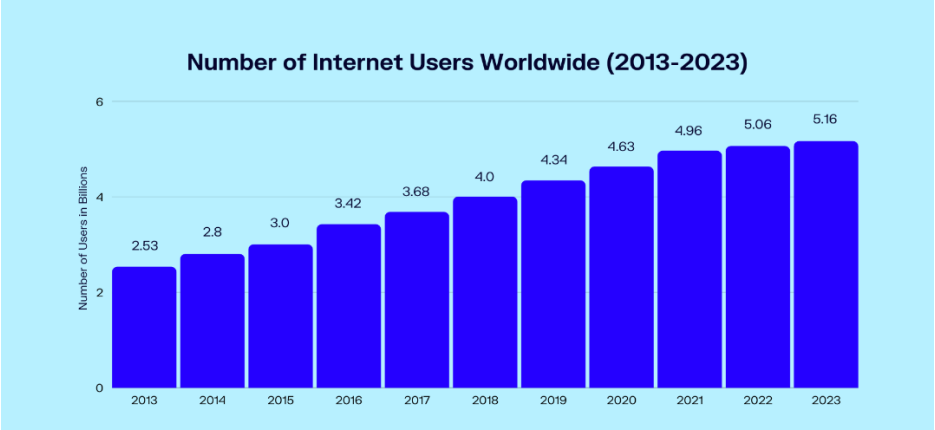


*Figure 1 Source: OBERLO, "How many people use the internet?"*

# 1 Sentiment Analysis: an overview

SA is a subfield of Natural Language Processing (NLP), and it is the field of study that analyses people's opinions, sentiments and emotions contained in a text (Natural Language Prcocessing - Sentiment Analysis 2021). Nowadays, the application spectrum for this technology is wide and can be utilized in many sectors such as marketing, politics, tourism and finance (which I will cover more in depth in Chapter 2).

SA tools enable companies to collect insights from disorganized and unstructured data that originates from online sources such as mails, blogs, web interactions, social media posts, forums, and comments to determine and classify the liking about a product, service, or concept. In customer service, it can assist organizations to promptly discover and address customers issues and concerns. In politics, it might help to analyse and evaluate the public opinion on political issues and on political campaigns. In social media monitoring, SA may help companies to determine how people feel about a certain topic or brand while in finance it can be used to predict stock movements after having analysed financial news about a particular company (Chapter 3).

SA has evolved notably in recent years, and it is now able to accurately determine the sentiment behind huge quantities of text using complex machine learning (ML), deep learning (DL) and artificial intelligence (AI) algorithms. Nevertheless, SA is still growing, and there are many challenges and limitations, which I will go over and discuss in Chapter 1.5, that researchers are still trying to solve.

## *1.1 History of Sentiment Analysis*

The history of SA can be divided into four important periods, each characterized by significant methodological and technical developments. The early research on SA dates back to the 60s with the publication named "The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information" by Stone, Philip J. and Earl B. Hunt (Stone 1962). This study focused on establishing ways to automatically determine the sentiment represented in text, using rule-based systems and lexicons, also known as sentiment dictionaries, which were created manually with words linked to their sentiment polarity, either positive, negative or neutral. However, these early attempts had limited accuracy because of the restricted computing capabilities available at the time, the absence of big datasets and the time-consuming as well as substantial manual effort needed to update the sentiment.

In the 80s – 90s, the discovery of Natural Language Processing (NLP) and Named Entity Recognition (NER) brought several opportunities for SA. Researchers started trying new methods to work with SA, such as lexical analysis, which exploited the knowledge of the relationships between words to deduce the sentiment of a text. During this time and thanks to these achievements, SA began to be applied in a variety of fields, such as customer service and market research.

Between the 2000s and 2010s, researchers were able to build more accurate SA models mainly thanks to the large quantity of data available but also to the development of ML algorithms such as Support Vector Machines (SVMs) and decision trees. ML-based approaches such as deep learning and supervised learning, enabled automatic feature extraction and modelling of complex interactions between words and feelings. The growing usage of social media, the Internet and the availability of enormous quantities of customer feedback data also contributed to the expansion of SA during this period. Businesses and organizations started to use SA to understand customers' needs and preferences, monitor their brand reputation, and make smarter business decisions. By this time, we started to see a transition of SA from being a manual rule – based task to an automated process that can manage a high volume and complexity of text data.

The deep learning revolution had a huge influence on SA. Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are much more powerful than the classic ML models on a wide range of NLP tasks, including SA. One significant feature of Deep Learning models is their capacity to automatically learn to "understand" a language, enabling them to capture the complex relationships between words and sentiments in a text. Furthermore, deep learning models permits to handle huge amounts of text input and can work

accurately for certain subjects and languages, allowing researchers to build deep learning models with high accuracy and resilience.

Another significant achievement during this period was the use of pre-trained language models, such as BERT and GPT (Ajayi 2020) (the model used by the famous chat bot “ChatGPT”). These models, which have been pre-trained on huge amounts of text data, may be fine - tuned for SA utilizing smaller amounts of labelled data. This has led to considerable improvements in the accuracy and efficiency of SA models, allowing them to do it on an even bigger scale. SA became widely utilized in a range of sectors and applications, and the discipline continues to expand, with current research focusing on enhancing the accuracy and interpretability of SA models and discovering new applications for this incredible tool.

To conclude, the history of SA demonstrates the relationship between technological innovations, methodological discoveries as well as the availability of new data sources and emphasizes the continuous expansion and growth of SA.

## *1.2 The four types of Sentiment Analysis*

Depending on the "type" of sentiment that need to be extracted, SA can be divided into four types: Fine-grained, Aspect-based, Intent-based, and Emotion detection (Nitor).

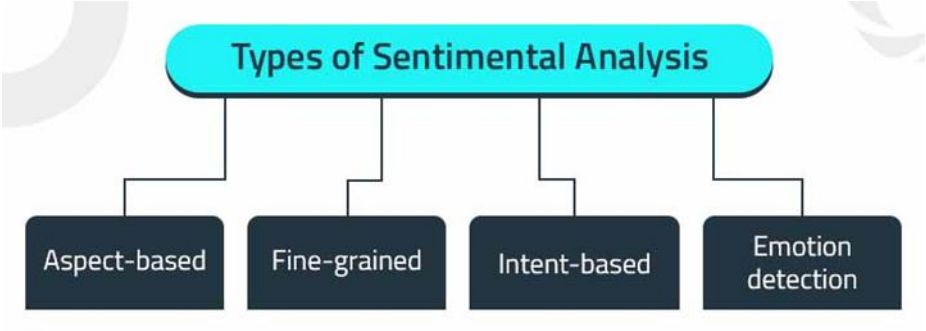


*Figure 2 Source: nitorinfotech.com, "Types of Sentiment Analysis", 04 May 2022*

Fine – grained SA normally includes two key tasks: element extraction and sentiment categorization. This model extracts the polarity from a piece of text classifying the polarity in the following categories: very negative, negative, neutral, positive and very positive. This model is widely used in the field of rating and reviews. For example, for a rating scale from 1 to 5, 1 should be considered as very negative and 5 as very positive. For example, let's take a look to the following sentence:

*"I love the food at this restaurant, but the service is slow"*

In the previous example, the sentiment expressed towards the "food" could be classified as 5 (very positive), while the sentiment expressed towards "service" could be classified as 1 (very negative).

The aspect – based (ABSA) type aims at recognizing the emotion expressed towards certain parts or characteristics of an entity, such as the battery life of a smartphone. This sort of SA is used to acquire a more fine-grained insight of client thoughts by examining the sentiment expressed towards various features of a product or service. The information obtained can be beneficial for products enhancement, as well as for brand reputation monitoring and customers' satisfaction. Considering the previous example, the sentiment expressed towards the "food" is positive, while the sentiment expressed towards "service" is negative. The goal of ABSA is to identify both aspect ("food" and "service") and the sentiment expressed towards each aspect.

Intent – based SA (IBSA) evaluates the overall sentiment indicated towards an activity or purpose, such as the feeling towards purchasing a goods. In this type of SA, the objective is to understand the sentiment behind an action or choice, rather than simply identifying the customers'

emotions. It may be valuable in industries such as marketing, customer service, and product development, since it gives insight into consumers' opinions and preferences. For example, an IBSA can help a company understand the intent of its customers, whether they intend to buy or not a product. With this information, the company can better assess the target for its advertisements, saving money and time. Using the previous example, the sentiment behind “food” and “service” is clear, but with IBSA we are also able to identify the intention behind that sentence, which, in this case, is to provide feedback to the restaurant or to make a recommendation to others.

Emotion detection SA (EDSA) aims at recognizing and classifying emotions of joy, anger, sadness, anxiety, and surprise conveyed in a text. This type of SA that goes beyond finding only positive or negative polarity to assess the sentiment of a text and can provide valuable insights about the emotional state of the writer. As a result, EDSA can be useful in applications such as customer SA, where understanding the emotions behind customer feedback can help businesses respond in more effective ways. Some common techniques for emotion detection include:

- Basic Emotion Detection: classifying a text into a set of basic emotions like happiness, sadness, anger, etc.
- Complex Emotion Detection: classifying a text into a set of complex emotions that are a mixture of fundamental emotions, such as love (a combination of both happiness and excitement).
- Dimensional Emotion Detection: classifying a text into continuous dimensions such as polarity (positive vs negative) and excitement (excitement vs tranquillity)

Using the same example, the sentiment expressed towards the “food” might be classified as happy or pleased, while the sentiment expressed towards “service” can be classified as frustrated or dissatisfied.

## *1.3 Sentiment Analysis approaches*

SA can be addressed using three approaches:

- Rule – based
- Machine Learning models
- Hybrid models

Rule – based SA uses a set of pre-defined rules to classify the sentiment of a text as positive, negative or neutral. These rules are based on lexicon features such as words or expressions found in a text. With this approach, the sentiment of the text can be deciphered as being positive or negative based on the presence of "positive" or "negative" phrases and words present in it . For example, two lists of polarized words are defined using positive terms such as "good" or "great" and negative words such as "poor" or "awful". After, the number of positive and negative words that appears in the text are counted and, if the positive count is greater than the negative one, the system will return a positive sentiment, and vice versa. This type of SA is well suited for applications that require a rapid and easy analysis, such as customer feedback analysis or social media monitoring. However, there are also some important limitations with this approach. Firstly, the creation and validation of lexicons is time-consuming. Another constraint is its lack of versatility because since the rules are predefined, they are not able to learn from the data and adapt to new circumstances. This may result in a substantial degree of inaccuracy, especially in situations where the text contains words or phrases that have different meanings or are used in multiple ways. Another weakness of rule - based SA is its limited capacity to handle sarcasm, and irony since the model only considers individual words and does not take into account the context in which they are used. Rule – based SA models are very simple but effective and they are more suitable for applications that do not require complicated analysis or a high degree of accuracy.

Machine learning approaches, differently from the rule – based ones, require training a model on a huge, labelled dataset of text to automatically learn how to recognize the sentiment represented in a text. To each piece of text is linked a sentiment label (e.g., positive, negative, or neutral). The model learns to identify the sentiment of new text by identifying patterns in the training data that are predictive of the sentiment labels. One of the best advantages of machine learning based methods is their ability to learn from the data and adapt to new situations. However, also machine learning approaches have limitations. The most important is the requirement for large amounts of labelled training data that can be time consuming and expensive to gather. Another drawback is that, unlike rule – based methods which provide predefined rules, machine learning models can be complex and

difficult to understand, making it complex to figure out how they arrived at their predictions. The most used machine learning algorithms for SA include Naive Bayes, Support Vector Machines (SVMs), and neural networks (CNN & RNN) (fig. 6). Depending on the dataset, the type of task and the requirements of the task, a different algorithm is used to perform SA.

Let's see the most used algorithms:

- Naïve Bayes: this classifier algorithm is very simple but surprisingly powerful for predictive models. It is based on Bayes' theorem with an assumption of independence between features. This algorithm calculates the probability of each class given an input and then classify the input into the class with the highest probability. There are three types of Naïve Bayes algorithms: Gaussian Naïve Bayes, Multinomial Naïve Bayes, Bernoulli Naïve Bayes.
- Support Vector Machine (SVM): it is a supervised learning algorithm that works by determining the hyperplane that best separates the data into distinct classes that, in the case of SA, the two classes are positive and negative sentiment (fig. 3). SVM is ideally suited for SA jobs because it is capable of processing high-dimensional data, such as text, and can deal complex non-linear relationships between the features and the target variable. Another benefit of SVM is that it is a computationally efficient algorithm. It uses a process called kernel trick, which allows it to map the input data into a higher dimensional space without calculating the mapping directly. This makes SVM significantly quicker and more efficient than other machine learning methods, such as neural networks, that would take a much higher number of computational resources to process the same amount of data. One of the main constraints is that SVM is a supervised learning algorithm, which means that it needs labelled training data to understand the connection between the features and the target variable and this might be time - consuming and costly to achieve.

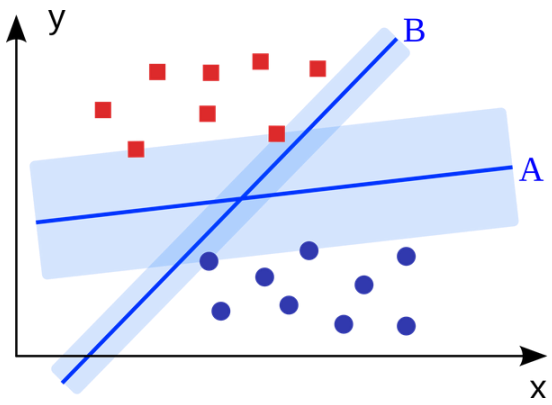


*Figure 3 Source: medium.com, Lucas Oliveira, "The SVM Classifier", 01 Aug 2017*

- Convolutional Neural Network (CNNs): is a type of deep learning algorithm that is frequently used for SA. CNNs are inspired by the structure and function of the human brain and are intended to process and analyse huge volumes of data, such as text, pictures, and audio (fig. 4). In SA, CNNs are used to analyse and classify text data, such as reviews or tweets, into various sentiment categories, such as positive, negative, or neutral. The algorithm works by first converting the text into a numerical representation, such as a one-hot encoding or word embedding, and then processing the data through several layers of neural networks to extract features and detect patterns in the data that are indicative of the sentiment. Two of the most important phases in CNN are convolution and pooling which aim at identifying the features in the input. One of the major benefits of CNNs for SA is their capacity to automatically learn and extract features from the data and they can manage huge amounts of information and can also process data in parallel, making them significantly faster and more effective than conventional machine learning methods. This makes CNNs suitable for SA jobs, where huge quantities of text data need to be processed quickly and efficiently. One of the biggest limitations is the need of large quantities of labelled training data to understand the relationship between the features and the target variable and the difficulty in understanding and interpreting the results. These limits could be minimized using proper data pre-processing.

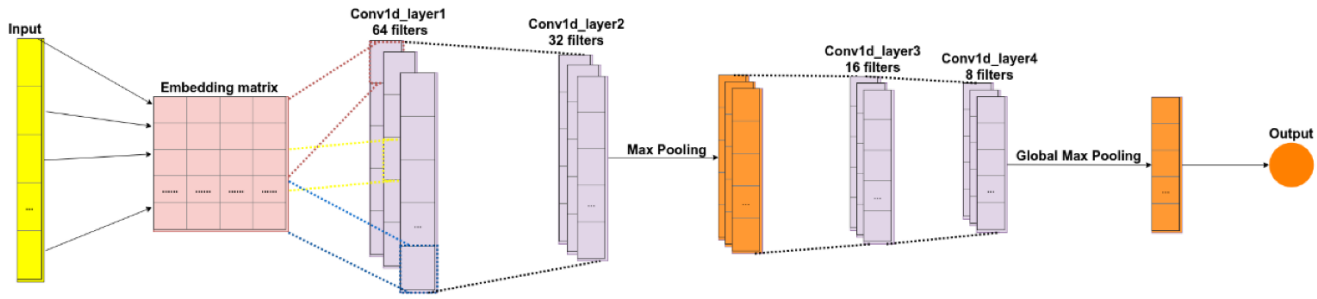


*Figure 4 A convolutional neural network. Source: Moreno-Garcia, M.N "Sentiment Analysis Based on Deep Learning: A Comparative Study" , Electronics 2020*

- Recurrent Neural Network (RNNs): RNNs are meant to process sequential data, such as time series data or text data, by processing the data one step at a time and using the output of one step as input to the next step. One of the main benefits of RNNs for SA is their ability to capture the context and relationships between words in text data. By analysing the data one word at a time, RNNs can capture the associations between words and phrases and utilize that knowledge to categorize the text data into multiple sentiment categories. Another strength of RNNs is that they can handle variable - length input sequences, which makes them well suited for SA jobs, where the length of the text data might vary substantially. One of the main drawbacks is the difficulty in training RNNs on long sequences of data, as the gradients might

become very small or disappear completely, which can make it difficult for the model to learn the relationships between the words and phrases in the text data. RNNs can have difficulty capturing long - term relationships in the data because the hidden state of the network is reset at the end of each sequence, which can cause the network to forget important information. Overall, RNNs are a strong and effective method for SA, especially for applications that necessitate the capacity to capture the context and relationships between words in the text data. These problems can be mitigated using appropriate data pre-processing and feature engineering techniques, such as using variants of RNNs, such as Long-Short Term Memory (LSTM) (fig. 5) or Gated Recurrent Units (GRU), that are designed to overcome these limitations (Best Machine Learning Algorithms for sentiment analysis? 2023).

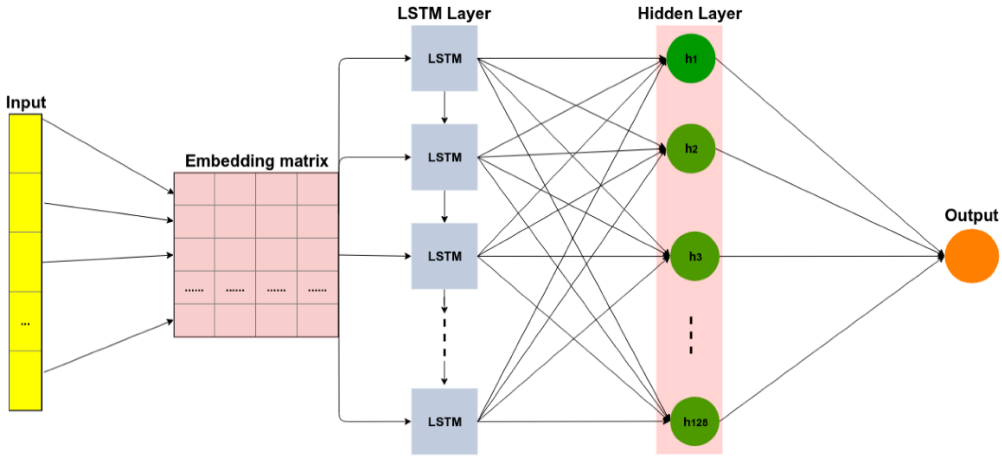


*Figure 5 A long short-term memory network. Source: "Sentiment Analysis Based on Deep Learning: A Comparative Study", Electronics 2020*

When it comes to hybrid SA models, the idea is to combine the strengths of rule-based and machine learning approaches to achieve better accuracy and overcome limitations of individual methods. Rule-based approaches can be very effective at detecting certain patterns and rules in language, while machine learning methods can detect more complex relationships and contextual nuances that can be challenging to capture with rule-based methods alone. In a hybrid model, the rule-based approach is often used as a pre-processing step to extract relevant features or to classify easy cases that can be accurately identified using simple rules. Then, machine learning methods are applied to analyse the more complex cases where rule-based methods are not sufficient, or to refine the results obtained from the rule-based analysis. This approach can yield more accurate results, especially for datasets with a wide range of language styles, tones, and contexts.

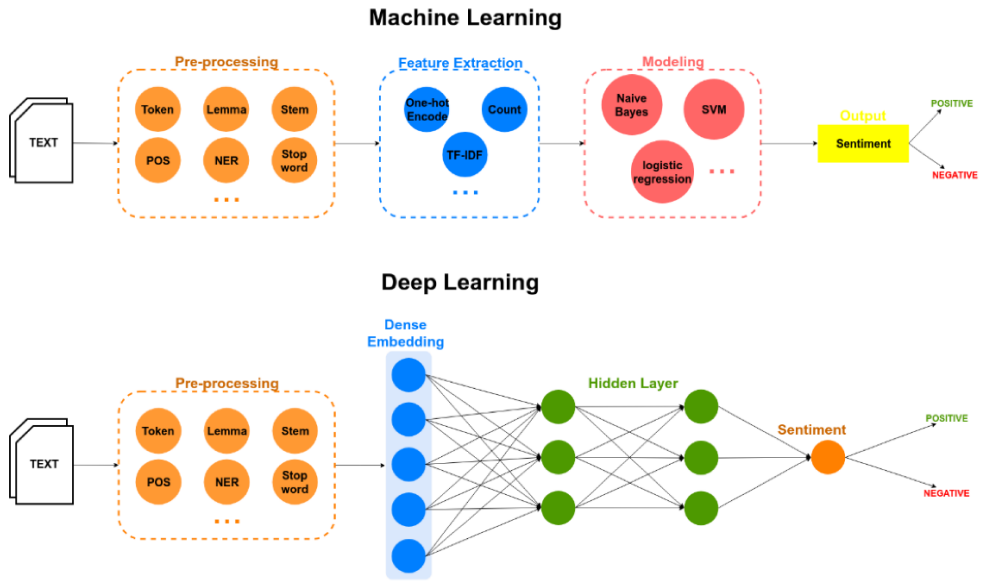


*Figure 6 . Differences between two classification approaches of sentiment polarity, machine learning (top), and deep learning (bottom). Source: "Sentiment Analysis Based on Deep Learning: A Comparative Study", Electronics 2020.*

## 1.4 How does Sentiment Analysis work?

### Rule-based SA

In rule-based SA, the process can be divided into five main phases. Everything starts with the creation of "lexicons" which are lists of both positive and negative words. After lexicons have been created, data needs to be prepared in order to be analysed and this is done in the text pre-processing phase. This step includes converting the text to lowercase, removing punctuation, special characters and removing stop words. This is a very important step because it helps to reduce the noise in the data and improve the accuracy of the SA model. After the text has been pre-processed, it is divided into individual sentences and, with the word tokenization phase, the sentences are broken into individual words enabling SA to be performed at the word level. Part-of-Speech tagging (POS) is the next step, and it aims at identifying the part of speech of each word in the sentence. This step is very important since it can provide important information as every word might have different implications for sentiment. For example, adjectives are typically more indicative of sentiment than nouns. The last step is Sentiment scoring where we use the lexicon list to assign a sentiment score to each word in the sentence. Sentiment scores can be binary (positive/negative), or they can be more nuanced, such as a 5-point scale from strongly negative to strongly positive. The sentiment scores of individual words can be combined to determine the sentiment of the entire sentence.

### Machine Learning-based SA

In ML-based SA, the process can be divided into seven key phases. Everything starts with data collection. This step aims at gathering text data in various forms (reviews, social media posts, or any piece of text that contains sentiments or opinions). The quality of the data gathered substantially influences the efficiency of the SA system, therefore it is essential to ensure that it is representative of the opinions and emotions that the SA system will encounter in real-world applications. The data gathered should be enough to train a SA model since using huge quantities of data will reflect in a better performing SA system. After obtaining the data, it should be cleaned, normalized and then prepared in order to be analysed and this is done in the pre-processing phase. This step is crucial for SA since better data means better performance and results of the SA system. We can divide the pre-processing phase into four steps:

- Text normalization: in this step we convert the text to lowercase, deleting special characters, numbers, and punctuation marks, and correcting syntax mistakes. The

purpose of this step is to normalize the text data and make it easier for the model to analyse it.

- Stop word removal: stop words are ordinary words, such as "and", "the", "of", etc., that do not have much meaning and can be omitted without compromising the sentiment analysis process. Stop word removal is performed to reduce the quantity of the text data and to improve the performance of the sentiment analysis algorithm.
- Stemming or lemmatization: these procedures are used to reduce words to their basic form, or stem, in order to group words with similar meanings together. This simplifies the process for the machine learning algorithm to discover patterns in the text data and improve the efficiency of the sentiment analysis system.
- Tokenization: this includes dividing the text into individual words, or tokens, for further analysis. The aim of tokenization is to simplify the process of determining the sentiment of individual words and to convert the text data into numerical representations.

After the data has been pre-processed, it is time for the feature extraction phase. This step aims at identifying and extracting important features from the pre-processed data that can be used to determine the sentiment. The features gathered may include the presence of some words or phrases (positive or negative), the overall sentiment of the text or other characteristics that may be significant to the SA task. This can be done using techniques such as TF-IDF, word embedding, bag of words, etc. The next step is the model selection. In this step, a machine learning algorithm is chosen in order to do the SA classification. There are several algorithms that can be used to address SA such as Naive Bayes, Support Vector Machines (SVMs), and Recurrent Neural Networks (RNNs). Depending on the difficulty of the SA task, on the purpose of the SA and on the type and quality of the data available for the analysis the most appropriate algorithm is used.

After choosing the model we can move on to the next step: model training. This step consists in training the machine learning algorithm on a labelled dataset to understand the relationship between the words and the sentiment of the text. The purpose of training is to find the model parameters that minimize the prediction error between the model predictions and the true sentiment labels. The quality of the training process, including the choice of features, the choice of algorithm, the quality and relevance of the labelled data, strongly influences the precision of the SA system. Once the model has been trained, it is ready to be tested.

In the model testing phase, the trained model is tested on a different dataset to measure its performance and accuracy. This requires comparing the model predictions with the true sentiment labels and determining performance metrics such as accuracy, precision, recall, and F1-score. The outputs of the model testing phase are used to analyse the performance of the SA system and to find areas for improvement and optimization. Finally, the trained model can be applied in a real-world application with the Deployment phase, where it may be used to predict the sentiment of new text data. The deployment step may include integrating the SA system into a larger system, such as a customer service or marketing application, or making the system available as a web service or API.

## *1.5 Sentiment Analysis limitations*

In recent years, SA has become a very powerful tool for businesses aiming to get insight of consumer opinions, preferences and sentiments. However, despite its increasing popularity, SA has its limitations. In this chapter, I will go over the limitations of SA and their effects.

One of the most crucial limitations of SA is its dependence on labelled training data. SA models are trained on huge datasets of labelled data, in which each piece of text is assigned a sentiment label. However, the quality and accuracy of SA models depend on the quality of the training data, as already said. If the training data is biased or does not represent the target population, the model's efficiency may be significantly affected. Another disadvantage of SA is its difficulty in understanding sarcasm and irony, and this may significantly change the sentiment of a piece of text.

*For example, a sentence like "What a great day it is today" said in a sarcastic tone can be misinterpreted as positive by a SA model.*

Complex syntax is another limitation for SA. SA models might have problems in processing complex syntactical structures, such as negation or multiple negations.

*For example, a sentence like "I don't dislike it, but I don't like it either" might be classified as neutral by a SA model, but the sentiment expressed is negative.*

Ambiguity is also another significant challenge for SA since the system can have problems with words that have multiple meanings or connotations (word sense disambiguation). This can bring to a significant change of the sentiment of piece of text. In order to mitigate this problem, it is very important to use additional techniques and methods to provide better accuracy to the SA model. These techniques include word sense disambiguation algorithms, topic modelling or human-in-the-loop method, where human annotators are involved in the process to ensure the accuracy of the SA model.

*For example, a word like "hot" could mean "spicy" or "attractive", depending on the context in which it is used.*

Culture also plays a key role in SA as cultural differences may significantly change the sentiment of a piece of text. This happens because some emotions and expressions may be more or less popular in different cultures. In some cultures, it may be considered impolite to express negative sentiments directly. For example, in Japanese culture people tend to communicate indirectly and may avoid giving a direct refusal or negative answer. Rather than saying "no," they may respond with phrases such as "I will consider it," even if they have no intention of considering the proposal (Cultural Atlas

2021). This can cause mistakes in SA models trained on data from one culture, when applied to data from a different culture.

Subjectivity is also a limitation since SA models are trained to identify subjective statements, but they might not work properly with texts that contain highly subjective or personal opinions.

*For example, a sentence like "I think this is the best movie ever" might be classified as positive, but it is a subjective statement that may not be true for everyone.*

SA remains a valuable tool for companies and organizations looking to gather and analyse public opinions. However, SA has many limitations that affect its accuracy, including dependence on labelled training data, difficulty in understanding sarcasm and irony, challenges with complex syntax, ambiguity in words and expressions, cultural differences, and subjectivity in personal opinions. To take the full advantage of SA, we must be aware of these limitations and learn to use SA in conjunction with other methods. In this way, SA can be of huge help to gain a more comprehensive understanding of the sentiments expressed in text data.

# *2 Sentiment Analysis:a very useful tool in many sectors*

The first chapter covered SA as a whole, and it is precisely the objective of this second chapter to describe and explain how the SA is used in its main fields of application, analyse its functions, its potential and illustrate the advantages that this tool can bring to its main field of applications.

For a marketer, for example, it could be very useful to find out what a customer really thinks about one of his products, while a politician would have a significant advantage if he knew in real time which topics are important to its voters. In the financial field SA can provide useful indicators for forecasting the future trend of a stock of interest on the stock market, which will be discussed in Chapter 3. In addition to the fields mentioned above I will also talk about the advantages SA has on tourism and other areas in which it can have an important role such as in the healthcare sector or in sports.

It must be mentioned that this thesis studies the functions and opportunities of the SA mainly from an economic - financial point of view; in each paragraph I will then evaluate the functions that SA makes available in a certain sector, and then explain what is the potential that it offers. Although it is intriguing, a technical - engineering study of this tool is not offered in this thesis. Instead, I will focus on the advantages that the usage of SA might deliver to its users and at what expense.

## *2.1 Sentiment Analysis in Marketing*

The connection between SA and Marketing appears almost immediately. Let's ask ourselves this question: which company is not interested in knowing the opinion of its customers in real time, even just for being able to know their preferences and beliefs or even for being able to evaluate the real effectiveness of an advertising campaign (Rambocas, Meena e Joao Gama 2013). Thanks to Web 2.0, the increasing number of users sharing reviews online, in addition to the opportunity provided by new tools to analyse Big Data, Marketing managers have unprecedented opportunities to be able to conduct market research or monitor the web reputation of their brand, in an easy way and without significant costs (Erevelles, Sunil e Fukawa 2015).

Although the topic and the studies are still at an early stage, additionally to the researchers, also the companies are increasingly interested in the use of the SA for their own marketing. Big corporations, having their own dedicated marketing department, undoubtedly manage to use the tool, but it is important to note that also many little organizations have decided to invest in this tool too, even if frequently in outsourcing, through specialized marketing agencies. The number of this kind of outsourcing organizations is expanding fast and they can provide many services and at cheap cost. Among them we can find for example:

- Track opinions and evaluations of services and goods
- Monitor that there are no online "issues" related to the business they could generate negative viral effects
- Assess market trends, competitor activities and current trends
- Measure public response to corporate or stakeholder actions (Rambocas, Meena e Joao Gama 2013)

To better describe how the SA may be useful in fulfilling the goals indicated earlier, it is appropriate to break these objectives into 3 categories:

- Market Research - In this case the goals might be diverse such as, understanding the opinion of the selected target on a certain product or subject of interest or wanting to segment the market and get to know better your consumers or the major stakeholders.
- Brand Reputation - The purpose here is to understand and control with social networks the public opinion of the stakeholders and of the consumers about their own brand, in order to avoid negative viral posts. There are two approaches used: Top Down and Bottom Up.

- Top-Down Approach: This is the most traditional and common approach. Here, businesses simply insert themself into social networks by building corporate profiles and managing their accounts with the help of a social media manager to be able to interact with individuals and enhance visibility and improve the reputation of the business.
- Bottom-Up Approach: This is the most innovative and probably the more successful strategy. Now organizations consider Social Networks as a huge square where everyone writes their own opinion about anything at all. It is consequently feasible to intercept in real time any issue that carries potentially negative viral news that could bring a loss in the market share.

- Marketing campaigns and Buzz Marketing - Being able to analyse the performance of a certain marketing activity.

The most intriguing aspect of all this, however, is not just the fact that you can understand the opinions of clients, but that it is possible to do this continuously throughout time and therefore before and after any event and, most importantly, in real time. In this way it is possible to make predictions about the future (forecasting) and projections about the present (nowcasting). Another important aspect of SA in marketing is that it can assist in improving customer experience by doing SWOT analysis, which stands for strengths, weaknesses, opportunities and threats. It can be used in product design and marketing to measure customer satisfaction rates and to see which aspects of a product are really liked by the customers and the ones that really need to be improved.

In confirmation of what a huge opportunity this is, even the CIA has created a project in collaboration with Google, called "Recorded Future", which aims to analyse the web and social networks in order to search for relationships between people, organizations, actions and events for creating tools capable of formulating reliable forecasts. (Condè 2010)

## 2.2 Sentiment Analysis in Politics

Since some years ago SA is a widely used method of capturing consensus from the politicians, suggesting them which are the hot topics on which to focus the electoral campaign. This is the most common application of SA in politics and it is done by analysing text data related to electoral campaigns so that SA can identify trends in public sentiment, such as changes in support for a particular candidate or party, or shifting public opinion on a specific issue. Both politicians and political parties uses this information to change their strategies to better align with public sentiment. SA in politics can also be used to track public opinion about specific political events, such as elections, debates, and major policy announcements or to track the reputation of political candidates, parties and organizations. Let's see two examples from the New York Times of 2010:

"...Crimson Hexagon, a technology company in Massachusetts, analysed expressions of public sentiment across the country about the oil spill in the Gulf of Mexico. Its analysis showed that people who lived near the gulf had a lower tendency to assign blame, focusing instead on the logistics of the relief efforts..." (Rambocas, Meena e Joao Gama 2013)

"Linguamatics, a British company, analysed posts from more than 130,000 Twitter accounts to gauge public opinion during the British elections this spring. The company's analysis yielded similar results to traditional political polling, and predicted within one point the percentage of votes the Conservative Party would win." (Rambocas, Meena e Joao Gama 2013)

It should be mentioned that an information asymmetry is frequently created between the politicians who use these means for their campaigns and those who don't and this difference in the long run can be significant for the election results. Also government agencies are exploiting SA by monitoring peaks of negative sentiment relating to a particular person, event or entity, in order to gather useful information that might be helpful to foil emerging threats. According to the New York Times, back in 2006, the US government would have spent approximately 2.5 million of dollars in research for a software to track internet activity;

"...Sentiment Analysis is meant to detect potential threats to the country. We want to comprehend the rhetoric that is being published and how intense it is." (Rambocas, Meena e Joao Gama 2013)

I would like to discuss quickly a problem that the usage of SA in politics might generate. The dilemma is this: how trustworthy and "genuine" is a politician who relies his electoral campaign on the trends of the moment and not on his actual and fundamental ideological convictions? I think that an excessive and careless use of the SA can lead to a class of politicians driven by social network trends, flattening any political discussion on issues and giving control to those who have the power to influence the opinion of the masses on the political class or even on the government of a nation.

## *2.3 Sentiment Analysis in Tourism*

It is not just the economic - political fields that are interested in using SA. The usage of SA in the tourism sector is becoming more and more popular. Indeed, tourists have now access to much more information about any destination thanks to the web, low-cost airlines and room booking companies such as Booking or Trivago; therefore, the number of tourists who go on holiday without buying bundles from tour operators is always growing.

As a further consequence, there are hundreds of reviews written by the users themselves every day and related forums/blogs for holiday suggestions. We can say that anybody who decides to travel to any location independently begins studying information about the holiday a long time before they leave and in the same way post images, videos and reviews, sharing their experience with other people. This dramatically affects the perspective of the tourist offer. So, it becomes essential for a hotel, as for a tourist attraction or a city itself to be able to attract as many tourists as possible by presenting themselves optimally on the web before the customer book. Another important aspect is to ensure that the customer is truly satisfied by posting positive or at least non-negative reviews, which might really affect the reputation of the attraction/city.

Which tool better than SA can perform all this? In fact, through SA, using the infinite amount of data offered mainly by platforms such as TripAdvisor, Expedia, VirtualTourist and Lonely Planet, it is possible to segment the tourist market by trying to define its own optimal tourist target following its own tastes and needs, using filters and keywords and analysing past reviews. SA enables you to monitor, using reviews and posts online, how much the client was satisfied with their experience and for which reason. An extra degree of analysis can be performed by focusing more on the destination rather than on the consumer, obtaining other useful insights such as:

- Understand the polarity of opinion regarding your offer.
- Understand which topics or attractions are the most mentioned by travellers. A great way to do so it by using word cloud. (fig. 8)
- Understand what they link the name of their facility with and why they book (Example: The hotel near the square...)
- Being able to compare your offer with that of your competitors
- Understand, utilizing historical data, one's trend through time

Going beyond the micro - economic view of the tourist industry and examining tourism in a wider way, we discover that the SA may also be highly beneficial for sociologists, public organizations or

other organizations engaged in the tourism sector. With geolocated postings and the SA it is possible to monitor in real time and discover the dynamics of the world of tourism, with specific reference to the quantity of visitors, from which region of the world they arrive and at which time of the year.

I cite as an example a study from 2017 (Neidhart, et al. 2017), which has highlighted various trends, including that the macro-regions more visited are Asia (32.6%) and then Europe (25.4%) and that most travellers come from Europe (60.5%) and then from Asia (24.0%). It also found that travellers from the UK mostly travel in Asia (31.7%), Australia and Oceania (26.3%) and Europe (24.2%), that travellers from Canada clearly favour tours to Asia (57.3%) followed by Europe (21.2%), while travellers from the USA prefer tours to the same two continents but with a fairer percentage (32.1% and 30.6%). Also, the majority of New Zealanders instead book tours in Europe (63.4%), followed by Asia (22.2%) and more than 80% of South Africans travel to Europe (82.2%) and that none of them participate in tour in their area of residence. Finally, it showed that German tourists prefer excursions to Asia (41.4%), followed by North America (18.8%) and Africa (18%). It was also determined that the scores of higher average ratings originate from Australian travellers; the same applies to travellers of the United Kingdom and Germany. On the other side, Canadians' feedback scores are significantly lower than the average, as well as those of Americans.

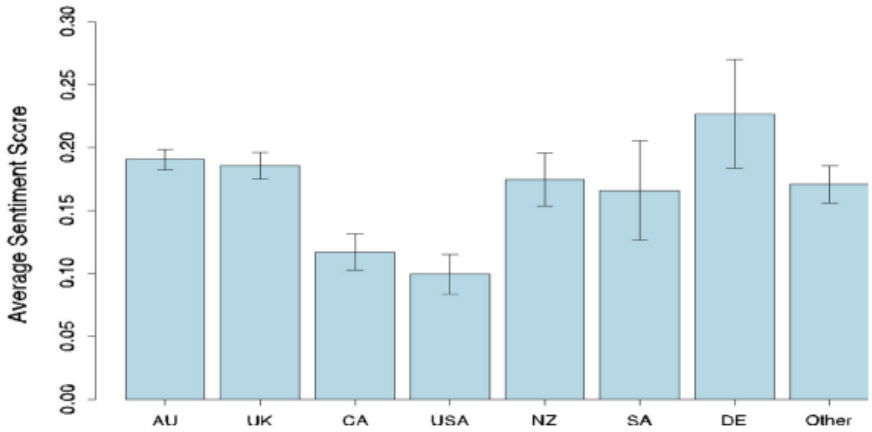


*Figure 7 Average sentiment scores and standard error per country. Source: "Predicting happiness: user interactions and sentiment analysis in an online travel forum", 2017*

Studies like this might be extremely valuable. In fact, in order to improve its tourist offer, info about how many visitors arrive, from where, in which period and for which reason, can be extremely beneficial to a specific region/city.


*Figure 8 Word Cloud of Rome's most mentioned attractions by tourists.*

## *2.4 Sentiment Analysis and other fields of application*

In the previous sections, I talked about how SA is widely used in marketing, tourism or politics, but it can be used in a lot of other fields of applications very different from each other. SA, for example, can also be applied in the Social Psychology field in order to analyse the sentiment of a population over time using social network posts or it is possible to understand the reaction of the population to specific episodes. Let's talk about a study from 2012 (A. Ceron, L. Curini e S.M. Iacus 2014) whose aim was to analyse the happiness of Italians using SA. It has been found that the happiest moments for Italians were Obama's victory in the American elections and Mario Balotelli's goal at the European Football Championship.

SA can also be used in the healthcare sector to monitor patient feedback, evaluate the effectiveness of treatments and even predict epidemics with the use of geo localized posts (fig. 9). By analysing large amounts of patient feedback from various sources, healthcare providers can gain a better understanding of what patients think about their services, staff and facilities. For example, healthcare providers use SA to find common issues and complaints raised by patients such as long wait times, poor communication or inadequate care, and address these issues proactively.

SA can be applied to the sports betting sector too since it can be used to analyse a player's sentiment in order to predict their performance based on their current mood, form and psychological state. It is also possible for sports organizations to gain a better understanding of fan opinions and sentiments regarding their team and players. This kind of information can help the sport organization to build a stronger connection with their fans and to give them a better fan experience. For example, it can identify common issues such as high-ticket prices, poor team performance or poor fan engagement.

SA can also be used in the education sector as well to analyse student feedback and to evaluate the effectiveness of teaching methods. SA enables educators to understand student opinions and sentiments towards their classes, professors and institutions. For example, SA can be used to identify issues raised by students such as poor course content or poor teaching method and to analyse the student feedback on a specific course, allowing teachers to identify problems in order to make improvements and improve the student experience.

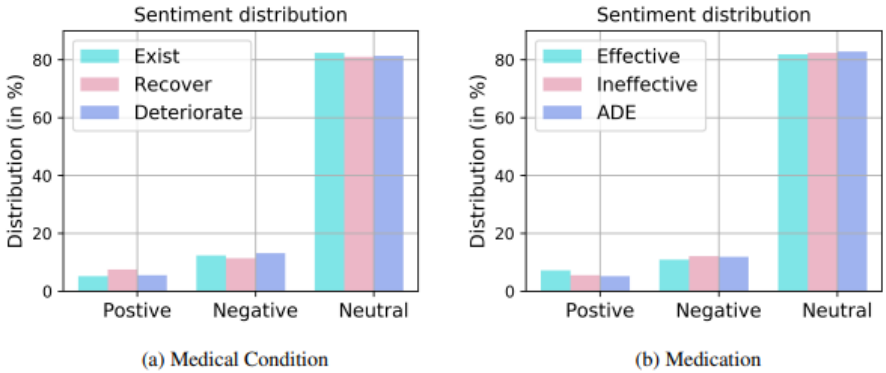


*Figure 9 Sentiment Analysis in Healtcare regarding the existance of medical condition (a) and the effectiveness of a medication (b)*

# 3 Sentiment Analysis in Finance

This chapter will talk about how SA is being used to predict the trend of stocks using financial news. Stock markets create large amounts of data every day and it is very hard to consider all the information for predicting future trends of a stock. We can predict market trends using two methods: technical analysis and fundamental analysis. The difference between these two methods is that the first uses past prices and volumes to forecast the future trend while the second one analyses its financial data to get some insights (investopedia 2022). What I want to answer in this chapter is: can financial news have an impact on stock trends?

The goal of SA in finance is to comprehend the feelings, opinions, and attitudes of people towards a certain financial asset or market and to use this information to generate predictions about its future performance and in order to make smart investment choices. There are different ways in which sentiment analysis can be used to predict stock prices. One strategy is to analyse news articles, press releases, and other types of publicly accessible text to detect the sentiment expressed about a certain stock as well as the overall sentiment of the market. For example, if a big number of tweets about a certain stock are negative, it may signal that investors are pessimistic about the stock, which might negatively influence its future performance. For example, in February 2018 social media influencer Kylie Jenner tweeted the following:

*"sooo does anyone else not open Snapchat anymore? Or is it just me… ugh this is so sad."*

With more than 39 million followers, Kylie Jenner's tweet had a large impact in the share price of SNAP, the parent company of Snapchat. In just one day, the share price decrease by 7% and SNAP lost about $1.3 billion in market value (Costa Tammy 2020).

The idea behind using SA in finance is that public opinion and news can have a significant impact on the stock market, and by analysing large volumes of financial news, it is possible to predict the future trend of a stock (fig. 10). There are several factors that might influence stock sentiment such as financial or industry related news, social media posts and financial sentiment indicators. The most relevant financial sentiment indicators of a stock are:

- Volatility: indicates the range of a price movement of a stock over time. Of course, trading on a market that is more volatile than others have the potential to make big profits, but also substantial losses.
- Put/Call Ratio: it is very important in order to understand the opinion of traders on a specific stock. It refers to the volume of put options to call options on a given stock.

For example, if 20000 investors have bought call options for a stock and 2000 traders have bought put options in the same trading session, then the sentiment would be bullish (price goes up) since most investors expect the price of that stock to go up and they use the call option in order to take advantage of the higher prices they are expecting.

- Client Sentiment: it is the percentage of investors who are long or short on an asset at a given time. A long position means that the investor owns the stock and will make a profit if the price of the stock rises. A short position is when an investor doesn't own the stock and will generate profit if the price goes down. This data can show when positioning is reaching extreme ends compared to the price of the stock in consideration. (Costa Tammy 2020)

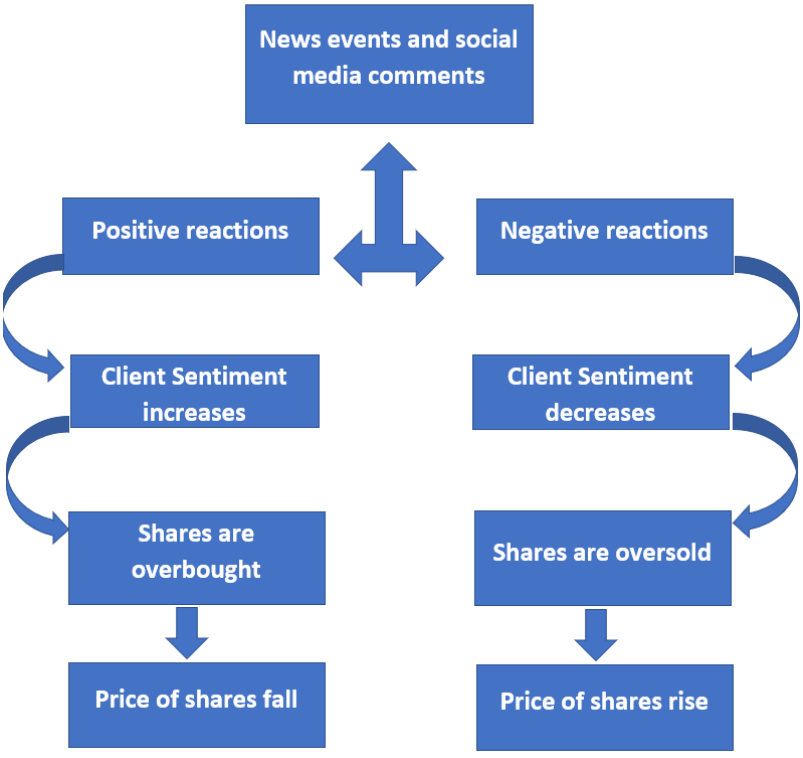


*Figure 10 The idea of SA in Finance. Source: dailifx.com*

## *3.1 Can SA help predict Stock Trends using financial news?*

To predict changes in Stocks using SA, we need to combine it with tools such as technical analysis and financial sentiment indicators. News and emotional events (negative comments on social media or bad news) can scare traders pushing them to massively sell a specific stock. The opposite happens when very positive news is released, letting traders be optimistic and boosting the price of the stock. A very recent example can be seen with Alphabet, Google's parent company, which is working on its AI chatbot called Bard, a competitor of ChatGPT from OpenAI. Alphabet AI chatbot produced a factual error during its first demo, and led Alphabet shares to drop by 9%, leading to a loss of $100 million in market value. Meanwhile, Google's rival Microsoft shares, rose by 3% after they announced that they would have incorporated ChatGPT into Bing search engine (Google Shares Drop $100 Billion after its new AI Chatbot makes a mistake 2023).

When technical indicators hit extreme values, traders may begin to consider an inverse effect as more likely. The same might be perceived when emotion is at extreme levels. For example, if 90% of retail customers are long a specific market or stock, this might possibly be interpreted as a bearish signal (downward trajectory).

Some researchers showed that there is a strong relationship between financial news about a company and its stock prices fluctuations. A study from 2016 (Kalayani, Joshi, H.N. Bherathi e Rao Jyothi 2016) shows how it is possible to predict stock trends by analysing financial news. They collected Apple Inc. data from 2013 to 2016 (from Reuters and Yahoo Finance): both financial historical data and financial news for the same period. Since the data is unstructured, it must be pre-processed in order to be classified by a classifier. In this phase, all the pre - processing tasks have been made (tokenization, stemming, text normalization and stop word removal) (Chapter 1.4). To automatically detect the sentiment of the financial news, they followed Dictionary based approach which uses a dictionary for text mining. They created two lists of positive and negative words with both general and finance specific sentiment words. In order to determine the polarity of financial news they considered both the title and the text of the news. An overview of their algorithms is given below:

1. Tokenize the document
2. Prepare the dictionary which contains words and financial terms with its polarity (positive or negative)
3. Check if each word of the news matches with one word from the dictionary (positive or negative)
4. Count the number of words that belongs to positive and negative polarity

5. Calculate score of document. They did this by subtracting the count of negative matches to the count of positive matches.
6. If score is 0 or more, they considered the news as positive, negative otherwise

As a result, they got a news dataset with sentiment score and polarity (positive or negative). In order to do the text classification, they used three algorithms: SVM, Random Forest and Naïve Bayes. Then, they split the dataset into train and test and created a new dataset in order to check the accuracy of the classifier using new data. They evaluated the performance of all three classifiers and saw that SVM classifier performs well for unknown data (90% accuracy), followed by Random Forest (80%) and Naïve Bayes (75%). After the classification of unknown data, they plotted the news score chart compared with the historical price chart (fig. 11).

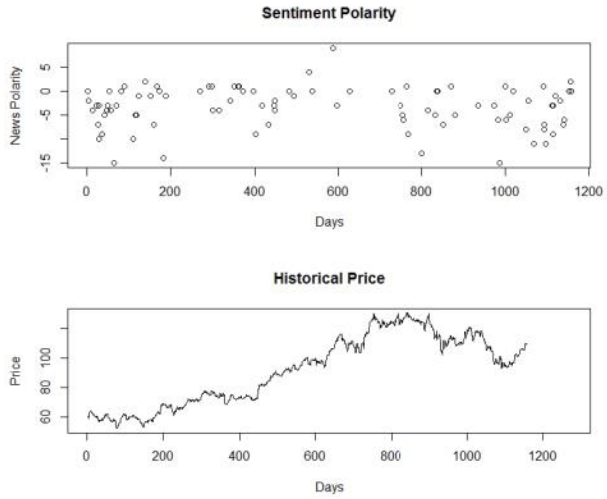


*Figure 11 The plot above is about the sentiment polarity of financial news compared with the chart of historical price. Source: "Stock trend prediction using news" by Kalyani Joshi, Prof Bharathi and Prof Jyothi, 2016*

After testing the models using different testing options such as 5/10/15 fold cross validation, changing the percentage of the data split (70% and 80%) and using new data, they compared their results. Random Forest has revealed to be the best one and its accuracy goes from 88% to 92%, followed by SVM with 86% and Naive Bayes algorithm with 83%. Once the model was ready, it could be used to predict the polarity of any news article and so, to predict the stock trend.

# *Conclusion*

SA has revealed to be a powerful tool in the field of data analysis and has found applications in various sectors such as marketing, social media monitoring, politics and finance. SA can provide valuable insights about customer's opinion, public sentiment and market trends. It is an effective tool that organizations use for identifying emerging trends and respond proactively to customer needs and preferences in order to improve operations and increase customer satisfaction. In this thesis I highlighted how SA can be used in the finance sector to predict stock trends by analysing financial news or social media posts related to a particular stock. For example, if many financial news or posts express positive sentiment about a stock, it is more likely that the stock's price will rise in the future and vice versa. However, it is important to say that SA should not be used as a basis for investment decision, but it should be seen as a tool which, integrated with technical analysis and financial sentiment indicators, can provide valuable information for a trader and provide a comprehensive investment strategy.

In conclusion, SA is a rapidly evolving tool used in several sectors and it has the potential to revolutionize the way organizations interact with their customers and make decisions. With continuous advancements in NLP and ML technologies, and the amount of data generated that is growing exponentially, I think that SA will become a widespread tool for both organizations and individuals helping to make the best decisions.

# *Bibliografia*


A. Ceron, L. Curini, e S.M. Iacus. «social media e sentiment analysis: l'evoluzione dei fenomeni sociali attraverso la rete.» *Springer*, 2014.

Ajayi, Demi. *How BERT and GPT Models Change the game for NLP.* 2020. https://www.ibm.com/blogs/watson/2020/12/how-bert-and-gpt-models-change-the-game-for-nlp/.

*Best Machine Learning Algorithms for sentiment analysis?* 2023. https://aiblog.co.za/ai-faq/best-machine-learning-algorithms-for-sentiment-analysis.

Condè, Nast. *Google, CIA invest in "future" of web monitoring.* 2010. https://www.wired.com/2010/07/exclusive-google-cia/.

Costa Tammy. *Using Sentiment Analysis to examine stocks.* 2020. https://www.dailyfx.com/education/understanding-the-stock-market/stock-market-sentiment-analysis.html.

Cultural Atlas. *Japanese culture - communication.* 2021 January 2021. http://culturalatlas.sbs.com.au/japanese-culture/japanese-culture-communication.

Erevelles, Sunil, e Fukawa. *Big Data Consumer analytics and the transformation marketing*, 2015.

H2OAI. *Natural Language Prcocessing - Sentiment Analysis.* s.d. https://h2oai.github.io/tutorials/natural-language-processing-sentiment-analysis/#0.

Kalayani, Joshi, H.N. Bherathi, e Rao Jyothi. «Stock trend prediction using news sentiment analysis.» 2016.

Neidhart, Julia, Daniel R. Fesenmaier, Tsvi Kuflik, e wolfgang Worndl. «Worskshop on recommenders in tourism. In proceedings of the eleventh ACM conference on Recommender Systems.» 2017.

Nitor, User. *A guide to sentiment analysis - part 1.* s.d. https://www.nitorinfotech.com/blog/a-guide-to-sentiment-analysis-part-1/.

*oogle Shares Drop $100 Billion after its new AI Chatbot makes a mistake.* 2023. https://www.tpr.org/2023-02-09/google-shares-drop-100-billion-after-its-new-ai-chatbot-makes-a-mistake.

Rambocas, Meena, e Joao Gama. In *Marketing research: the role of sentiment analysis*. 2013.

Stone, P.J., Bales, R.F., Namenwirth, J.Z and Ogilvie, D.M. In *The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information*, 484-498. 1962.

*The difference between fundamental vs technical analysis?* 2022. https://www.investopedia.com/ask/answers/difference-between-fundamental-and-technical-analysis/.

Zippia. *How many people use the internet? [2023]: 35 facts about internet usage in america and the world.* 12 January 2023. https://www.zippia.com/advice/how-many-people-use-the-intrnet/.