2022-01

Sentiment Analysis for Marketing: A Case Study on Twitter

Tsekoura, Maria

Digital Marketing Program, School of Economic Sciences and Business, Neapolis University Pafos

JANUARY 2022

Digital Marketing-Distance Learning

Sentiment Analysis for Marketing: A Case Study on Twitter

Maria Tsekoura

JANUARY 2022

Digital Marketing-Distance Learning

Sentiment Analysis for Marketing: A Case Study on Twitter

Dissertation submitted for distance learning Master of Digital Marketing at the University of Neapolis

Maria Tsekoura

JANUARY 2022

Copyright

Student's name: Maria Tsekoura

Postgraduate Thesis Title: Sentiment Analysis for Marketing: A Case Study on Twitter

This Master Thesis was prepared in the context of the studies for the distance master's degree at the University of Neapolis and was approved on…… [date of approval] by the members of the Examination Committee.

Examination Committee:

Research Supervisor: (University of Neapolis Paphos) …… [name, rank, signature]

Member of the Examination Committee: …… [name, rank, signature]

Member of the Examination Committee: …… [name, rank, signature]

Ή ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

Η Μαρία Τσεκούρα, γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Sentiment Analysis for Marketing:A Case Study on Twitter.», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Η Δηλών /σα

Μαρία Τσεκούρα

List of Content

List of Tables

List of Figures

Acknowledgments

I would like to convey my heartfelt appreciation to my research supervisor, Assistant Professor Stavros Souravlas, for allowing me to do research and providing me with guidance during this process. Working under his leadership was a tremendous privilege and honor.

I would like to express my heartfelt gratitude to Sotiris Tsatsos for his acceptance and patience during the dissertation's preparation. Additionally, I am indebted to my parents for their love, care, and sacrifices for my education and for my future preparation, as well as their understanding and continuous support in completing my research endeavor.

Finally, I'd like to express my gratitude to the members of my dissertation committee and the University of Pafos.

I would like to dedicate my dissertation to my family

Summary

Understanding the consumer public's perspective is a well-known issue that affects all businesses. Nowadays, social media precisely reflects the public's sentiments and thoughts on current events. Twitter, in particular, has garnered considerable attention from experts conducting research on the public's emotions. The utilization of social media data for marketing purposes is rising daily.

Given the foregoing, in this dissertation we examined the sentiment analysis of tweets according to their polarity. We chose a highly popular product, the iPhone13, which is manufactured by Apple, the world's most successful technology firm. More precisely, we gathered 1303 tweets about iPhone13, classified them as positive, neutral, or negative, and after processing the data, we applied the Naive Bayes classifier as well as 10 - fold cross-validation for better accuracy in the results. The experimental results have shown that the preferred classification model received relatively high evaluation results, with an average accuracy of 82.7%.

Keywords: Sentiment analysis, Naïve Bayes, Twitter API, iPhone

Περίληψη

Η κατανόηση της οπτικής του καταναλωτικού κοινού είναι ένα πολύ γνωστό ζήτημα που επηρεάζει όλες τις επιχειρήσεις. Σήμερα, τα μέσα κοινωνικής δικτύωσης αντικατοπτρίζουν με ακρίβεια τα συναισθήματα και τις σκέψεις του κοινού για τα τρέχοντα γεγονότα. Ειδικότερα το Twitter, έχει συγκεντρώσει μεγάλη προσοχή από ειδικούς που διεξάγουν έρευνα για τα συναισθήματα του κοινού. Η χρήση των δεδομένων των μέσων κοινωνικής δικτύωσης για σκοπούς μάρκετινγκ αυξάνεται καθημερινά.

Δεδομένων των προαναφερθέντων, στην παρούσα διατριβή εξετάσαμε την ανάλυση συναισθημάτων των tweets σύμφωνα με την πολικότητα τους. Επιλέξαμε ένα εξαιρετικά δημοφιλές προϊόν, το iPhone13, το οποίο κατασκευάζεται από την Apple, την πιο επιτυχημένη εταιρεία τεχνολογίας στον κόσμο. Πιο συγκεκριμένα, συγκεντρώσαμε 1303 tweets για το iPhone13, τα ταξινομήσαμε ως θετικά, ουδέτερα ή αρνητικά και μετά την επεξεργασία των δεδομένων, εφαρμόσαμε τον ταξινομητή Naïve Bayes καθώς και 10 - fold cross-validation για καλύτερη ακρίβεια στο αποτέλεσμα. Τα πειραματικά αποτελέσματα δείχνουν ότι το προτιμώμενο μοντέλο ταξινόμησης έλαβε σχετικά υψηλά αποτελέσματα αξιολόγησης, με μέση ακρίβεια 82,7%.

Λέξεις Κλειδιά: Sentiment analysis, Naïve Bayes, Twitter API, iPhone

Chapter 1

1. Introduction

Because of the widespread availability of the Internet in the twenty-first century, the globe converted into a global neighborhood. According to the most recent data, the total number of internet users worldwide was 4.66 billion in January 2021, a rise of 7.3 percent (+316 million) over January 2020. Indeed, 59.5 percent of the world's population now uses the Internet. Thus, social media users now account for 53% of the global population("Internet Users in the World 2021 _ Statista" 2020). As a result of the above, more than half of the world's population has a voice in social media.

Over the last several years, the function of social media has grown significantly beyond just facilitating our social lives. Social media sites are now essential parts of how we engage with politicians and the rest of the world, and we can't live without them. Social media also plays a crucial economic role because of the direct interaction with consumers enabled by social media platforms. Many businesses incorporate social media into their marketing strategies and take advantage of the direct interaction with customers those social media platforms enable. According to a study published by the Content Marketing Institute in North America in 2010 (Helal, Ozuem and Lancaster, 2018), 96 percent of business-to-consumer content marketers use social media for marketing purposes. Some businesses, such as Apple, even use social media as a customer service component.

The success of marketing initiatives is critical to the businesses that launch them. With the growth of social media's importance in marketing, social media management services have already developed, simplifying the design and analysis of social media marketing campaigns. Additionally, several social media networks research and advise on their platforms' marketing tactics. The effectiveness of advertising campaigns is frequently measured in terms of brand and campaign awareness through metrics such as an increase in followers and mentions following the campaign, view rate and view time, as well as brand sentiment, which refers to the brand's overall perception of social media, not to mention the companies' actual sales figures.

Nowadays, social media is a reflection of what people think about current events and news stories. Any positive or wrong public perception of a company may have a cascading effect on the value of its shares(Mankar *et al.*, 2018). Millions of individuals increasingly post information about their lives on social networking sites like Facebook and Twitter. They get interactive content from online communities where members share knowledge and influence one another. Tweets, status updates, blog posts, comments, and reviews are just some of how social media generates sentiment-rich data. Businesses may also use social media to interact with their consumers for marketing purposes. People make internet purchasing choices based on user-generated content. The sheer volume of user-generated material is staggering. As a result, Sentiment Analysis is required; Sentiment analysis (SA) informs consumers whether or not the information about a product is acceptable before purchase. Marketers and businesses utilize this analytical data to understand better their goods or services to tailor them to the user's needs (A. and Sonawane, 2016).

1.1. Microblogs & Twitter

Microblogging is a kind of blogging in which users post short messages online. A microblog's content is smaller than a conventional blog, both in actual file size and the aggregated file size. The success of micro-blogs may be partly since they enable users to share tiny pieces of information such as brief words, individual pictures, or video links(Kaplan and Haenlein, 2011). Micro-posts are a term for these little communications(Kaplan and Haenlein, 2011; Lohmann *et al.*, 2012). Twitter is one of the most famous and "busy" Microblogs; Twitter is one of the most popular and "active" Microblogs; incredibly, Twitter.com had 5.6 billion global visits in June 2021(Clement, 2021). Additionally, in 2021, the worldwide monetizable daily active international Twitter users reached 169 million, and the social network earned over 3.2 billion US dollars in advertising service revenue, up from almost three billion US dollars the past year(Statista Research Department, 2021).

Because Twitter started as an SMS-based network, the 140-character restriction was first required, then doubled to 280 characters. People create simple websites and write about anything they want, whether politics, sports, cuisine, fashion, etc. A tweet is a message posted. People connect through following one other's Twitter accounts. If you click follow, everything they say will show on your timeline. To tweet someone, use @ before their username. Twitter also allows retweeting; website users retweet other users' tweets to their

followers. Hashtags are widely used on Twitter. These usernames are used to group related tweets. As an example, many individuals attending a conference may want to know what the speakers said. Thus they would use the # sign followed by the conference's name.

Tweets are immediate. A tweet may notify the globe in seconds. Like when Mike Wilson initially tweeted about a Denver aircraft accident in 2008. So, he knew? Twitter is a message. On Twitter, however, users may broadcast their remarks to the whole site rather than just one person. It's also free.

From the above facts regarding Twitter, we may deduce that it is the home of social media news, journalistic discourse, and business analysis. Twitter is critical for information dissemination and has developed into a valuable source of information and a news feed. People are interested in what is occurring on Twitter and in the news, it delivers daily. Additionally, Twitter serves as a forum for sharing perspectives and expressing feelings about current events, news, and goods. Users seek to express their true beliefs on Twitter, making it a valuable source of viewpoints.

Twitter data may be examined and utilized for commercial, policy, and journalism purposes. Often, individuals do not need to read news websites as long as they are logged into their Twitter account, and hot subjects reach the top quicker owing to the high volume of users who refer to them. (Sakaki, Okazaki and Matsuo, 2010) showed via a case study that Twitter may operate as a bursting event monitor, enabling information to spread even quicker than conventional news channels. Individuals and businesses, as predicted, quickly recognized the importance of this data and began using it to make choices.

1.2. The Importance for Sentiment Analysis in Twitter

Knowing what others think about a subject or a product has long been an interest in society. Sentiment Analysis is motivated by two factors. Consumers and businesses alike place a premium on "customer opinion" on goods and services. Thus, both companies and academics have made significant investments in Sentiment Analysis.

We see people expressing their opinions on blogs and forums concerning the customer. These are now actively read by those seeking information about a specific organization. As a result, there are several perspectives accessible on the Internet. Extracting thoughts on a particular entity is critical from a consumer standpoint. However, people cannot possibly wade through such a massive volume of information to comprehend the

general public. As a result, the requirement for a system that distinguishes between positive and negative evaluations is self-evident. Additionally, emphasizing the emotion of these papers provides readers with a short assessment of public opinion regarding an institution.

On the business side, the explosion of Web 2.0 platforms has helped consumers have a platform to share their experiences and views on the brand, positive or negative, about any product or service. Consumer voices can have a massive impact on shaping the opinions of other consumers and, ultimately, on brand loyalty, market decisions, and advocating for their brand. These views shape the future of the product or service. Suppliers need a system to identify trends in customer reviews and use them to improve their products or service and determine future requirements. Therefore, there is a need for Sentiment Analysis systems to detect such phenomena and help both parties.

1.3. Motivation

The primary purpose of the work is to find emotions or polarity in tweets about the famous brand i-phone13to help consumers have a complete picture of it, and companies learn the collective point of view of their audience and become better. An emotion analysis algorithm will be used to achieve this goal. Twitter will do the data collection through API, the language of the comments will be English, and the process will be done in R language.

1.4. Purpose And Outlines

This dissertation consists of five other Chapters. In Chapter 2, we introduce a background of the relevant literature. This chapter will provide in-depth knowledge of research in emotion analysis by first defining the term. We examine each category of emotion analysis in order, i.e., document level, phrase level, and emotion analysis at aspect level. In addition, two effective methods of emotion analysis will be explored in detail: the dictionary-based approach and the machine learning approach. Furthermore, we will look at other research and how they used Machine Learning methods to address the issue of Emotion Analysis using Twitter as a database.

In Chapter 3, the following is the theoretical section, which discusses the algorithms that underpin each machine learning approach and the advantages and disadvantages. We address classification using the supervised machine learning approach - which we also use at work - and the two primary algorithms that offer this technique.

The following part contains thorough information on how we gather, handle, and analyze data from Twitter. The results are reported in Chapter 5, which includes a complete description and discussion of the experimental findings. The last chapter -Chapter 6- summarizes the work, and some suggestions are given for further study.

Chapter 2

2. Literature Review

Sentiment analysis (SA) is the process of extracting sentiment information from natural language text automatically. Over the past decade, there has been a surge of interest in academics(Dave, Lawrence and Pennock, 2003; Andrea Esuli, 2005; Liu and Hu, 2005; Waila *et al.*, 2012; Balage Filho and Pardo, 2013; Goel, Gautam and Kumar, 2017). Sentiment analysis systems and apps have been used for various reasons, including assessing the polarity of customer evaluations, monitoring political sentiments, and forecasting stock market moves.

Assuming that there is a wealth of study material available in "Sentimental Analysis," a bibliography was compiled for this research based on our point of view. More precisely, we sought out research texts that have been cited in many papers. The search process was as follows: initially, we looked for broad characteristics of "Sentiment Analysis," and then narrowed our search to keywords like "Twitter" and "machine learning methods" that frame the field of "Sentiment Analysis," and it's also the field that is going to be discussed in this research.

This chapter will offer in-depth knowledge of the research in sentiment analysis by first defining the term. Since sentiment analysis research has been performed at different levels, this chapter examines each class of sentiment analysis in turn, namely document level, phrase level, and aspect level sentiment analysis. Additionally, two effective sentiment analysis methods will be explored in detail: the lexicon-based and machine learning approaches. Moreover, we will look at other research and how they used Machine Learning methods to solve the issue of Sentiment Analysis.

2.1. Sentiment Analysis

Sentiment analysis is the systematic identification, extraction, quantification, and study of emotional states and personal information via natural language processing, text analysis, computational linguistics, and biometrics. The Merriam-Webster dictionary (*Merriam-Webster*, 2005) defines' sentiment' as 'an attitude, idea, or judgment inspired by emotion. It

is also described as a particular perspective or notion: opinion and emotion (*Merriam-Webster*, 2005). The term 'opinion' refers to a mental perspective, judgment, or assessment made about a specific subject. Opinions are often subjective statements that reflect people's emotions, evaluations, or feelings about things, events, and their characteristics, according to (Li and Liu, 2010). Although the area of sentiment analysis and opinion mining has garnered considerable attention from academics and marketers in recent years, there has been a consistent undercurrent of interest in understanding views. Much of the early research on textual information processing concentrated on factual information mining and retrieval, such as information retrieval, text categorization, or text clustering. One of the primary causes for the dearth of sentiment analysis research is the scarcity of opinionated material before the advent of the World Wide Web. A few years ago, an individual often used to consult their friends or family before making a choice, while a company typically conducts opinion polls and surveys in focus groups to ascertain the general public's views about its goods or services(Li and Liu, 2010) instead today people can share their ideas and feelings online.

It is important to note that three factors contributed to the explosion of emotion analysis research: first, the rise of machine learning methods in natural language processing and information retrieval. Second, the availability of datasets for the development of machine learning algorithms due to the development of the World Wide Web and, more specifically, the development of revision aggregation isotopes. And the third manifestation of the area's intriguing intellectual problems, commercial uses, and intelligence applications (B Pang, 2008).

People now have unparalleled outlets and power to express their views and brand experiences about any product or service, thanks to the development of Web 2.0 platforms such as blogs, Twitter, Facebook, and many other kinds of social media. Additionally, businesses may adapt their marketing tactics due to social media monitoring and research. However, locating and monitoring view sources on the World Wide Web may be a daunting job, given the abundance of different sources such as online forums, discussion groups, and blogs. Additionally, each source may include a significant amount of user-generated material expressing sentiments or emotions.

Sentiment Analysis developed due to the massive amount of information exchanged on the Internet. Nasukawa was the first to introduce the concept of sentiment analysis(Nasukawa and Yi, 2003). To begin, the SA is used for natural language processing

(NLP), which analyzes the views, emotions, and responses of individuals and authors on the Internet through social networking and commercial websites about various goods and services(Hussein, 2018). Sentiment analysis is a vast area of study for many academics and is often referred to as opinion mining since it assists in categorizing thoughts and views as positive, negative, or neutral. Some researchers argue that "the history of the term emotion analysis somehow reflects the view of opinion mining." (B Pang, 2008), as many academics have used the terms emotion and "view" alike in their publications on automated analysis. Evaluative texts (Peter D. Turney, 2002; Nasukawa and Yi, 2003; Ghose, Ipeirotis and Sundararajan, 2007).

Sentiment Analysis is a textual analysis often utilized in online reviews, surveys, and social media. It manages replies and customer feedback on commercial websites to ascertain a customer's approval or rejection of a product; this assists the business in increasing sales by indicating the consumer's preference. With the emergence of diverse viewpoints through social networking sites, new ideas are produced by systems, politicians, psychologists, manufacturers, and researchers for analysis and implementation. Sentiment analysis is an efficient process that utilizes natural language processing, statistics, and machine learning techniques to extract and characterize sentiment information inside a text unit fund in social media.

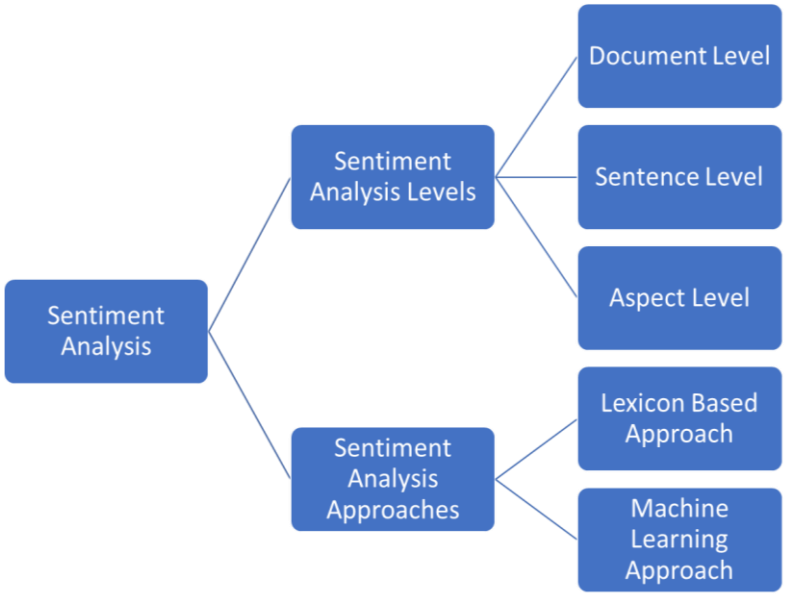

Figure 1: Sentiment Analysis Techniques

2.2. Sentiment Analysis Levels

Sentiment analysis has mainly been studied at three distinct levels, based on the granularity of prior research: document level, sentence level, and feature level(B Pang, 2008).

2.2.1. Document Level

At this level, the goal is to determine if a whole opinion paper conveys a favorable or unfavorable emotion. For instance, the system evaluates if the review reflects an overall favorable or unfavorable impression of the product given a product review. This is often referred to as document-level sentiment categorization. This level of analysis presupposes that each document represents a single entity's viewpoint. As a result, it is inapplicable to papers that assess or compare numerous entities. Numerous academics have conducted document-level sentiment analysis(Pang *et al.*, 2002; Melville, Ox and Lawrence, 2009). They mainly discuss how to automatically distinguish positive and negative texts and offer several techniques for increasing the accuracy(Peter D. Turney, 2002).

2.2.2. Sentence Level

At this level, the goal is to examine each phrase and decide if it conveys a positive, negative, or neutral viewpoint (Aue and Gamon, 2005). The reality is that there is no fundamental difference between document-level and sentence-level sentiment analysis since sentences are just a subset of longer documents. One frequent erroneous assumption made by academics about sentence-level analysis is that a sentence often includes a single viewpoint (although not valid in many cases). Typically, a document contains several perspectives (Liu, 2012).

The task of Sentence level is usually divided into two sub-tasks. The first is to determine if the statement is subjective or objective; the second is to determine whether the sentence conveys a positive or negative view if it is personal (Liu, 2012).

2.2.3. Aspect Level

Both document-level and sentence-level analyses do not reveal precisely what individuals liked and disliked. Aspect level analysis is more precise. Previously, the aspect level was referred to as the feature level(Liu and Hu, 2005). Instead of examining linguistic structures (documents, paragraphs, sentences, clauses, or phrases), the aspect level examines the

viewpoint directly. It is predicated on the premise that an opinion is composed of positive or negative emotion and a target of a statement (Liu, 2012; Joshi and Itkat, 2014).

The aspect level of sentiment analysis focuses on individual views rather than the document's structure, such as paragraphs, sentences, and phrases. It is not sufficient to determine the polarity of the opinions; it is also necessary to identify the opinion targets(Steinberger, Brychcín and Konkol, 2015). The sentiment analysis at the aspect level may be divided into two subtasks (Liu, 2012): aspect extraction and sentiment categorization. Aspect extraction may also be thought of as an information extraction job since it attempts to extract the aspects about which views exist. The fundamental method of removing parts is to look for often occurring nouns or noun phrases specified as aspects. Based on factors, the text is categorized as good, harmful, or neutral (Steinberger, Brychcín and Konkol, 2015). However, sentiment accuracy at the aspect level remains poor due to current algorithms' inability to handle complicated phrases effectively. Thus, sentiment analysis at the aspect level is more challenging than categorization at the document or sentence level(Liu, 2012).

2.3. Sentiment Analysis Approaches

Both document-level and sentence-level analyses do not reveal precisely what individuals liked and disliked. Aspect level analysis is more precise. Previously, the aspect level was referred to as the feature level(Liu and Hu, 2005). Instead of examining linguistic structures (documents, paragraphs, sentences, clauses, or phrases), the aspect level examines the viewpoint directly. It is predicated on the premise that an opinion is composed of positive or negative emotion and a target of a statement (Liu, 2012; Joshi and Itkat, 2014).

The aspect level of sentiment analysis focuses on individual views rather than the document's structure, such as paragraphs, sentences, and phrases. It is not sufficient to determine the polarity of the opinions; it is also necessary to identify the opinion targets(Steinberger, Brychcín and Konkol, 2015). The sentiment analysis at the aspect level may be divided into two subtasks (Liu, 2012): aspect extraction and sentiment categorization. Aspect extraction may also be thought of as an information extraction job since it attempts to extract the aspects about which views exist. The fundamental method of removing parts is to look for often occurring nouns or noun phrases specified as aspects. Based on factors, the text is categorized as good, harmful, or neutral (Steinberger, Brychcín and Konkol, 2015). However, sentiment accuracy at the aspect level remains poor due to

current algorithms' inability to handle complicated phrases effectively. Thus, sentiment analysis at the aspect level is more challenging than categorization at the document or sentence level(Liu, 2012).

2.3.1. Lexicon Based Approach

Opinion mining may be classified into two distinct approaches: lexicon-based and Machine Learning. The lexicon-based method begins at the word level and works its way up to the text's polarity. This method utilizes a sentiment lexicon to determine the polarity of a word. The dictionary-based approach analyzes the text and gives a rating of emotions concerning a dictionary of predetermined emotions (Taboada, Brooke and Voll, 2011). On the other hand, the Machine Learning method starts at the text level and builds a model that gives a polarity score to the whole text; this approach requires a tagged corpus. Essentially, the machine learning method trains classifiers in hand-labelled data (Ruz, Henríquez and Mascareño, 2020). Because learning algorithms depend on the coverage and quality of training data, this method is more demanding and expensive than the dictionary-based approach. Also, the machine learning method often goes beyond the dictionary-based approach to performance.

lexicon-based methods use sentiment lexicons to determine the polarity of a text(Gamallo and Garcia, 2015). A text's sentiment is determined by the words shared between the text and the sentiment lexicons. This function may be the number of positive words divided by the number of negative words; if the ratio is higher than 1, the text is regarded positive; if the ratio is equal to 1, the text is deemed neutral; otherwise, the text is considered negative. Additionally, the semantic orientation (SO) of adjectives, phrases, or grammatical patterns found in the text is a frequently utilized function(Peter D. Turney, 2002).

Much of the early research on lexicons concentrated on using adjectives as markers of text's semantic orientation(Hatzivassiloglou and McKeown, 1997; Hu and Liu, 2004; Kamps *et al.*, 2004). First, a lexicon of adjectives and their associated semantic orientation values is constructed. Then, all adjectives are retrieved and tagged with their semantic orientation value for each given text using the dictionary scores. The semantic orientation ratings are then combined to provide a single text score.

The primary disadvantage of this method is that it lacks a mechanism for dealing with context-dependent terms. For instance, the word "Long" may communicate both a good

and negative impression, depending on the context in which it is used. For instance, consider the following two sentences: "This phone takes a long time to charge," which expresses a bad view, while "This phone has a long battery life" expresses a good attitude(Rahman Khan, Siddique and Basir, 2021). As a result, we argue that the Lexicon-Based method necessitates the availability of robust linguistic resources(Devika, Sunitha and Ganesh, 2016).

2.3.2. Machine Learning Approach

Machine learning is a field of computer science(Wang, 2010) that was created by studying pattern recognition and computational learning theory in the context of artificial intelligence(William L. Hosch, 2021). Arthur Samuel described machine learning in 1959 as "an area of research concerned with the capacity of computers to learn without being explicitly programmed"(Simon, 2013). Machine learning is the study and development of algorithms capable of learning from and making predictions about data. These algorithms operate by building models from experimental data to make predictions or make results-based choices.

The proliferation of machine learning methods in natural language processing has increased the prevalence of sentiment analysis research. In the machine learning method, a textual feature representation was combined with various algorithms such as Naive Bayes, Support Vector Machines (SVM), and Maximum Entropy, often used to construct sentiment analysis classifiers. These classifiers, built using various algorithms, may be trained on training data to learn the rules or decision criteria for sentiment classification and then used to automatically perform sentiment analysis (Ghiassi, Skinner and Zimbra, 2013).

The supervised method is a kind of machine learning. Classification of emotions may be seen as a text classification issue. To categorize a text into distinct subjects, such as politics, economics, science, and sports, topic-related terms are the defining characteristics. While in sentiment categorization, words indicating positive or negative views such as good, terrible, happy, and sad are more significant(Bo, Pang. Lillian, Lee Shivakumar, 2002; Liu, 2012).

This shows that the sentiment analysis machine learning method is a kind of supervised learning, in which a significant amount of labelled training data is needed to train the classifier before it is used to classify new information (B Pang, 2008). The rationale behind the sentiment analysis method based on machine learning is simple. It is built on the framework of supervised classification and consists of two stages: learning the model

through algorithms from a corpus of labelled training data and categorizing new data using the learned model(S, Bird, E, Klein, E, 2009). In practice, the classification job is broken down into many subtasks, including data preparation, feature selection, representation, classification, and post-processing(Khairnar and Kinikar, 2013).

The disadvantage of machine learning-based techniques is often concentrated on the human labelling needed across large datasets of tweets. Additionally, the labelling must be done for each unique area of interest to ensure that the classifier has sufficient training for that domain(Aue and Gamon, 2005).

Typically, machine learning algorithms are classified as supervised, unsupervised, semi-supervised and reinforcement.

Supervised Learning

Supervised techniques need a training set of texts with manually given polarity values, and they learn the characteristics associated with the value from these instances(Neri *et al.*, 2012).

Unsupervised Learning

Unsupervised learning: The learning algorithm is not given labels, leaving it to its own devices to discover structure in the data. Unsupervised learning may be used as a means to an end or as a means of achieving(Peter D. Turney, 2002).

The primary difference between the two methods is the label datasets. Simply stated, supervised learning algorithms make use of labelled input and output data, while unsupervised learning algorithms make use of unlabeled data. In supervised learning, the algorithm "learns" from the training dataset by generating predictions and adjusting for the proper response repeatedly. While supervised learning models are more accurate than unsupervised learning models, they involve manual data labelling in advance. In a survey conducted and using both techniques in film reviews, the results showed that supervised techniques yielded about 85% accuracy, while unsupervised methods yielded about 77%(Chaovalit and Thou 2005).

Semi-Supervised

Semi-Supervised machine learning is a middle ground between supervised and unsupervised machine learning. The system is taught using a mixture of labelled and unlabeled data in this kind of learning. This combination will often include a small quantity of labeled data and a

significant amount of unlabeled data. This is advantageous for many reasons. To begin, categorizing large amounts of data in preparation for supervised learning is often prohibitively time consuming and costly. Additionally, excessive labelling may introduce human biases into the model. Adding a large amount of unlabeled data during the training process enhances the final model's accuracy while decreasing the time and cost required to construct it(Gamon, Michael, Anthony Aue, Simon Corston-Oliver, 2005).

Reinforcement

Reinforcement learning happens when the algorithm is presented with unlabeled instances, as in unsupervised learning. However, you may complement an example with favorable or negative comments based on the algorithm's recommended answer. Reinforcement learning is associated with applications that need the algorithm to make choices with consequences. In the human world, this is analogous to trial and error learning. Errors aid in learning because they impose a penalty (cost, loss of time, regret, and suffering, for example), demonstrating that one line of action is less likely to succeed than another(Grosan and Abraham, 1997).

2.4. Sentiment Analysis Using Machine learning in Twitter

In a study, researchers used machine learning methods to develop a model for detecting emotion on Twitter, successfully using feature sets and improving accuracy, namely bigram, unigram, and object-oriented capabilities. The tweets are classified using two methods, namely Naive Bayes classifiers and support vector machines (SVM); the accuracy is determined by calculating the accuracy, recall, and F rating, which also exhibits the same accuracy(Andrea Esuli, 2005).

Furthermore, another research created a dataset by Twitter API and collected all tweets regarding the topic of the blue whale game. Their main aim is to perform analysis on sentimental tweets. They have used Naïve Bayes, Support vector machines, Maximum entropy, and Ensemble classifier. SVM and Naive Bayes classifiers are implemented using MATLAB built-in functions. Maximum Entropy classifier is implemented using Maxent software. Based on comparative results, Naïve Bayes has better precision and slightly lower recall, and accuracy, i.e., 89%, and other classifiers have similar accuracy levels, i.e., 90%. (Sayali P. Nazare, Prasad S. Nar, Akshay S. Phate, 2018). The result shows the pie chart representing the positive, negative, and neutral hashtags with percentages

Pak et al. built a Twitter corpus by collecting tweets and automatically annotating them with emoticons using the Twitter API. They developed a sentiment classifier using that corpus based on the multinomial Naive Bayes classifier and features N-grams and POS-tags. There is a possibility of inaccuracy with that approach since the emotions of tweets in the training set is classified simply by the polarity of emoticons. Additionally, the training set is inefficient since it includes only emoticon-containing tweets(Pak and Paroubek, 2010).

Chapter 3

3. Theoretical Section

After examining the theory of machine learning in the previous chapter, we will mention why we chose the machine learning method for emotion analysis; we will describe some classification techniques and algorithms created for Emotion Analysis. We will focus on the naive Bayes method, which we used to develop our approach.

3.1. Sentiment Analysis as a Classification problem

Many researchers have used machine learning for a Sentiment Analysis problem like (Andrea Esuli, 2005; Pak and Paroubek, 2010; Sayali P. Nazare, Prasad S. Nar, Akshay S. Phate, 2018); the truth is that customer sensitivity and underlying opinions may be recorded and analyzed more effectively with the use of machine learning. Another significant advantage of Machine Learning is the algorithms' training capability. Natural Language Processing (NLP) is used in conjunction with sentiment libraries, sentiment corpora, and other human-annotated sentiment rules to continually enhance algorithms, making them quicker and more accurate.

Classification of emotions is essentially a text classification issue. However, conventional text categorization primarily categorizes publications according to politics, science, or sports. The crucial aspects of these categories are the terms associated with the subject. When classifying emotions, it is more vital to choose words that convey positive or negative feelings, such as fantastic, exceptional, shocking, dreadful, harmful, or worse(Liu, 2015).

Perhaps the most thoroughly researched subject is the classification of emotions(B Pang, 2008). Since sentiment classification is a text classification problem, it may be handled using any known supervised learning methodology; however, there are also unsupervised techniques. Sentiment regression has been primarily performed using supervised learning.

Numerous current supervised classification methods are capable of being used in Sentiment Analysis. These strategies are discussed in detail below, along with the benefits and drawbacks associated with each. Sentiment analysis may use any of the following

categorization algorithms: K-NN, Support Vector Machines, Decision Trees , Logistic Regression, Naive Bayes.

3.1.1. K - Nearest Neighbors

In practice, closest neighbor classifiers are well-suited for classification problems in which the links between the characteristics and the target classes are many, intricate, or very difficult to comprehend. However, the objects belonging to the same class type are relatively homogenous. Another way to express it is that closest neighbors may be acceptable if a notion is difficult to describe but intuitively understood. On the other hand, if the data is noisy and no apparent differentiation between the groups exists, the closest neighbor algorithms may have difficulty identifying the class borders.

K-Nearest-Neighbors (KNN) is a basic but effective nonparametric supervised classification technique. The KNN classifier is the most widely used pattern recognition due to its high performance, efficiency, and simplicity. It is extensively utilized in pattern recognition, machine learning, text classification, data mining, and object identification, among other applications (Pedro J., Garcı´a-Laencina a, Jose -Luis, Sancho-Gomez, Anıbal R. and Verleysen, 2009). The KNN algorithm classifies by analogy, that is, by comparing the unknown data point to comparable training data points. Euclidean distance is used to quantify similarity. The attribute values are standardized to ensure that characteristics with more comprehensive ranges do not overwhelm those with lower capacities. KNN classification assigns the unknown pattern to the most prevalent class among its nearby neighbours' classes. If there is a tie for the practice between two classes, the class with the shortest average distance to the unknown pattern is allocated. A global distance function "dist" may be computed by combining many local distance functions depending on specific parameters(Hota and Pathak, 2018). As seen in Equation (1.1), the easiest method is to add the values together:

$$dist.(X^T, X) = \sum_{i=1}^{n} distA_i(X^T \cdot A_i, X \cdot A_i) \qquad (3.1)$$

X^T is the test tuple, X denotes the closest neighbor, and A_i (i=one to n) represents the attributes of the data points. Global distance is defined as the weighted sum of local distances. Specific weights w_i may be provided to the characteristics A_i to indicate their relative relevance in determining the proper classes for the samples. Consequences are

typically between 0 and 1. Weight 0 is allocated to irrelevant qualities. As a result, Equation (1.1) may be rewritten as Equation (1.2)(Hota and Pathak, 2018):

$$dis(X^T, X) = \sum_{i=1}^{n} W_i \times distA_i(X^T \cdot A_i, X \cdot A_i) \quad (3.2)$$

The Equation gives the average weighted distance:

$$avgdist(X^T, X) = \sum_{i=1}^{n} w_i \times distA_i(X^T \cdot A_i, X \cdot A_i) \quad (3.3)$$

$$\sum_{i=1}^{n} w_i \quad (3.4)$$

The following table lists the strengths and weaknesses of the K-NN algorithm.

Table 1: Advantages and disadvantages of K-NN algorithm.

Advantages	***Disadvantages***
K-NN is very quick since it does not need training compared to other algorithms that do. Rather than that, it reloads training data and continuously learns from it (Lazy Learner).	*When dealing with massive datasets, the cost of computing the distance between a new point and an old issue is enormous, degrading the algorithm's speed.*
Because the KNN algorithm does not need training before producing predictions, new data may be supplied without affecting the system's accuracy.	*The KNN technique does not perform well with high-dimensional data because it becomes more challenging for the algorithm to compute the distance in each dimension as the number of dimensions increases.*
To implement KNN, just two parameters are required: the value of K and the distance function.	*KNN is very susceptible to noise in the training set. Manually imputing missing numbers and eliminating outliers is required.*

3.1.2. Support Vector Machines

Support Vector Machines (SVMs) are a supervised machine learning technique often used to solve classification issues. Many researchers have used SVM as a sentiment method (Pak

and Paroubek, 2010; Devika, Sunitha and Ganesh, 2016; Ruz, Henríquez and Mascareño, 2020). SVM method displays each data item as a point in n-dimensional space (where n is the number of features we have), with the value of each element being the coordinate value. Then, we accomplish classification by identifying the hyperplane that best distinguishes the two classes.

The fundamental principle behind SVM for sentiment classification is to identify a hyperplane that separates documents, or in our instance, tweets, according to their sentiment, with a maximum margin between the classes(Bhuta *et al.*, 2014). For example, if we need to express a training set mathematical, we have:

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_{m,} y_m)\} \quad (3.5)$$

Xi is an n-dimensional real vector, and yi is either 1 or -1, representing the class where xi belongs.

The following table lists the strengths and weaknesses of the SVM algorithm.

Table 2: Advantages and disadvantages of SVM algorithm.

Advantages	***Disadvantages***
In high-dimensional spaces, SVM is more effective.	*The SVM method is inefficient for big data sets.*
Support Vector Machines may be slow while training, particularly with massive datasets, but they are relatively quick when predicting.	*SVM performs poorly when the data set contains more noise, i.e., when target classes overlap.*
SVM is a memory-efficient algorithm.	*There is no probabilistic justification for the classification since the support vector classifier operates by placing data points above and below the classifying hyperplane.*

3.1.3. Decision Trees

Decision trees are considered one of the most common methods of solving problem categorization, and many researchers have included them in their research (Prasad *et al.*, 2015; Bayhaqy *et al.*, 2018; Mankar *et al.*, 2018).

Decision tree classification is a nonparametric technique that divides a dataset into classes by recursive partitioning (Quinlan, 1986; Rokach and Maimon, 2005). A decision tree is a set of rules that divides a dataset into increasingly smaller groups for categorization or prediction. A decision tree is a very efficient way for generating classifiers from data. The Decision Tree is a flowchart structure like a tree, with each internal node testing an attribute, each branch representing the test's findings, and each leaf node stating the class label. At the same time, the root node is the node at the top of the decision tree.

The following table lists the strengths and weaknesses of the Decision trees algorithm.

Table 3: Advantages and disadvantages of Decision trees algorithm.

Advantages	***Disadvantages***
Pre-processing data for decision trees is less time consuming than for other techniques.	*The decision tree structure might be unstable due to a tiny change in the data.*
The process of generating a decision tree is unaffected by missing values in the data.	*The cost of decision tree training is high because of the intricacy and time it takes.*
A decision tree model may be explained to technical teams and stakeholders without the need for analytical, mathematical, or statistical know-how straightforwardly and understandably.	*Regression and continuous value prediction cannot be made well using the Decision Tree method.*

3.1.4. Logistic Regression

Logistic Regression is an extended regression model for binary dependent variables (such as true or false, yes or no) (Kantardzic, 2011). For example, the Likelihood that the dependent variable would be positive or negative may be estimated using logistic Regression.

Table 4: Advantages and disadvantages of Logistic Regression algorithm.

Advantages	***Disadvantages***
Logistic Regression is a straightforward prediction algorithm. It is also transparent, in contrast to more complicated processes that are more difficult to watch, in that we can see through them and comprehend what is going on at each stage.	*Not all issues can be handled this way; logistic Regression cannot tackle non-linear situations. As a result, converting these non-linear issues to linear problems may be time consuming and inefficient.*
Logistic Regression is less susceptible to overfitting in low-dimensional datasets with sufficient training samples.	*Data upkeep is difficult in logistic Regression due to the lengthy nature of data preparation.*
When the dataset contains characteristics that can be separated linearly, Logistic Regression performs well.	*Because logistic Regression is not as robust as other algorithms, it is likely to have trouble capturing complicated correlations.*

3.2. Bayes' Theorem

The Bayes algorithm is a machine learning classic. An English mathematician and statistician, Thomas Bayes, created Bayesian classifiers (1702, London - 1761, Tunbridge Wells, United Kingdom). He made the first mathematical framework for calculating the possibility of a recurrence based on the number of occurrences.

It is necessary to keep in mind that Bayes inference methods do not generally provide crisp, unequivocal answers of the type 'an unlabeled event a_i belongs just to a class C_j with the probability $p(a_i) = 1.0$ and certainly not to another class $C'_k \neq j$. Usually, for such a probability $p(a_i)$, it holds that $0.0 < p(a_i) < 1.0$, even if either $p(a_i) = 0.0$ or $p(a_i) = 1.0$ is also quite possible but not typical in practice(Žižka, Dařena and Svoboda, 2019).

Commonly, Bayesian classifiers are used to determine an event's overall probability of occurrence utilizing data from several attributes. Unlike many machine learning algorithms, Bayesian techniques make minor adjustments to predictions based on all

available evidence. If multiple qualities have a negligible effect, their combined effect may be relatively significant(Brett, 2015).

A well-known foundation for classification is given by a basic probability theory referred to as Bayes' theorem or Bayes's rule(Aggarwal and Xhai, 2012). Before discussing Bayes' Theorem, first, let recall two essential laws of probability theory (Aggarwal, 2015):

$$p(X) = \sum_{Y} p(X, Y) \tag{3.6}$$

$$p(X, Y) = p(Y \mid X)\, p(X) \tag{3.7}$$

In which the sum rule is the first Equation, and the product rule is the second Equation. Here, p(X,Y) denotes a joint probability, p(Y|X) denotes a conditional probability, and p(X) denotes a marginal likelihood. The following Bayes' theorem is easily obtained using the product rule and the symmetry condition p(X,Y) = p(Y,X).

$$p(Y \mid X) = \frac{p(X \mid Y)\, p(Y)}{P(X)} \tag{3.8}$$

The $p(Y \mid X)$ is the Posteriori probability, $p(X \mid Y)$ is the Likelihood, $p(Y)$ is the Class prior probability, and $p(X)$ the Prediction prior probability. Machine learning, particularly classification, relies heavily on Bayes' theorem. Bayes' theorem denominator may be stated in terms of the numerator's numbers with the sum rule $p(X) = \sum_{Y} p(X \mid Y)\, p(Y)$. Bayes' theorem's denominator may be thought of as the necessary normalization constant to guarantee that the total of the conditional probability on the left-hand side of Equation (3.8) across all possible Y values equals one.

3.3. Naïve Bayes Classifier

Naive Bayes is a classification approach based on Bayes' Theorem that assumes predictor independence. Said, the existence of a specific feature inside a class does not imply the presence of another feature within a Naive Bayes classification. Because the Nave Bayes classification technique is successful and straightforward, it is suitable for classifying even large data sets. The naive Bayes classifier is the simplest Bayesian classifier, and it has developed into a significant probabilistic model, despite its high independence

requirement(Bishop, 1995; Di Nunzio, 2009). Simple Bayesian classifiers have gained attention in recent years and are astonishingly effectively implemented by many academics. Let's describe the model of Naïve Bayes classifier. Assume we have a collection of training sets $\{(x^{(i)}, y^{(i)})\}$ containing N examples, where $x^{(i)}$ signifies a d-dimensional feature vector and $y^{(i)}$ gives an example's class label. Assume that Y and X are random variables with components $X_1,...,X_d$ corresponding to the label y and a feature vector $x = \langle x_1, x_2, ..., x_d \rangle$. Nota bene, the superscript indexes training samples for $i = 1,..., N$, while the subscript indexes each feature or random variable in a vector. In general, Y is a discrete variable that belongs to precisely one of the K possible classes $\{C_k\}$ $k \in \{1,..., K\}$ and $X_1,..., X_d$ may have any discrete or continuous characteristic. Our objective is to train a classifier that outputs the posterior probability p(Y|X) for all possible Y values. Bayes' theorem states that $p(Y = C_k | X = x)$ may be expressed as(Aggarwal, 2015):

$$p(Y = C_K \Big| X = x) = \frac{p(X = x | Y = C_k)\, p(Y = C_k)}{p(X = x)}$$
$$= \frac{p(X_1 = x_1, X_2 = x_2, ..., X_d = x_d | Y = C_k)\, p(Y = C_k)}{p(X_1 = x_1, X_2 = x_2, ..., X_d = x_d)} \quad (3.9)$$

For several reasons, this formula is computationally challenging to solve. As additional features are added, tremendous amounts of memory are needed to store probabilities for all possible intersecting events; Imagine how difficult it would be to draw a Venn diagram for just four words, much more for hundreds or thousands(Brett, 2015).

The study is enhanced further when we use the fact that Naive Bayes maintains case equality. Which mean that the Naive Bayes classifier reduces this complexity by making a conditional independence assumption that the features X_1,..., X_d are all conditionally independent of one another, given Y. Thinking conditional independence allows us to simplify the formula using the probability rule for independent events, which is: P(A ∩ B) = P(A) * P(B). The denominator is assumed to be a fixed value and can be ignored for the time being. Consider the likelihood p(X = x|Y = C_k) of Equation (3.9), we have:

$$
\begin{aligned}
& p(X_1 = x_1, X_2 = x_2, ..., X_d = x_d \mid Y = C_k) \\
&= \prod_{j=1}^{d} p(X_j = x_j \mid X_1 = x_1, X_2 = x_2, ..., X_{j-1} = x_{j-1}, Y = C_k) \\
&= \prod_{j=1}^{d} p(X_j = x_j \mid Y = C_k)
\end{aligned} \tag{3.10}
$$

The second line is derived from the chain rule, a general property of probabilities, and the third line is derived directly from the preceding conditional independence, which states that the value of the random variable X_j is independent of all other feature values, $X_{j'}$ for $j' \neq j$, when conditioned on the label Y identity. This is referred to as the Naive Bayes hypothesis. It is a reasonably robust and very valuable assumption. When Y and Xj are boolean variables, the definition of p(Xj|Y = C_k) requires just two double arguments(Aggarwal, 2015). By replacing Equation (3.10) for Equation (3.9), we may derive the Naive Bayes classifier's basic Equation.

$$
p(Y = C_k \mid X_1 ... X_d) = \frac{p(Y = C_k) \Pi_j p(X_j \mid Y = C_k)}{\Sigma_i p(Y = y_i) \Pi_j p(X_j \mid Y = y_i)} \tag{3.11}
$$

If all we care about is the most likely value of Y, then we have:

$$
Y \leftarrow \arg\max_{C_k} \frac{p(Y = C_k) \Pi_j p(X_j \mid Y = C_k)}{\Sigma_i p(Y = y_i) \Pi_i p(X_j \mid Y = y_i)} \tag{3.12}
$$

The above formula (3.12) can be simplified since the denominator is not dependent on C_k.

$$
Y \leftarrow \arg\max_{C_k} p(Y = C_k) \prod_j p(X_j \mid Y = C_k) \tag{3.13}
$$

3.3.1. Types of Naive Bayes Classifier

Perhaps the simplest and most often used generative classifier is the Naive Bayes classifier. It uses a probabilistic model with different assumptions about the distributions of different words to describe the distribution of documents within each class. There are two types of models often employed in naive Bayes classification the Bernoulli and the Multinomial models. Both approaches calculate a class's posterior probability based on the word distribution in the text. These models do not attempt to account for the actual placement of the words in the text, but they rely on the "bag of words" assumption. The essential distinction between these two models is the assumption of the inclusion (or exclusion) of word frequencies and the corresponding strategy to sample the probability space.

On the one hand, the Bernoulli model represents a document using the presence or absence of words in a text document as attributes. Thus, word frequencies are not used in document modelling, and the characteristics of the words in the text are assumed to be binary, with the two values indicating the presence or absence of a word in the text. Conversely, we use the Polynomial model to keep track of the Frequency of words in a text by representing it with a bag of words. Therefore, each class's documents may be described as samples from a polynomial word distribution. On the other hand, the conditional probability of a document belonging to a class is just the product of the possibilities of each observed word inside that class(Aggarwal and Xhai, 2012).

While the multivariate Bernoulli model works well with small vocabulary sizes, the multinomial model often performs better with bigger vocabulary sizes, giving a 27 per cent decrease in error over the multivariate Bernoulli model at any vocabulary level(Mccallum and Nigam, 1998).

3.4. Naïve Bayes Pros & Cons

As previously stated, the Naive Bayes Classifier is a straightforward method that may tackle various classification issues. Below are the advantages and disadvantages of Naive Bayes.

Advantages:

- The Naive Bayes method is rapid and can accurately predict the class of a test dataset.
- We may use Naïve Bayes to handle issues involving several class predictions since it is helpful with them.
- We need a small quantity of test data to estimate the training data using the Naive Bayes method. So we have a shorter training duration results.
- A Naive Bayes classifier outperforms other models because the assumption of independent predictors is valid.

Disadvantages:

- The premise of independent predictors is the most striking characteristic of Naive Bayes. All of the qualities are assumed to be independent in Naive Bayes. But entirely independent predictors are very hard to find in the actual world.

- If a categorical variable in the test data set contains a category not seen in the training data set, the model will give it zero probability and be unable to predict. This is often referred to as Zero Frequency.

Sentiment Analysis issues may be binary or multiclass classification problems, implying that an event may have two or more possible outcomes. This kind of issue dictates that we use classification techniques such as K-NN, Logistic Regression, Support Vector Machines, and Naive Bayes, among others. When Naive Bayes is used compared to other algorithms, it produces the best results.

Chapter 4

4. Analysis of Data

This chapter has to do with the data we used, the data includes tweets that refer to the iPhone13; the reason for choosing this hashtag was that there is a fairly large number of tweets every day. More specifically, this chapter describes in detail the process of collecting Tweets and characterizing them for their polarity. It also describes the preparation and cleaning process required to be done on the data before it applies Naive Bayes.

4.1. Data collection via API

Twitter is not only a flexible communication platform for people around the world; it is also an incredible storehouse of knowledge and information. Researchers from different backgrounds use Twitter data to answer various questions, ranging from simple information about people or events to more complex ones. Twitter is a snapshot of what is happening globally and discussed right now. Anyone can access Twitter via the web or a mobile device. Twitter allows programmable access to Twitter data (application programming interfaces). At a high level, APIs are the means through which computer applications "communicate" with one another to request and provide information. This is accomplished by enabling a software program to access what is referred to as an endpoint: an address associated with a particular sort of information we supply. Twitter provides API access to several aspects of the service to enable developers to create software that connects with Twitter(Twitter, Guide and Us, no date).

Twitter API subscriptions are available in three levels: Standard, Premium, and Enterprise. The Standard is entirely free and includes all the necessary functions for our research purposes. Specifically, Twitter allows us to export comments for the current day or the last seven days. The tweets are returned in JSON format,16 a widely used data storage and exchange standard, and hence readily interpreted by a wide variety of computer languages.

We accessed Twitter via API using the R (R is a free software environment for statistical computing and graphics) and retrieved 1300 comments referring to # iPhone13

using the *rtweet* package (Kearney, 2020). We downloaded comments for about a month and almost every day. The raw data was then cleaned to eliminate duplicate entries and tweets comprised entirely of hashtags, photos, or URLs. After this phase, there were still 1.300 tweets remaining. The comments were moved to a CSV file that can be modified and interpreted using R statistical software. The CSV file had 15 columns. the following table summarizes the available data. (Workshop, Aug and Fraley, 2016)

Table 5: Information that is available in the CSV.

Column Name	***Interpretation***
Text	This is the text that was included in the tweet.
Favorited	Likes
FavoriteCount	Count how many times the tweet has been liked.
ReplyToSN	If the tweet was a response to another user, the screen name is written.
Created	Date & Time the tweet was posted
Truncated	When a Tweet is longer than 140 characters. Uses a True/False values
ReplyToSID	If this tweet responded to a previous tweet, this provides the ID of the tweet in question.
Id	The tweet's unique identifier. Each tweet should be assigned a unique ID number.
ReplyToUID	If this tweet was a response to a previous tweet, this contains the ID for the account in question.
StatusSource	What is the origin of the tweet? For example, iPhone, Android.
ScreenName	The user's screen name
RetweetCount	How many times the tweet has been retweeted
IsRetweet	If the tweet is a retweet. Uses a True/False values
Longitude	Coordinates the tweet's origin if the user has geocoords enabled.
Latitude	Coordinates the tweet's origin if the user has geocoords enabled.

4.2. "Labeling" Process

After data collection, the comments are characterized by their polarity as *positive*, *neutral*, and *negative*. Although the API helps to collect tweets, the "labeling" process is complicated and must be done with extreme care (Giachanou and Crestani, 2016) the difficulty lies in the fact that in a large amount of data, it is easy to make a mistake if there is no re-control and it is therefore very likely that this will have a negligible effect on the outcome.

In our case, we noticed that while adding tags to Tweet: First, most of the tweets made many times Retweets were from individuals or companies who wanted to influence - to entice the public to buy the product or follow the steps to win an award. All of these Tweets had a positive polarity due to the words "I win," "I offer," "I give you," and a sense of urging. In addition, many Tweets were linked to an image or video that did not exist on our CSV; most of these tweets were deleted because we could not name them without all meaning. Last but not least, Twitter users often use idiosyncrasies and abbreviations that do not always make sense; in such cases, the comment was ignored. In Table 6, we can see details about the labeled tweets (any name or nickname used from the users has been omitted).

Table 6: Example for manual coding of polarity.

Code	***Percentage of Polarity (Out of 1300)***	***Examples***
Positive	63,4%	Omg. I love everything about this Iphone13 proMax .. it definitely a big upgrade from the 12.
Neutral	17%	The logo on #apple #iPhone13 should be the fingerprint sensor. What do you think
Negative	19,6%	I have to say… I'm not impressed by the #iPhone13

4.3. Data Preparation

Data preprocessing is a critical stage in Machine Learning because it contributes to data quality improvement, facilitating the extraction of relevant insights from it. The term "preprocessing" in Machine Learning refers to preparing raw data for creating and training models. To put it in a simpler way, data preprocessing is a machine learning technique for data mining that transforms raw data into a legible and understandable format. Data from real-world sources (especially from Twitter because of the informal type of communication and the length limitation) tend to be inconclusive, inconsistent, imprecise, and lacking. Preprocessing raw data helps clean, prepare, and arrange it in Machine Learning algorithms. Let's look at the different phases in our data preparation.

The first stage in developing our classifier is to prepare the raw data for analysis. Text data are difficult to prepare since they must be transformed into a machine-readable format. We will convert our data into what is known as a bag-of-words model, which ignores word order and merely offers a variable indicating if the term exists at all. In practice, the Bag-of-words model is mainly used for feature generation. After transforming the text into a "bag of words", we can calculate various measures to characterize the text. The most common type of characteristics or features calculated from the Bag-of-words model is term frequency; namely, the number of times a term appears in the text.

Once we turn the data into a "word bag", it's time to clean it up. We must explore how to eliminate numbers and punctuation, deal with meaningless words like and, but, and or, and split sentences down into individual words. The first stage in text data processing is to create a corpus, which is a collection of text documents. The corpus in this example will be a collection of Tweets. To create a corpus, we'll use the *VCorpus*() function in the *tm* package, which refers to a volatile body (the *tm* function is going to help us in the data cleaning process).

- Our initial priority will be to standardize the messages such that they include just lowercase letters. To do this, R has the *tolower*() function, which returns text strings in lowercase.
- Following is the removal of Url by *removeURL* function.
- The removal of Twitter usernames by *removeTWusers* function.
- The removal of all hashtags by *removeHashtag* function.
- The removal of all numbers by *removeNumbers* function.

- The removal of all punctuation by *removePunctuation* function.
- Our next objective is to eliminate filler terms from our Tweets, such as to, and, but, and or by *removeWords* function. Stop words are terms that usually are eliminated before text mining. This is because they do not supply much helpful information for machine learning despite their frequency of occurrence.
- Finally, another prevalent method of standardizing text data is to reduce words to their root form via a process called stemming, and we use *stemDocument* function. The steaming process removes the suffix from terms such as offered, offering, and offers, transforming them into the offer's basic form. This enables machine learning algorithms to consider similar phrases together rather than discover a pattern for an individual version. Table 7 presents an example of tweets before and after data preparation.

Table 7: Tweets before and after preparation.

Before Preparation	***After Preparation***
Finally! New phone for me after 6yrs! #iphone13 https://t.co/pWKAV0hUBX...	final phone yrs
NO RESPECT MEANS NO PUDDY https://t.co/0CDzgt2BJf.#iphone13...	respect mean puddi
2 months and 3 hours later ?□□... @dbrand #iPhone #iPhone13 https://t.co/Pfxh3rs...	month hour

Chapter 5

5. Analysis and Results

In this section, we describe our suggested model for identifying positive, negative, and neutral tweets, as well as the outcomes of our experiments and evaluations. Using the Naive Bayes algorithm, we divide our tweets into a training and testing set, with 70% of the tweets being used for training and the remaining 30% being used as a testing set. We ignored words that appeared in less than three evaluations; instead, we used the training set to build our model and the testing set to verify its predictions. In order to analyze the overall methodology process, we conducted cross-validation ten times to evaluate the Naive Bayes algorithm.

5.1. Visualization of data

Dataset was created by crawling the comments from 1,300 different tweets. The difference between positive, negative, and neutral tweets is shown in the bar graph in Figure 2. Initially, we notice that this is an unbalanced set of data. There are so many more tweets, positive than neutral or negative. More specifically, our sample consists of 63.4% positive, 17% neutral, and 19.6% negative tweets.

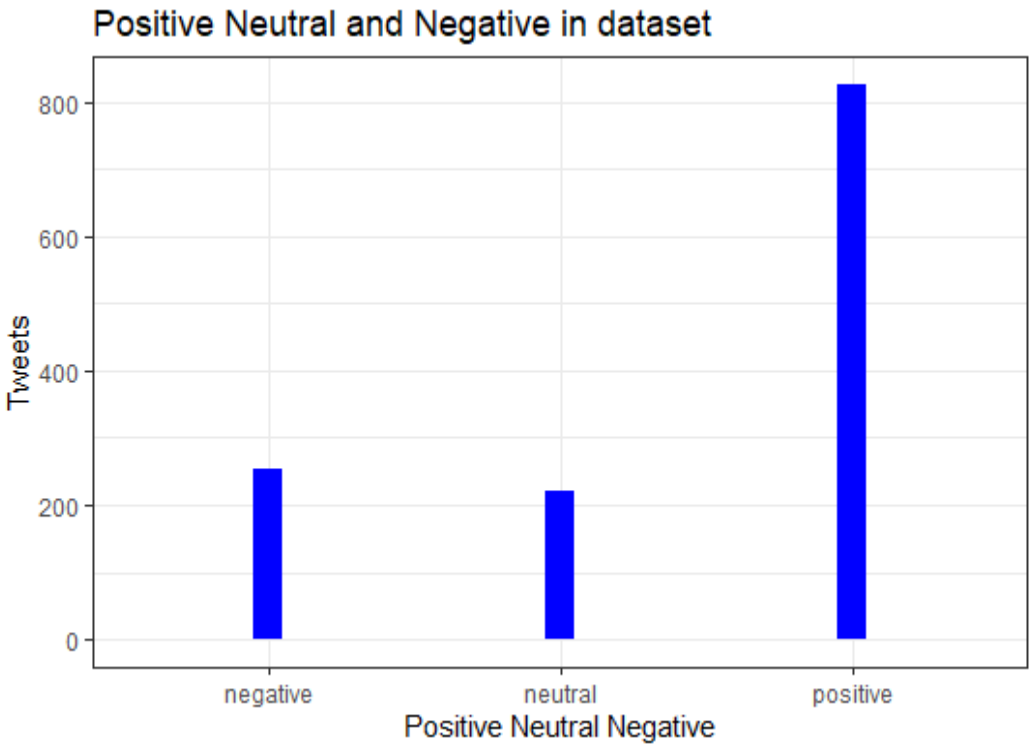

Figure 2: Positive, Neutral and negative tweets in the dataset.

We preface the analysis of emotions with a word cloud from the data set of the top hundred words that appear at least 15 times in our dataset (Figure 3). The word cloud indicates that negative words do not dominate. The most common words are either neutral ("today", "check", "time") or positive ("victory", "good", "love", "gift") this fact is justified as we saw from Figure 2, the positive tweets are much more and may also contain neutral words.

A word cloud is a visualization of the frequency with which certain terms appear in tweets. The cloud is composed of words that have been randomly distributed around the chart. The text's most frequently occurring words are printed in a larger font, while less regularly occurring phrases are written in a smaller font. Recent years have seen an increase in the importance of this type of statistics since it enables tracking current trends on social media platforms. Word clouds constructed for a corpus of literature can serve as a jumping-off point for further investigation. The purpose of the word clouds is to highlight statistically significant terms.

Figure 3: Hundred most common words in the dataset.

The following stage in data visualization is to display the five most frequently used words in all tweets. This is illustrated in Figure 4. A frequently used word diagram is another technique for displaying data that supply us with necessary information via an extensive database. The following graph depicts the five most often occurring terms in our data. What we note is that each of the five terms (win, chance, buy, give, repair) has an impulse component; they are intended to produce an impact, which suggests they are most likely derived from advertising tweets.

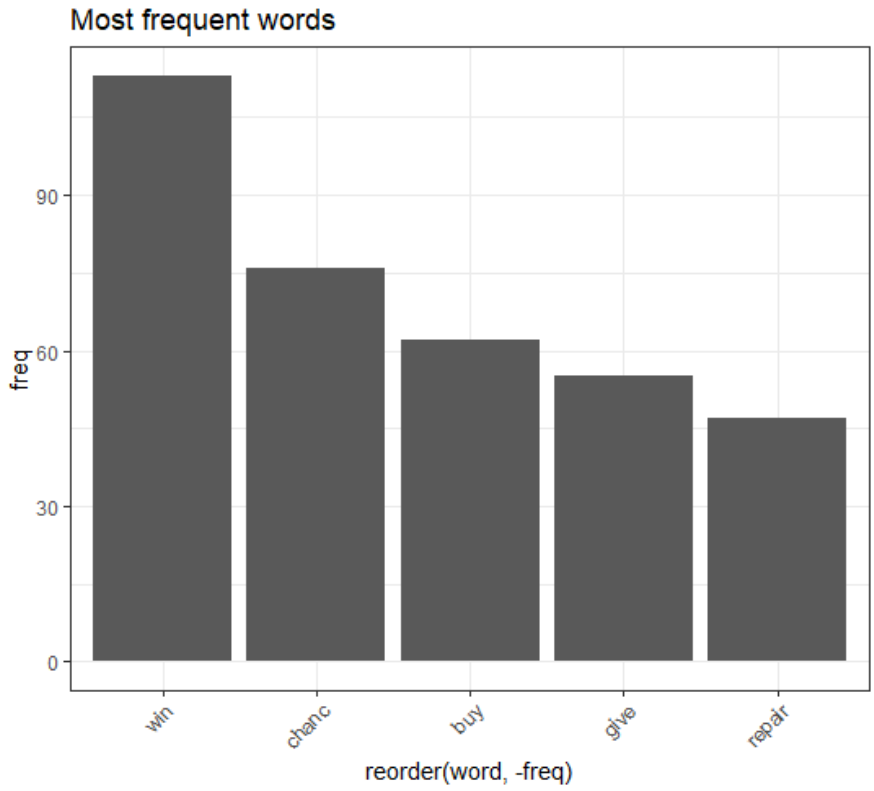

Figure 4: 5 tops of most frequent words.

5.2. Experimental Results and Performance Evaluation

The initial idea was to research with all three polarizations (positive, negative, neutral), but we discovered that the algorithm was not very efficient, so to make it easier for the algorithm, we only worked with negative and positive polarity.

The main goal of any machine learning model is to predict the effect of real-time outcomes, and therefore evaluating the performance of the applied machine learning model becomes crucial in determining whether the model is significant enough to predict the creation of an unknown data point. To evaluate the result of the algorithm, we will use K-Fold Cross-Validation 10 times.

5.2.1. K-Fold CV

Cross Validation is used mainly in applied machine learning to estimate a machine learning model's skill on previously unseen data. That is, to use a small sample to assess how the

model will perform in general when used to generate predictions on data that was not utilized during the model's training.

Cross-validation is a statistical technique that allows for evaluating and comparing learning algorithms by separating data into training and testing sets. The training set is used to acquire knowledge and refine a model. The test set is used to validate the model. Indeed, the preparation and assessment sets must be bridge-examined sequentially to ensure that no data point is overlooked. (Payam Refaeilzadeh, Lei Tang, 2014). The data is separated into k folds of equivalent (or nearly identical) dimensions for k fold validation. The k iterations are then conditioned and verified so that each iteration has a discrete folding of data for validation, while the remaining k-1 folds are used for learning. Until the data is separated into k – folds, it is stratified in such a way that each fold represents the entire dataset rather well.(Payam Refaeilzadeh, Lei Tang, 2014). The following is the procedure we followed, we shuffled the data set, we split the dataset into k groups, next randomly selected the training and testing set, after that, we used the training set to train the naive Bayes classifier, we used the test set to test the prediction and we calculated the accuracy of the forecast. In the end, the mean value of the accuracy will be shown.

Generally, the folds are stratified to ensure that cross-validation appropriately assigns each instance to k distinct folds. The stratification of the folds ensures that the proportion of predictor marks is nearly equal to that of the initial dataset. The provided data set S was randomly partitioned into k mutually exclusive subsets $(S_1,\ldots,S_k)$ of about similar size for k-fold CV. The classifier is trained and tested k times. Each time $t \in \{1,2...,k\}$, it is trained on all but one-fold S_t and tested on the remaining single fold S_t. The overall precision of the cross-validation calculation is determined only as the average of the individual precision measurements k as seen in the Equation 5.1.

$$CVA = \sum_{i}^{k} A_i \qquad (5.1)$$

where CVA stands for cross-validation accuracy, k is the number of folds and A is the accuracy measure of each fold.(Delen, 2009)

5.2.2. Validation, Recall, and F-measure

The true-positive rate, false-positive rate, true-negative rate, and false-negative rate all contribute to the effectiveness of any machine learning model. We utilized the most frequently reported yield metric in this study: accuracy (equation 5.2), which is determined by the criteria mentioned previously.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.2}$$

The positive predictive value (PPV), often called precision, is defined in equation 5.3. Whereas a "true positive" occurs when the test makes a positive prediction, and the subject responds positively, a "false positive" occurs when the test makes a positive prediction, but the subject responds negatively. The PPV's best test value is 1 (100%), and its worst value is zero. As illustrated in Equation 5.4, the negative predictive value is defined as follows. Whereas a "true negative" occurs when the test generates a negative prediction and the subject provides a negative response, a "false negative" occurs when the test generates a negative prediction but the subject provides a positive response. The NPV value of a great test that produces no false negatives is 1 (100%); the NPV value of a poor test, one that produces no true negatives, is nil. (Updating, 2020)(Table, 2020).

$$PPV = \frac{Number \text{ of true positives}}{Number \text{ of true positeves+Number of false positives}} = \frac{Number \text{ of true positives}}{Number \text{ of positive calls}} \tag{5.3}$$

$$NPV = \frac{Number \text{ of true negatives}}{Number \text{ of true negatives} + Number \text{ of false negatives}} = \frac{Number \text{ of false negatives}}{Number \text{ of negative calls}} \tag{5.4}$$

The recall metric is the percentage of all relevant instances that were retrieved (equation 5.5). The F-measure is a metric used to determine the validity of a test. It is decided by the test's accuracy and recall. The F_1 score (equation 5.6) is a harmonic notation used to achieve precision and recall. F_1 can have a maximum value of 1, indicating flawless accuracy and memory, and a minimum of 0, indicating that either precision or recall is zero.

$$\text{Recall} = \frac{\text{Number of true positive}}{\text{Number of true positive+ Number of false negative}} \tag{5.5}$$

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (5.6)$$

5.3. Results

After using the training set to train the naive Bayes classifier and the test set to check the prediction, we then calculated the accuracy of the forecast (for the test set). This study tested our model using the accuracy mentioned above criterion (classification accuracy). For Naive Bayes, the results were obtained using tenfold cross-validation using the mean findings from the test data set (10 times).

Our training model confirms the validity of the Naive Bayes as a classification method, which is proved by previous surveys as well(Bo, Pang. Lillian, Lee Shivakumar, 2002; Maria Arista Ulfa, Budi Irmawati, 2018), achieving a high degree of accuracy. More specifically, we found that the Naive Bayes model had a classification accuracy of 80.8%. The complete set of results is presented in tabular form in Table 8.

Table 8: Classification results.

Naïve Bayes	
Fold	**Accuracy**
1	79%
2	79.6%
3	79.3%
4	80.6%
5	79%
6	81.5%
7	85.5%
8	81.2%
9	79.6%
10	82.7%
mean	80.8%

The comprehensive estimation results of the validity sets are given in confusion matrices. A matrix of uncertainty reflects the effects of classification. In a two-class

prediction problem, such as in our research, the upper-left cell represents the number of positive samples when they are positives (i.e. true positive), and in the lower right section, the number of negative samples, which they are negatives (i.e., true negatives). The two other cells (lower left and lower right) represent the misclassified number of samples. In particular: the lower left-hand cell indicates the number of samples classified as positives when they are negatives (i.e., false positive), while the upper-right cell indicates the number of samples classified as negatives when they are positives (i.e. false negative). Once the confusion matrix is constructed, the accuracy of the fold is calculated using the respective formula presented in equation 5.2. Table 2 shows the results of the confusion matrix from the test set. Also, from the confusion matrix table, we can find the F_1, *PPV*, *NPV*, and *Recall*. The results are shown in Table 3. It is observed that a high positive predictive value (PPV = 87%) which means that most of the positive results from this control process are valid. On the other hand, we observe an average percentage for the negative prognostic value (NPV = 55.3%) which means that only 55.3% of the negative results from this control process are authentic. In addition, the effectiveness of the positives becomes even more apparent if we observe both other metrics (Recall =92% and F-score=89.4).

Table 9: Confusion table and evaluation metrics for the classification model.

	Actual	
predicted	Positive	Negative
Positive	242=TP	35=FP
Negative	21=FN	26=TN

Table 10: The evaluation results using the Naive Bayes method.

Polarity	*Precision*	*Recall*	*F-score*	*Accuracy*
Positive	87%	92%	89,4	82,7%
Negative	55,3%	42,6%	49%	

Chapter 6

6. Conclusions and Future Research

Web-based marketing and communication are influenced by the internet, a vast virtual space where people may express and share their ideas with others. Online reviews are becoming increasingly influential in evaluating products and services by potential customers. When it comes to product or service reviews, consumers are more likely to trust the opinions of other customers, particularly those who have already used the product or service, than they are to believe the marketing messaging of the corporation. Social media's influence on people's views and behavior has an impact on their purchasing decisions. People's shopping habits are becoming increasingly affected by the internet, mainly through social networking sites like Twitter.

A massive amount of data is generated every day, most of which contains valuable information that has yet to be discovered and analyzed. To read and learn from this information, Artificial Intelligence can be utilized in conjunction with Emotion Analysis as a scientific field preference of Artificial Intelligence can be used to evaluate, operate, and finally extract this information conclusion that will be of great importance to everyone interested.

This paper was prepared to study the public's attitude towards i-phone 13, a much-discussed product from the most successful multinational technology company. The aim was to explore the possibility of recognizing the trends of the public from his posts on the social network Twitter.

For our research, we collected an API dataset with tweets listed on i-phone 13; the dataset contained 1303 tweets and was collected over a period of almost 3 months, the collection was done almost daily, but many of them had to be deleted (Chapter 4). After manually marking the data as "Positive", "Negative" and "Neutral" (63.4% positive, 17% neutral and 19.6% negative), we entered it in the R program for pre-processing. The first step was the pre-processing of the data, their cleaning, and the removal of data that we did not need, which would hinder our research (Chapter 4). The next step was to look at the data more specifically, visualizing our data and finding the most common words that refer to it,

diagram and word cloud. We removed the "Neutral" tags to facilitate the algorithm to continue our research. We divided our data into two random samples for training and test set. We used the training set to train the Naive Bayes classifier and the test set to test his predictions. Finally, we applied cross-validation ten times to predict the success of our model. The results show that our model is well constructed and achieves relatively high accuracy with an average accuracy of 82.7%.

After the end of our research, we come to the following conclusions. Initially, the relatively high average accuracy rate justifies our choice to use the Naive Bayes classifier. We see that it successfully serves its purpose and effectively addresses the problem we solve. In addition, the success rate of the proposed model in the present study seems to confirm the relevant literature in which many researchers have proposed the Naive Bayes classifier. Finally, the results show that the polarity of tweets remains a big challenge, as the search for people's opinions is different on each site and for each topic and period, as there is a difference in vocabulary and features that you need to focus on each time. Thus, a powerful polarity technique must take into account the aspects and characteristics of the data and the model it chooses.

6.1. Future Work

There is no doubt that in the field of sentiment analysis still has many opportunities and challenges that can face becoming even more efficient. Several of these may include the following:

Initially, the data is categorized using various people for labeling. Using a single person to determine the polarity of a tweet does not guarantee the quality of the entire training set. Generally, people who make the labeling are likely to categorize tweets under the influence of their personal beliefs, style and character, educational and social status, and ability to follow the same evaluation rules whenever necessary strictly. More independent professionals would significantly improve the overall training, especially if they ultimately agreed with their choices. In addition, a team helps to label and get more feedback as the process gets faster. So maybe we had a better picture and a better result.

Another challenge is combining more Emotion Analysis models. Supervised machine learning requires large sets of manually tagged training that need time and effort and conceal subjective and objective errors. On the other hand, using a dictionary has finite

dictionaries that cannot cover the whole of each study object. Also, a word can give a different meaning to two sentences depending on the context and the theme or even its place in the sentence. Combining the advantages of each model can result in a significant improvement in classifier accuracy and performance.

Finally, another thought for continuation is comparing similar technology products from different companies and different countries to know what they prefer in each market.

Chapter 7

References

A., V. and Sonawane, S. S. (2016) ‘Sentiment Analysis of Twitter Data: A Survey of Techniques’, *International Journal of Computer Applications*, 139(11), pp. 5–15. doi: 10.5120/ijca2016908625.

Aggarwal, C. C. (2015) *Data Classification Algorithms and Applications*. New York: CRC Press Taylor & Francis Group.

Aggarwal, C. C. and Xhai, C. (2012) *Mining Text Data*. Edited by C. Z. Charu C. Aggarwal. New York. doi: 10.1007/978-1-4614-3223-4.

Andrea Esuli, F. S. (2005) ‘Determining the Semantic Orientation of Terms through Gloss Classification’, (July), pp. 617–624.

Aue, A. and Gamon, M. (2005) ‘Customizing Sentiment Classifiers to New Domains : a Case Study’.

B Pang, L. lee (2008) ‘Opinion mining and sentiment analysis’, 22, pp. 1–135. doi: 10.3748/wjg.v22.i45.9898.

Balage Filho, P. P. and Pardo, T. A. S. (2013) ‘NILC USP: A hybrid system for sentiment analysis in twitter messages’, **SEM 2013 - 2nd Joint Conference on Lexical and Computational Semantics*, 2(SemEval), pp. 568–572.

Bayhaqy, A. *et al.* (2018) ‘Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes’, *2018 International Conference on Orange Technologies, ICOT 2018*. IEEE, pp. 1–6. doi: 10.1109/ICOT.2018.8705796.

Bhuta, S. *et al.* (2014) ‘A review of techniques for sentiment analysis of Twitter data’, *Proceedings of the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT 2014*, pp. 583–591. doi: 10.1109/ICICICT.2014.6781346.

Bishop, C. M. (1995) *Neural Network for Pattern Recognition, CLARENDON PRESS • OXFORD*. doi: 10.1109/RusAutoCon49822.2020.9208207.

Bo, Pang. Lillian, Lee Shivakumar, V. (2002) ‘Thumbs up? Sentiment Classification using Machine Learning Techniques’, *American Journal of Orthodontics and Oral Surgery*, 31(9), pp. 481–482. doi: 10.1016/0096-6347(45)90048-2.

Brett, L. (2015) *Machine Learning with R, Machine Learning*. doi: 10.1002/9781119642183.ch14.

Clement, J. (2021) *Total global visitor traffic to Twitter . com 2021 Worldwide visits to Twitter . com from January to June 2021*, *Statista*. Available at: https://www.statista.com/statistics/470038/twitter-audience-reach-visitors/.

Dave, K., Lawrence, S. and Pennock, D. M. (2003) ‘Mining the peanut gallery: Opinion extraction and semantic classification of product reviews’, *Proceedings of the 12th*

International Conference on World Wide Web, WWW 2003, pp. 519–528. doi: 10.1145/775152.775226.

Delen, D. (2009) 'Analysis of cancer data: a data mining approach', *Expert Systems, The journal of Knowledge Engineering*, 26(1). doi: 10.1111/j.1468-0394.2008.00480.x.

Devika, M. D., Sunitha, C. and Ganesh, A. (2016) 'Sentiment Analysis: A Comparative Study on Different Approaches', *Procedia Computer Science*. The Author(s), 87, pp. 44–49. doi: 10.1016/j.procs.2016.05.124.

Gamallo, P. and Garcia, M. (2015) 'Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets', (SemEval), pp. 171–175. doi: 10.3115/v1/s14-2026.

Gamon, Michael, Anthony Aue, Simon Corston-Oliver, E. R. (2005) 'Pulse: Mining Customer Opinions from Free Text', *Advances in Intelligent Data Analysis VI*, 3646(May 2014), pp. 121–132. doi: 10.1007/11552253_12.

Ghiassi, M., Skinner, J. and Zimbra, D. (2013) 'Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network', *Expert Systems with Applications*. Elsevier Ltd, 40(16), pp. 6266–6282. doi: 10.1016/j.eswa.2013.05.057.

Ghose, A., Ipeirotis, P. G. and Sundararajan, A. (2007) 'Opinion mining using econometrics: A case study on reputation systems', *ACL 2007 - Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, (June), pp. 416–423.

Giachanou, A. and Crestani, F. (2016) 'Like it or not: A survey of Twitter sentiment analysis methods', *ACM Computing Surveys*, 49(2). doi: 10.1145/2938640.

Goel, A., Gautam, J. and Kumar, S. (2017) 'Real time sentiment analysis of tweets using Naive Bayes', *Proceedings on 2016 2nd International Conference on Next Generation Computing Technologies, NGCT 2016*, (October), pp. 257–261. doi: 10.1109/NGCT.2016.7877424.

Grosan, C. and Abraham, A. (1997) *Machine Learning, Intelligent Systems Reference Library*. doi: 10.1007/978-3-642-21004-4_10.

Hatzivassiloglou, V. and McKeown, K. R. (1997) 'Predicting the semantic orientation of adjectives', pp. 174–181. doi: 10.3115/979617.979640.

Helal, G., Ozuem, W. and Lancaster, G. (2018) 'Social media brand perceptions of millennials', *International Journal of Retail and Distribution Management*, 46(10), pp. 977–998. doi: 10.1108/IJRDM-03-2018-0066.

Hota, S. and Pathak, S. (2018) 'KNN classifier based approach for multi-class sentiment analysis of twitter data', *International Journal of Engineering & Technology*, 7(3), pp. 1372–1375. doi: 10.14419/ijet.v7i3.12656.

Hu, M. and Liu, B. (2004) 'Mining and summarizing customer reviews', *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177. doi: 10.1145/1014052.1014073.

Hussein, D. M. E. D. M. (2018) 'A survey on sentiment analysis challenges', *Journal of King Saud University - Engineering Sciences*. King Saud University, 30(4), pp. 330–338. doi: 10.1016/j.jksues.2016.04.002.

Internet users in the world 2021 _ Statista (2020) *Statista*. Available at: https://www.statista.com/statistics/617136/digital-population-worldwide/.

Joshi, N. S. and Itkat, S. A. (2014) ‘A Survey on Feature Level Sentiment Analysis’, 5(4), pp. 5422–5425.

Kamps, J. *et al.* (2004) ‘Using WordNet to measure semantic orientations of adjectives’, *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, pp. 1115–1118.

Kantardzic, M. (2011) *Data mining: concepts, models, methods, and algorithms.* Kentucky,USA: IEEE PRESS A JOHN WILEY & SONS, INC., PUBLICATION.

Kaplan, A. M. and Haenlein, M. (2011) ‘The early bird catches the news: Nine things you should know about micro-blogging’, *Business Horizons*, 54(2), pp. 105–113. doi: 10.1016/j.bushor.2010.09.004.

Kearney, M. M. W. (2020) ‘Package “ rtweet ”’.

Khairnar, J. and Kinikar, M. (2013) ‘Machine Learning Algorithms for Opinion Mining and Sentiment Classification’, *International Journal of Scientific and Research Publications*, 3(6), pp. 1–6. Available at: www.ijsrp.org.

Li, G. and Liu, F. (2010) ‘A clustering-based approach on sentiment analysis’, *Proceedings of 2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2010*, pp. 331–337. doi: 10.1109/ISKE.2010.5680859.

Liu, B. (2012) *Sentiment Analysis and Opinion Mining*. doi: 10.2200/S00416ED1V01Y201204HLT016.

Liu, B. (2015) *Sentiment analysis: Mining Opinions, Sentiments, and Eemotions*. Edited by U.-C. University of Illinois. Cambridge University Press. doi: https://doi.org/10.1017/CBO9781139084789.

Liu, B. and Hu, M. (2005) ‘Opinion Observer: Analyzing and Comparing Opinions on the Web’. Available at: http://arxiv.org/abs/2010.07523.

Lohmann, S. *et al.* (2012) ‘Visual analysis of microblog content using time-varying co-occurrence highlighting in tag clouds’, *Proceedings of the Workshop on Advanced Visual Interfaces AVI*, (January 2014), pp. 753–756. doi: 10.1145/2254556.2254701.

Mankar, T. *et al.* (2018) ‘Stock Market Prediction based on Social Sentiments using Machine Learning’, *2018 International Conference on Smart City and Emerging Technology, ICSCET 2018*. IEEE, pp. 2–4. doi: 10.1109/ICSCET.2018.8537242.

Maria Arista Ulfa, Budi Irmawati, and A. Y. H. (2018) ‘Twitter Sentiment Analysis using Naıve Bayes Classifier with Mutual Information Feature Selection’, *Journal of Computer Science and Informatics Engineering (J-Cosine)*, 2(2), pp. 106–111.

Mccallum, A. and Nigam, K. (1998) ‘A Comparison of Event Models for Naive Bayes Text Classification’, *AAAI Workshop on Learning for Text Categorization, 1998.*

Melville, P., Ox, O. and Lawrence, R. D. (2009) ‘Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification’, pp. 1275–1283.

Merriam-Webster (2005). Available at: https://www.merriam-webster.com/dictionary/opinion#synonyms.

Nasukawa, T. and Yi, J. (2003) ‘Sentiment analysis: Capturing favorability using natural language processing’, *Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP 2003*, (January 2003), pp. 70–77. doi: 10.1145/945645.945658.

Neri, F. *et al.* (2012) 'Sentiment analysis on social media', *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, pp. 919–926. doi: 10.1109/ASONAM.2012.164.

Di Nunzio, G. M. (2009) 'Using scatterplots to understand and improve probabilistic models for text categorization and retrieval', *International Journal of Approximate Reasoning*. Elsevier Inc., 50(7), pp. 945–956. doi: 10.1016/j.ijar.2009.01.002.

Pak, A. and Paroubek, P. (2010) 'Twitter as a corpus for sentiment analysis and opinion mining', *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, pp. 1320–1326. doi: 10.17148/ijarcce.2016.51274.

Pang, B. *et al.* (2002) 'Thumbs up ? Sentiment Classification using Machine Learning Techniques', (July), pp. 79–86.

Payam Refaeilzadeh, Lei Tang, H. L. (2014) 'Cross-Validation', *Encyclopedia of Database Systems*, pp. 141–148. doi: 10.1201/b17767-13.

Pedro J., Garcı´a-Laencina a, Jose -Luis, Sancho-Gomez, Anıbal R., F.-V. and Verleysen, M. (2009) 'Neurocomputing K nearest neighbours with mutual information for simultaneous classification and missing data imputation', *Neurocomputing*, 72, pp. 1483–1493. doi: 10.1016/j.neucom.2008.11.026.

Peter D. Turney (2002) 'Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews', *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424. Thumbs*, 25(4), p. 52. Available at: http://www.google.com.

Prasad, S. S. *et al.* (2015) 'Sentiment classification: An approach for Indian language tweets using decision tree', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9468, pp. 656–663. doi: 10.1007/978-3-319-26832-3_62.

Quinlan, J. R. (1986) 'Induction of Decision Trees', *6 Kluwer Academic Publishers, Boston - Manufactured in The Netherlands Induction*, pp. 81–106.

Rahman Khan, M. A., Siddique, M. R. and Basir, M. S. (2021) 'A Comparative Analysis of Machine Learning Classifiers for Medical Datasets', *SSRN Electronic Journal*, 110, pp. 71–83. doi: 10.2139/ssrn.3814755.

Rokach, L. and Maimon, O. (2005) 'Top-down induction of decision trees classifiers - A survey', *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 35(4), pp. 476–487. doi: 10.1109/TSMCC.2004.843247.

Ruz, G. A., Henríquez, P. A. and Mascareño, A. (2020) 'Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers', *Future Generation Computer Systems*. Elsevier B.V., 106, pp. 92–104. doi: 10.1016/j.future.2020.01.005.

S, Bird, E, Klein, E, L. (2009) *Natural language processing with python, O'Reilly Media*. doi: 10.17509/ijal.v1i1.106.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010) 'Earthquake shakes Twitter users: Real-time event detection by social sensors', *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, (April), pp. 851–860. doi: 10.1145/1772690.1772777.

Sayali P. Nazare, Prasad S. Nar, Akshay S. Phate, P. D. D. R. I. (2018) 'Sentiment Analysis in Twitter', *International Research Journal of Engineering and Technology(IRJET) e-ISSN:*

2395-0056, 05(1), pp. 880–886. Available at: http://dx.doi.org/10.1053/j.gastro.2014.05.023%0Ahttps://doi.org/10.1016/j.gie.2018.04.013%0Ahttp://www.ncbi.nlm.nih.gov/pubmed/29451164%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5838726%250Ahttp://dx.doi.org/10.1016/j.gie.2013.07.022.

Simon, P. W. and (2013) 'Too Big to Ignore : The Business Case for Big Data', *Journal of Chemical Information and Modeling*, 53(9), pp. 1689–1699. doi: 10.1017/CBO9781107415324.004.

Statista Research Department (2021) *Twitter : annual revenue 2010-2020 , by segment Annual revenue of Twitter from 2010 to 2020 , by segment*. Available at: https://www.statista.com/statistics/274566/twitters-annual-revenue-by-channel/.

Steinberger, J., Brychcín, T. and Konkol, M. (2015) 'Aspect-Level Sentiment Analysis in Czech', pp. 24–30. doi: 10.3115/v1/w14-2605.

Table, E. (2020) 'Confusion matrix', pp. 9–12.

Taboada, M., Brooke, J. and Voll, K. (2011) 'Lexicon-Based Methods for Sentiment Analysis', *Computational Linguistics*, (December 2009). doi: 10.1162/COLI_a_00049.

Twitter, U., Guide, G. S. and Us, C. (no date) *About Twitter ' s APIs*. Available at: https://help.twitter.com/en/rules-and-policies/twitter-api.

Updating, B. (2020) 'Positive and negative predictive values', pp. 1–5.

Waila, P. *et al.* (2012) 'Evaluating Machine Learning and Unsupervised Semantic Orientation approaches for sentiment analysis of textual reviews', *2012 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2012*, (c). doi: 10.1109/ICCIC.2012.6510235.

Wang, A. H. (2010) 'Don't follow me - Spam detection in twitter', *SECRYPT 2010 - Proceedings of the International Conference on Security and Cryptography*, 2010, pp. 142–151.

William L. Hosch (2021) *Machine Learning*, *Encyclopedia Britannica*. A Britannica Publishing Partner. Available at: https://www.britannica.com/technology/machine-learning.

Workshop, I. R., Aug, D. and Fraley, R. C. (2016) *Using R to Mine Data from Twitter: An Introduction for Psychological Scientists*. Available at: https://www.yourpersonality.net/R/R_Notes_4.html.

Žižka, J., Dařena, F. and Svoboda, A. (2019) *Text Mining with Machine Learning*, *Text Mining with Machine Learning*. doi: 10.1201/9780429469275.