Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Information Technology

Bachelor Thesis

Sentiments Analysis on Twitter Data

Priyanshu Sharma

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

BACHELOR THESIS ASSIGNMENT

Bc. Priyanshu Sharma

Informatics

Thesis title

Sentiment Analysis on Twitter Data

Objectives of thesis

The main objective of the thesis is to design and implement a sentiment analyzer for twitter dataset with data mining approach towards classifying data (tweets) as positive, negative, or neutral with respect to a query term.

Partial objectives:
- data collection using Twitter APIs,
- pre-processing of collected data (classification and parallel processing using for example multinomial Naïve Bayes Classifier or Support Vector Machines),
- sentiment scoring module,
- design and perform experimental measurements of sentiment scores for different query terms.

Methodology

The theoretical part of the thesis is based on the study and analysis of professional and scientific information sources.

Following steps will be taken to conduct the sentiment analysis. Firstly, data will be collected using libraries in python, text processing, testing training data, and text classification using the Naïve Bayes or other appropriate machine learning method. The Naïve Bayes method can be used to help classify classes or the level of sentiments of society. After training, the model will be tested against the custom query parameter. Results of sentiment analysis on twitter data will be displayed as different sections presenting positive, negative, and neutral sentiments.

Based on the synthesis of theoretical knowledge and the results of the practical part, the conclusions of the work will be formulated.

The proposed extent of the thesis

40 – 50 pages

Keywords

sentiment analysis; sentiment classification; polarity detection; machine learning; social network; twitter

Recommended information sources

A. P. Jain and V. D. Katkar, "Sentiments analysis of Twitter data using data mining," 2015 International Conference on Information Processing (ICIP), 2015, pp. 807-810, doi: 10.1109/INFOP.2015.7489492.

A. Shelar and C. -Y. Huang, "Sentiment Analysis of Twitter Data," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 2018, pp. 1301-1302, doi: 10.1109/CSCI46756.2018.00252.

H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016, pp. 416-419, doi: 10.1109/ICATCCT.2016.7912034.

M. Venugopalan and D. Gupta, "Exploring sentiment analysis on twitter data," 2015 Eighth International Conference on Contemporary Computing (IC3), 2015, pp. 241-247, doi: 10.1109/IC3.2015.7346686.

M. Wongkar and A. Angdresey, "Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter," 2019 Fourth International Conference on Informatics and Computing (ICIC), 2019, pp. 1-5, doi: 10.1109/ICIC47613.2019.8985884.

Expected date of thesis defence

2022/23 WS – FEM

The Bachelor Thesis Supervisor

Ing. Jan Masner, Ph.D.

Supervising department

Department of Information Technologies

Electronic approval: 27. 7. 2021	Electronic approval: 5. 10. 2021
doc. Ing. Jiří Vaněk, Ph.D.	**Ing. Martin Pelikán, Ph.D.**
Head of department	Dean

Prague on 07. 03. 2023

Declaration

I declare that I have worked on my bachelor thesis titled "**Sentiments Analysis on Twitter Data**" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break any copyrights.

In Prague on 28 November 2023

Acknowledgement

I would like to express my deep gratitude and appreciation to the following individuals and organizations who have supported me to the completion of this thesis:

First and foremost, I would like to thank my thesis advisor, **Ing. Jan Masner**, for his guidance, encouragement, and support throughout the thesis process. His insights, feedback, and patience were invaluable and greatly appreciated.

I am indebted to Czech University of Life Sciences for providing me with the necessary resources and facilities to carry out my research. Without their support, this thesis would not have been possible.

Finally, I would like to express my heartfelt gratitude to my Parents, for their unwavering love, support, and encouragement. Their sacrifices and belief in me have been the driving force behind my academic success.

Thank you all for your invaluable contributions to this thesis.

Sentiments Analysis on Twitter Data

Abstract

In this thesis we proposed a model for sentiment analysis on twitter data. We conducted sentiment analysis on tweets in order to give some business intelligence predictions. We applied our algorithm to different datasets consisting of over 1 million tweets form different peoples of different regions. This dataset was particularly challenging due to its greater variety of sentiment and complex sentence structures. The first stage is to use Python language code to extract tweets on a certain pattern, followed by the clean-up process, and finally the construction of a bag of phrases. These sacks of words are afterwards sent into the algorithms as input. Finally, when the algorithms have been trained, we shall get public opinion on the proposal. We were pleased to find that our model still managed to achieve an accuracy of over 95% in classifying sentiments as positive or negative in nature, proving that our algorithm is versatile and capable of handling different types of data. Moving forward, we plan to further refine our algorithm and improve its performance even more. To further improve the performance of our sentiment analysis algorithm, we also tested it on Twitter data of larger datasets. The results were quite promising, with our model achieving an accuracy of over 97% on the Twitter dataset. We believe that this demonstrates the potential of our algorithm for sentiment analysis on social media data and can be used for a wide range of applications.

Keywords: sentiment analysis; sentiment classification; polarity detection; machine learning; social network; twitter

Analiza pocitů na Twitteru

Abstrakt

V tomto článku jsme navrhli model založený na pro analýzu sentimentu na twitterových datech. U tweetů jsme provedli analýzu sentimentu, abychom poskytli nějaké předpovědi business intelligence. Náš algoritmus jsme aplikovali na různé datové sady skládající se z více než 1 milionu tweetů od různých národů různých regionů. Tento soubor dat byl obzvláště náročný kvůli jeho větší rozmanitosti sentimentu a složitým větným strukturám. První fází je použití kódu jazyka Python k extrahování tweetů na určitém vzoru, následuje proces čištění a nakonec sestavení pytle frází. Tyto pytle slov jsou poté odeslány do algoritmů jako vstup. Nakonec, až budou algoritmy natrénovány, získáme veřejné mínění o návrhu. S potěšením jsme zjistili, že náš model stále dokázal dosáhnout přesnosti více než 95 % při klasifikaci sentimentů jako pozitivní nebo negativní povahy, což dokazuje, že náš algoritmus je všestranný a dokáže zpracovávat různé typy dat. Do budoucna plánujeme náš algoritmus dále vylepšovat a ještě více zlepšovat jeho výkon. Abychom dále zlepšili výkon našeho algoritmu analýzy sentimentu, testovali jsme jej také na datech větších datových sad na Twitteru. Výsledky byly docela slibné, přičemž náš model dosáhl na datovém souboru Twitteru přesnosti přes 97 %. Věříme, že to demonstruje potenciál našeho algoritmu pro analýzu sentimentu na datech sociálních médií a může být použit pro širokou škálu aplikací.

Klíčová slova: analýza sentimentu; klasifikace sentimentu; detekce polarity; strojové učení; sociální síť; cvrlikání

Table of Contents

1 Introduction

Genuine text component processing is the focal point of natural language processing (NLP). NLP changes over the message component into a machine-meaningful portrayal. Artificial intelligence (AI) examinations information from natural language processing and performs broad numerical tasks to assess assuming something is positive or Negative. There are multiple ways of removing a creator's point of view regarding a matter from natural language text-based information. An AI procedure or some likeness thereof is utilized, with various levels of progress. Assessment mining is a kind of natural language processing that spotlights on measuring public feeling toward a particular item or subject. This product tracks mentalities and sentiments on the web as well as naturally extricates thoughts, sentiments, and feelings from message. People discuss a wide range of issues in blog posts, comments, reviews, and tweets to share their opinions. Opinion mining has a variety of names and slightly varied duties such as affect analysis, subjectivity analysis, sentiment mining, opinion extraction, analysis of emotions, review mining, etc. Combined with artificial intelligence, computer science, and linguistics, natural language processing. The field of computational linguistics focuses on how computers and people communicate (natural) languages. NLP is related with the field of human-PC communication accordingly. Numerous challenges in NLP includes natural language processing, which empowers PCs to decipher human or others entail the formation of natural language without any preparation.

Sentiment analysis (SA) illuminates' clients whether the material is positive, negative, or nonpartisan in about the item's quality prior to causing a buy Advertisers and organizations to use this analysis information to find out about their items or administrations in a way that permits them to be presented according to the client's details. Text based Data Recovery Strategies Focus Fundamentally On processing, looking, and deciphering the present verifiable realities. There is an objective part to realities, however there are extra text-based data that communicates emotional characteristics. These items are fundamentally conclusions, considerations, evaluations, perspectives, as well as sentiments, which are the underpinning of Sentiment Analysis. It gives various hard potential outcomes to make new applications, principally because of the gigantic extension of available data accumulated from online sources like websites and informal organizations.

2 Objectives and Methodology

Sentiment analysis is a powerful technique in natural language processing that involves extracting subjective information from text data to determine the sentiment or emotional tone expressed. The primary goal is to understand whether a piece of text, such as a tweet, review, or comment, conveys a positive, negative, or neutral sentiment. This analysis is applied on twitter with millions of tweets, using different tags, keywords, topics, and trends.

2.1 Objectives

The main objective of the thesis is to design and implement a sentiment analyser for twitter dataset with data mining approach towards classifying data (tweets) as positive, negative, or neutral with respect to a query term.

Partial Objectives -:

- Data collection using Twitter APIs.
- Pre-processing of collected data (classification and parallel processing using for example multinomial Naïve Bayes Classifier or Support Vector Machines)
- Sentiment scoring module
- Design and perform experimental measurements of sentiment scores for different query terms.

2.2 Methodology

The theoretical part of the thesis is based on the study and analysis of professional and scientific information sources. Following steps will be taken to conduct the sentiment analysis.

Firstly, data will be collected using libraries in python, text processing, testing training data, and text classification using the Naïve Bayes or other appropriate machine learning method. The Naïve Bayes method can be used to help classify classes or the level of sentiments of society. After training, the model will be tested against the custom query parameter. Results of sentiment analysis on twitter data will be displayed as different sections presenting positive, negative, and neutral sentiments.

3 Literature Review

Modern times has adjusted how individuals convey their perspectives and suppositions today. It is presently directed essentially by means of blog postings, web discussions, item survey sites, and online entertainment, and so on. As of now, a huge number of people utilize web-based entertainment to utilize long range informal communication destinations like Facebook, Twitter, Google+, and so on express their sentiments, feelings, and offer their point of view on daily schedules hrough web-based networks, we gain admittance to Intuitive media in which clients can illuminate and impact others utilizing gatherings [28]. Virtual entertainments are creating a lot of opinion rich information got from tweets, notices, and blog entries postings, remarks, surveys, and so on. Besides, web-based entertainment gives organizations an open door by giving a stage to speak with their clients for the end goal of publicizing. If someone wishes to purchase a product or utilize a service, they first check up its before purchasing a product, consult online reviews and social media the volume of content created by people is excessive for a typical user to examine. This amount of data that includes reviews, survey, opinions is almost impossible to understand from a normal human without categorising it into meaningful categories. Thus, it is necessary to automate consequently, numerous sentiment analysis approaches are extensively employed. This not only help business to gain the valuable insight of custom feedback, but this data can be used in numerous good and bad ways, including providing good customer service or manipulating the thoughts and behaviours of individuals.

3.1 Assess Reaction on Twitter

Paying attention to clients is fundamental for finding ways of upgrading an item or administration. Although there are different wellsprings of criticism, as overviews and public surveys, Twitter gives unedited, genuine contribution on your crowd's thought process of your administration. By inspecting how individuals examine your image on Twitter, you can decide whether they partake in a recently reported highlight. You can likewise decide if your cost is straightforward to your objective market. To make business decisions, you can likewise figure out what parts of your administration are the most adored and detested, For example clients cherishing the straightforwardness of the User Interface of the product, yet disdain how slow client service is.

3.2 Applications of Sentiment Analysis

Sentiment analysis offers insightful information about customer preferences, opinions, and feedback that can be used by businesses to make smart decisions about product development, marketing, and customer support. There are potentially innumerable use cases of sentiment analysis, but here are some of the real-life use cases in which Sentiment analysis can be applied.

Voice of the Customer Application

Voice of Customer (VoC) Programs are the input obtained to better comprehend the feelings and worries of a brand's customers. This is fundamental for improving the client experience. Sentiment analysis can help in categorizing and organizing the data to help brands to detect latest trends and complaints. This data will aid in risk avoidance and the advancement of procedures for settling worries with the organization's items or administrations. Assessing how clients feel and their thought process can help you distinguish and deal with arising concerns.[29]

Customer Feedback Analysis

An exceptional client service experience may make or break a business. Estimation analysis and content analysis can both be linked to client feedback discussions. By using sentiment analysis, companies can get reviews of their products from users. Some examples can be if the customer understood the product clearly or not or there are any questions about any component of the product. If a user is not satisfied with the product, their problem ticket can be prioritised over others. By connecting questions with the proper individuals, assumption examination can save handling times and boost productivity.[30]

Analysis of Social Media Sentiment

Social media can monitor client behaviour seven days a week, throughout the clock. It can not only show real time data if there is some issue going on with the product (e.g., Twitter Hashtags) but can also help the brand to track the customers to provide the appropriate solution. Not to forgot, it also can be one of the most effective strategies to get new customers and retain existing ones. Customers are encouraged to buy from a firm by positive evaluations and social media posts. On the other side, negative reviews and comments can be among the most damaging forms of advertising.[29]

Product Experience

The use of a sentiment analysis could uncover how your clients feel about the highlights and advantages of your items or administrations. This might offer help and shed light on already obscure areas of chance. A good example of this approach can be seen in Google Chrome's popularity. During initial launch days, Chrome's development team was consistently monitoring the users' opinions, irrespective of criticism or praise, they figured out the needs of the users regarding UI, Optimization, security, and various areas. This approach not only helped Google to Launch a more stable product, but they crushed the market. Google still follows this approach and the results of this can be seen in the market share of Chrome. [30]

Brand Reputation Analysis

For a positive customer experience, brand sentiment is one of the most crucial variables to consider. Evaluating customers' views toward their brand, product or service is very crucial for industries like Fashion, Marketing, Food, and other businesses. Depending on brand perception, sales could increase or decline. This is also apparent in brand loyalty, as positive attitudes result in favourable reviews and referrals and negative sentiments raise consumer churn rates. Sentiment analysis offers brands with the means to track their customers' attitudes. To be aware of your brand's reputation, it is prudent to examine communities on forums and social media platforms. Companies must also monitor brand, product, and competitor references to comprehend the brand's overall image [29]. This allows businesses to evaluate the impact of a public relations campaign or new product introduction on the overall brand perception.

Stock Market Prediction

One of the advantages of Sentiment Analysis is predicting the stock price. It is usually done by gathering and analysing the data (about a brand) from various sources such as Twitter, Facebook, news articles, blogs, etc. Sentence level analysis can be performed on the collected data (news) and then overall polarity of the text about the company can tell us about how well the stock will perform in future.[29]

However, in the work of Kraaijeveld and De Smedt (2020), it shows that businesses which applied sentiment analysis in areas of Blockchain/Cryptocurrency, the results were infrequent and predicted very inconsistent results in different tests.

Market Research, Competitor Analysis

Sentiment analysis can aid businesses in identifying new trends, analysing competitors, and testing prospective markets. Companies may need to evaluate the review scores of their competitors. Using sentiment analysis to examine this data can assist in determining how clients feel about competition.

AI, especially factual AI, is the underpinning of contemporarily calculations. The AI worldview is unmistakable from that of most of prior endeavours at language processing. Before, language processing exercises were frequently carried out by straightforwardly hand-coding colossal arrangements of rules. The AI worldview prescribes rather utilizing general learning calculations to naturally learn such standards through the assessment of tremendous corpora of normal genuine examples. These calculations are oftentimes, however not generally, grounded in measurable deduction. An assortment of reports (or sometimes, individual sentences) that have been physically labelled with the important qualities to be learned is alluded to as a corpus (plural, "corpora"). AI techniques from various classes have been utilized for NLP issues.

Measurable surmising procedures can be utilized via programmed growing experiences to make models that are strong to novel information, (example, words or designs that have never been seen) and to misleading info (for example with incorrectly spelled words or words incidentally discarded). By and large, it is truly challenging, blunder inclined, and tedious to deal with such information effortlessly utilizing transcribed rules, or more for the most part, to make frameworks of manually written decides that pursue delicate choices. By just adding more information, frameworks that naturally get familiar with the standards can be made more precise. Nonetheless, written by hand rule-based frameworks must be made more precise by making the guidelines more confounded, which is an impressively really testing undertaking. Specifically, there is an intricacy edge past which hand-made rule-based frameworks become progressively unmanageable.

Sites for microblogging have formed into a wellspring of a wide range of sorts of data. This is a consequence of the idea of microblogs, where clients distribute messages continuously in regard to a scope of points, recent developments, complaints, and positive sentiment for items they use consistently. As a matter of fact, organizations that cause these things to have started to overview these microblogs to check general assessment of their merchandise. These organizations often screen client criticism and answer individuals on microblogs.

3.3 Sentiment Analysis

Sentiment analysis is classified as a branch of machine learning and natural language processing. It classifies positive, neutral, and negative views extracted, recognized, or depicted from various content formats, such as media, reviews, and publications [24]. Twitter is one of the most popular social media platforms for expressing one's opinion on a certain issue. Twitter allows tweeples (Twitter users) to express their thoughts on politics, entertainment, industry, the stock market, and other topics. Twitter has a large quantity of useful data, which allows numerous researchers to dive through tweets and create predictions about events, products, industries, financial markets, and other topics using various categorization algorithms. The vast amount of data available has piqued the interest of numerous researchers interested in studying Twitter data quantitatively in a broad or in a more particular sense scientifically and meaningfully [25].

Text mining frequently employs the technique of sentiment analysis. By analysing the sentiment of the text (in this case, a tweet), using sophisticated text mining techniques, it is possible to determine if it is positive, negative, or neutral. It is often referred to as "opinion mining" and is largely used to analyse discussions, opinions, and viewpoints (all expressed in the form of tweets) in order to determine commercial strategy, conduct political analysis, and evaluate societal behaviour. Some of the well-known technologies used for sentiment analysis on twitter include ingenuity, revealed context, meaning cloud, and social mention. Twitter sentiment analysis dataset is frequently used in R and Python. Sentiment analysis is the most common way of finding and ordering the feelings addressed in a message source. When dissected, tweets can deliver a lot of sentiment information. These insights assist us with understanding how people feel about a scope of issues.

Social Systems

In today's society, societal evolution is accelerating Networks have become an integral and invaluable component of human life. As technology advances, as anticipation grows, a growing number of new types of social websites are emerging with more sophisticated, freshly added features. There are numerous social networking sites like Facebook Twitter, WhatsApp, WeChat, Instagram, Pinterest, Snapchat, LinkedIn, and Q Zone are examples of social media platforms. These are the social networks embraced by commercial companies that utilize this communication network to reach consumers in order to promote their products.

Social Network Analysis

A test based on principles would reveal to you all entries containing a specific word (for example, "Tide" AND "cleanser" BUT "sea"). While this may function to show you posts regarding your requirement, it does not permit you to reveal much in the way of business-related information. Social Network Analysis (SNA) is a demonstrative technique for revealing the links within an informal group in order to better accurately predict the relationships. It employs a variety of techniques and devices to provide a visual representation of interpersonal relationships and functions as a "authoritative X-ray" into the informal operations of a group of people. Rather of focusing on the characteristics of the individuals (such as a skills profile or socioeconomic status), the emphasis is on the relationships between them.[32]

In this way, to ascertain the shopper viewpoint, we want to make a computerized AI sentiment analysis model. It is trying to apply models to them since they contain both important information and non-valuable attributes (alluded to aggregately as commotion).

Twitter

With 330 million months to month dynamic clients and 500 million tweets posted every day, twitter is one of the most generally utilized person to person communication locales on the planet. We might find a ton about how individuals feel about specific subjects via cautiously looking at the sentiment of these tweets, whether it be ideal, negative, or impartial, for instance. For various purposes, including corporate showcasing, legislative issues, the exploration of public way of behaving, and data gathering, fathoming the feeling of tweets is significant. Marketing professionals can learn through sentiment analysis of twitter data how consumers react to new products and marketing initiatives, and political organizations can use it to see how the public reacts to announcements or changes in legislation. However, analysing twitter data is not an easy operation. Approximately 6000 tweets are published per second. That's a ton of twitter information! Human sentiment analysis is also not scalable, even though it is simple for humans to determine a tweet's emotional tone.

3.4 Understanding Machine Learning and Sentiments

Using computational methods to transform empirical data into useful models is a topic of research known as "machine learning." Traditional statistics and artificial intelligence groups gave birth to the machine learning area. There are a variety of applications for machine learning algorithms, including gathering knowledge about the cyber phenomenon that generated the study's data, abstracting that knowledge into models, forecasting future value changes, and spotting anomalous behaviour displayed by the phenomena under observation. Many machine learning techniques have open-source implementations that may be utilised with API calls or non-programmatic applications [22].

People's comments, reviews, and opinions are crucial in determining if a specific population is happy with the product or service. It aids in anticipating the emotion of an extensive range of individuals on a specific event of interest, such as a movie review or their opinion on numerous topics circulating throughout the world. These details are critical for sentiment analysis [23]. The speaker's or writer's attitude toward a subject or the broad contextual polarity of work defines the sentiment of the writer. It is currently mostly accomplished through blog postings, internet forums, product review websites, social media, and other means. Thousands of individuals use social networking sites such as Facebook, Twitter, Google Plus, and others to express their feelings, opinions, and share perspectives regarding their everyday lives. Thus, we get a participatory medium where consumers engage and impact others through discussions and web forums. Tweets, status updates, blog posts, comments, reviews, and other forms of social media data generate a significant volume of sentiment-rich data. Furthermore, social media gives a platform for businesses to communicate with their consumers to advertise. To a large extent, people rely on user-generated content on the internet to make decisions. For example, if someone wants to buy a product or service, they will first research it online before deciding. The amount of material created by users is just too large for a typical user to examine. As a result of the necessity to automate this, several approaches for sentiment analysis are extensively employed [26].

There is a rising need for automated sentiment/opinion detection systems. Businesses are excited to know popular criticisms of their product lines and learn the results of their marketing campaigns; elected officials are keen to evaluate their picture among the voting population; and ordinary people, who are overwhelmed by the amount of text on the internet about almost any topic, with lots of different voices but little trustworthiness, needs some automated assistance. However, the lack of standardization, even the absence of precise definitions of the key subjects under discussion, hampered any meaningful progress [27].

Sentiment analysis of miniature contributing to a blog is an ongoing exploration subject, so there is still space for additional examination in this field. Respectable measures of earlier work were analyzed based on client input, papers, web journals/articles and general expression level sentiment analysis. These varieties are for the most part because of the limit of characters per tweet, which requires the client to offer viewpoints packed in exceptionally short text. The best outcomes in feeling grouping utilize managed learning procedures like Naive Bayes and backing vector machines. Some work has been finished on unaided methodologies and semi-directed approaches, and there is a lot of room for progress. Different specialists testing new highlights and grouping procedures likewise contrast their discoveries and basic line results. Legitimate and formal correlations between these outcomes are expected across various elements and grouping methods to pick the best highlights and best order strategies for explicit applications.

3.5 Reaearch Contributions

A scientific approach is adopted to identify the various features like time, joviality, tranquilly, surprise, anxiety, depression, aggression, shyness, fatigue, attentiveness and etc. For various tags. The work in the thesis considered the hash tags related to politics, stock exchange, movies, sports and technology.

To distinguish various boundaries and for dissecting the sentiment analysis of different hash labels, we have utilized vocabulary-based AI furthermore, profound learning approaches, for example, naive bayes, random forest from ai and convolution neural networks from deep learning applied to a wide assortment of tweets from movies, sports, politics, technology furthermore, stock exchange individually.

3.6 Variouus Sentiment Analysis Algorithms

Social computing is an increasingly innovative example of interpretation and modelling of social experiences on multiple networks. Intellectual and collaborative applications are developed to deliver productive results. The universal availability and use of social media platforms encourages individuals to express their views on a single event, substance, or issue. It is very useful for drawing conclusions about various events, topics, issues, products etc. by mining these informal and homogeneous results. Nonetheless, the highly unstructured format of opinions on the web presents challenges to the mining process. The texts published on the web are mostly divided into one of the two groups: data of facts and data of feelings [33]. The objective terminology related to different entities, problems and events constitutes factual data. While feeling information is the subjective term that defines the views or beliefs of individuals for a specific entity, product, or event. Analysis of sentiments is the process of identification and classification of individuals' different feelings online, so that their response to a particular product, subject or event is decided by the author whether it is positive, negative, or neutral.

Social processing is an undeniably imaginative illustration of translation and social encounters on numerous organizations. Scholarly and cooperative applications are created to convey useful outcomes. The texts distributed on the web are generally isolated into one of the two gatherings: information of realities and information of sentiments. The objective phrasing connected with various substances, issues and occasions is genuine information. While feeling data is the emotional term that characterizes the perspectives or convictions of people for a particular element, item, or occasion. Analysis of sentiments is the course of recognizable proof and characterization of people's various sentiments on the web, so their reaction to a specific item, subject or occasion is chosen by the creator whether it is positive, negative, or impartial. The feelings communicated in the text are both immediate and near in the sentimental analysis strategy.

Tokenization makes sense of the system by which a text body is isolated into individual components for contribution of different natural language calculations. Other treatment steps, like disposal, stemming or lemmatization of stop words and accentuation characters, and the production of n grams, are commonly trailed by tokenization. The design of the momentous brain networks is unmistakable from the typical brain organizations. Standard brain networks convert inputs into obscure layers. This layer comprises of a succession of neurons wherein each layer is associated completely with the past layer's neurons. Eventually, the last completely associated yield layer addresses the objectives. There are fairly unmistakable brain organizations. The layers are isolated into three aspects in any case: distance, level, and profundity.

3.7 Model Architecture

3.7.1 Supervised Machine Learning Classifiers

Supervised machine learning is a technique whose task is to deduce a function from marked samples of training. The training samples for supervised learning provide a wide variety of examples for a particular subject. Each example of training data comes in a pair of input and output value in supervised learning [35]. These algorithms process data and create an output function to map new sets of data to each category. The various classifiers for machine learning that use to construct classifiers are:

- Naïve-bayes classifier
- Multinomial NB classifier
- Bernoulli NB classifier
- Logistic regression classifier
- SGDC (stochastic gradient decent classifier)
- SVC (support vector classifier): linear SVC and nu SVC

1. Naïve-bayes classifier

Naïve-bayes classifiers are probabilistic classifiers. This characterization frameworks are in view of the use of Bayes' theorem with an unmistakable free suspicion among the various highlights [36]. We should assume that there's a x1 to xn subordinate vector furthermore 'y' type variable. Corresponding to bayes:

$$p(y)|\, x_{1,\ldots\ldots,}x_n = \frac{p(y)p(x_{1,\ldots.,}x_n|y)}{p\ (x_{1,\ldots.,}x_n|y)}$$

Now in relation to assumption

N

$$p\big(\, x_i\, |y, x_{1,\ldots.,}x_i - 1, x_i\ + 1, \ldots., x_n\big) = P(x_i|y)$$

This function becomes for any i,

$$p(y)|\, x_{1,\ldots\ldots,}x_n = \frac{p(y)\ \prod_{i=1}^{n} p(x_i\,|u)}{P(x_{1,\ldots..,}x_n)}$$

P (x1,, xn)is constant on the given input, the classification rules can be applied:

$$p(y|x_{1,\dots\dots,}x_n)\ \alpha\ P(y)\prod_{i=1}^{n} p\ (x_i\,|y)$$

$$\Downarrow$$

$$\hat{y}\ \underset{y}{argmax}\ p(y)\prod_{i=1}^{n} p\ (x_i\,|y)$$

And for approximation use maximum-a-posterior estimation p(y) and p (xi | y);

2. Multinomial NB classifier

Multinomial NB further develops NB algorithm utilization. It applies NB for multinomially dispersed information and use one of its variations to distinguish text [36]. The conveyance information was defined by vectors $\boldsymbol{\theta_y} = (\boldsymbol{\theta_{y|,\dots\dots,}\theta_{ym}})\boldsymbol{f}$for each, where „n‟ gives the whole features and likelihood $p\ (xi\mid y)$ that show in the example of class „y‟ $is\ \theta yi$. Smoothed rendition of most extreme opportunity for assessment of boundary y, which is relative recurrence of counting:

$$\hat{\boldsymbol{\theta}}_{yi} = \frac{N_{yi} + \alpha}{N_y\ + \alpha n}$$

Where nits and for number of times „i‟ come into view in any sample of class „y‟ which fit into training sample test.

3. Stochastic gradient decent classifier

In multiclass, the classification coefficient only consists of a 1d array of shape = [classes] and a 2d array of form = [classes, features]. The ith class ova classifier weight quantity is included in the ith coefficient matrix row. Classes are being scheduled more and more.[36]

4. Bernoulli NB classifier

In correlation, Bernoulli NB applies the NB preparing and order algorithm. Nb is utilized for Bernoulli multivariate information circulation; for example, a few highlights can be utilized, however, every one of the highlights has a binary worth or a valid or bogus boolean variable. Thusly each class needs tests to be communicated in binary worth factors. Bernoulli NB will binaries its feedback assuming one more sort of information is given.

$$p(x_i\,|y) = p(i\,|y)x_i + \big(1 - p(\,i|y)\big)(1 - x_i)$$

This regulation, which varies from multinomial NB's standard, unequivocally rebuffs any unavailability of a capability that capabilities as a class highlight, where as in multinomial, it plainly disregards any non-happening highlights. [38]

5. **Logistic regression classifier**

It is a straight order model. Different titles additionally call this model maximum-entropy order. This model purposes a calculated capability, where likelihood addresses the result of one preliminary. It tends to be applied from python's scikit-learn library with a class called logistic relapse. This execution suits multi-class one-versus rest relapse with optional L1 or L2 regularization [36].

L2 penalized logistical regression assist minimize the cost function:

$$\underset{w,c}{min} \frac{1}{2} w^T w + c \sum_{i=1}^{n} \log\left(\exp\left(-Y_i \ (X_i^T \ w + c)\right) + 1\right)$$

L1 regularize logistics regression can resolve the subsequent optimization problem:

$$\underset{w,c}{min} \ \| w \|_1 + c \sum_{i=1}^{n} \log\left(\exp\left(-Y_i \ (X_i^T \ w + c)\right) + 1\right)$$

1. **SVC (support vector classifier): linear SVC and nu SVC**

Machine preparing strategies for gathering relapse and recognition models are managed by SVM. For high dimensional space, SVM are more effective. SVCs are fit for gathering of various grades. Linear SVC depends on straight parts while SVC and nu SVC are indistinguishable. These SVCs all take the two information clusters: a size x exhibit [samples, characteristics], what's more, a size y cluster [samples]. Nu SVC carries out the multi-class 'one-against-once' plot, offering a straightforward connection point to other people. In correlation, linear SVC utilizes 'one-versus rest' nu SVC is based on the library 'libSVM,' while the finishing of linear SVC is based on the library. A SVM gathering, relapse and different exercises are completed utilizing hyper planes. [39]

3.7.2 Unsupervised Machine Learning Classifiers

Unsupervised learning algorithms are used to categorise instances based on similar characteristics or naturally occurring trends, patterns, or connections in the data. Self-organizing maps are another name for these concepts. Clustering methods and self-organizing maps are examples of unsupervised models. Different algorithms employ various techniques for categorizing data. Some approaches are quite simple, rapidly grouping instances into groups based on shared features or some other similarity. [37]

3.8 Evaluation Metrics

We will analyse the results using Confusion matrix**.** A confusion matrix summarizes the performance of a classification model by comparing the predicted and actual class labels of a set of data points. It is a square matrix with the same number of rows and columns as the number of classes in the classification problem. The rows represent the actual or true classes of the data points, while the columns represent the predicted classes. Each cell contains a count of the number of data points that belong to a particular combination of actual and predicted classes. There are four terms that are commonly used in a confusion matrix:

True Positives (TP): the number of data points that are correctly classified as positive (i.e., the model correctly predicts the positive class).

False Positives (FP): the number of data points that are incorrectly classified as positive (i.e., the model incorrectly predicts the positive class when it should be negative).

True Negatives (TN): the number of data points that are correctly classified as negative (i.e., the model correctly predicts the negative class).

False Negatives (FN): the number of data points that are incorrectly classified as negative (i.e., the model incorrectly predicts the negative class when it should be positive).

Using these four terms, we can calculate several evaluation metrics to assess the performance of a classification model, such as accuracy, precision, recall, and F1 score.

Accuracy: the proportion of correctly classified data points out of the total number of data points. It is calculated as (TP + TN) / (TP + TN + FP + FN).
Precision: the proportion of true positives among all predicted positives. It is calculated as TP / (TP + FP).
Recall: the proportion of true positives among all actual positives. It is calculated as TP / (TP + FN).
F1 score: the harmonic means of precision and recall. It is calculated as 2 * (precision * recall) / (precision + recall).

A confusion matrix is a useful tool for understanding the strengths and weaknesses of a classification model. It provides a detailed breakdown of the types of errors the model is making and can help identify areas for improvement.

3.9 Motivation Of The Study

Analysis of the opinion demonstrates to the usage of NLP, text analysis, machine phonetics and biometrics in the precise identification, extraction, evaluation and investigation of effects and the abstract information. Feeling analysis generally centres around recognizing whether the statement of assessment about an individual, news story, and computerized content is positive or negative. Knowing the feelings communicated by the clients helps organizations, specialist co-ops and individuals to introspect their systems to work on their administrations. It additionally helps organizations in getting the genuinely necessary criticism on their items and areas of progress. Opinion analysis tells the world moving right now, evolving designs, tastes and assessments of individuals across the world. Currently traditional techniques exist to play out the opinion analysis for organized information. Anyway, the equivalent isn't true with unstructured information like the sentiments communicated on different virtual entertainment stages. This has given us enough inspiration to research on tweaking the current methods and to make a superior one.

The writing contains a choice of outstanding scholastic writing on feeling examination. The applicable issues, procedures, and arrangements introduced in the writing have featured. As a result, this segment planned to depict the technique used to choose writing. Santos, et al. (2014) [40] in his work, proposed another profound convolutional brain network which takes advantage of from character-to sentence-level data for performing opinion investigation of short texts. This approach is applied for two corpora of two unique areas: the Stanford sentiment treebank (SSTB) that contains sentences from film audits; and the Stanford twitter sentiment corpus (STS) that contains twitter messages. For the SSTB corpus, this approach accomplishes cutting edge results for single sentence feeling expectation in both parallel positive/negative characterization, with 85.7% precision, and fine-grained order, with 48.3% exactness. For the STS corpus, this approach accomplishes an opinion expectation precision of 86.4%.

3.10 Related Work

The literature contains a selection of notable academic literature on sentiment analysis. The relevant issues, problems, strategies, and solutions presented in the literature have been highlighted. As a consequence, this section intended to describe the method used to select literature.

Pak A, & Paroubek P., (2010) [1] conducted a study on the use of Twitter as a corpus for sentiment analysis and opinion mining. They analyzed a large collection of tweets related to different topics to determine the sentiment polarity of the tweets. They proposed a method for preprocessing Twitter data, including the removal of hashtags, URLs, and user mentions, and the use of emoticons and punctuation marks for sentiment analysis. They also evaluated different machine learning techniques for sentiment analysis and compared their results to a manually annotated dataset. The study concluded that Twitter is a valuable source for sentiment analysis and opinion mining and can be used to build accurate models for sentiment analysis.

Go et al. (2009) [2] proposed a method for Twitter sentiment classification using distant supervision. They used a large dataset of tweets and used emoticons as noisy labels to train a classifier for sentiment analysis. The study used machine learning algorithms and evaluated their performance on a manually annotated dataset of tweets. The results showed that their approach was able to achieve high accuracy in predicting sentiment on a large dataset of tweets.

Agarwal et al. (2011) [3] conducted a study on sentiment analysis of Twitter data. They collected a large dataset of tweets related to different topics and used machine learning algorithms to classify them into positive, negative, and neutral sentiment categories. They proposed a method for feature selection, including n-grams, POS tags, and sentiment lexicons, to improve the performance of sentiment analysis. The study evaluated the performance of their approach on a manually annotated dataset and showed that their method was able to achieve high accuracy in sentiment classification.

Liu et al. (2012) [4] provided a comprehensive overview of sentiment analysis and opinion mining, including different approaches and techniques for sentiment analysis. The book covered different domains of sentiment analysis, including social media, product reviews, and political texts. The author discussed the challenges of sentiment analysis, such as ambiguity, sarcasm, and irony, and provided solutions for these challenges. The book is a valuable resource for researchers and practitioners in the field of sentiment analysis.

Saif et al. (2012) [5] proposed a method for semantic sentiment analysis of Twitter using semantic information from WordNet and FrameNet. They used a semantic lexicon to map tweets into sentiment categories and evaluated their approach on a manually annotated dataset of tweets. The study showed that incorporating semantic information into sentiment analysis can improve the accuracy of sentiment classification. The authors also discussed the limitations and challenges of their approach, such as the ambiguity of words and the need for domain-specific lexicons. The study is a valuable contribution to the field of sentiment analysis and can be used to improve the accuracy of sentiment classification on social media platforms like Twitter.

Taboada et al. (2011) [6] proposed lexicon-based methods for sentiment analysis, which are a rule-based approach that relies on a pre-defined list of words with their corresponding sentiment polarities. The authors demonstrated that such lexicon-based approaches can produce good results on a variety of text genres, including news articles, product reviews, and movie reviews, and that the use of additional features such as part-of-speech tags and negation handling can further improve performance.

Duan et al. (2012) [7] provided an overview of sentiment analysis with big data, discussing the challenges and opportunities of analyzing large volumes of text data. The authors highlighted the importance of efficient algorithms and distributed computing frameworks, as well as the need for domain-specific knowledge and human annotation to improve accuracy.

Pak, A., and Paroubek, P. (2011) [8] presented a case study on cross-lingual sentiment analysis, focusing on English and French tweets. The authors evaluated different machine translation methods and found that a phrase-based approach yielded the best results. They also discussed the challenges of cross-lingual sentiment analysis, such as linguistic differences between languages and the availability of labeled data.

Mohammad, M., and Turney, D., (2013) [9] proposed a crowdsourced word-emotion association lexicon, which is a list of words with their associated emotions. The authors showed that such a lexicon can achieve good performance on several benchmark datasets and can be used for various applications, such as sentiment analysis and emotion detection.

Tumasjan et al. (2010) [10] studied the use of Twitter for predicting political sentiment and election outcomes. The authors collected and analyzed over 100,000 tweets related to the 2009 German federal election and found that the sentiment expressed in tweets can be a good predictor of election results. They also compared the performance of different sentiment analysis methods and found that lexicon-based approaches were the most effective.

Read, et al. (2009) [11] In this paper, the authors propose a weakly supervised technique for sentiment classification that does not require labeled training data. Instead, the technique relies on a set of seed words and a large unlabeled corpus to automatically induce a sentiment classifier. The authors show that this technique can achieve performance comparable to that of fully supervised classifiers on several different domains, including movie reviews, product reviews, and political blogs. The proposed approach is based on a generative model that represents each document as a mixture of topics, each of which has an associated sentiment polarity. The authors use an Expectation-Maximization algorithm to learn the parameters of the model.

Gao, et al. (2012) [12] Microblog sentiment classification using a coupled generative/discriminative approach. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 169-173).
This paper proposes a novel approach to sentiment classification of microblogs (such as tweets), which combines a generative model with a discriminative model. The generative model captures the underlying topic structure of the microblogs, while the discriminative model uses lexical and syntactic features to classify the sentiment of each microblog. The proposed approach is evaluated on a dataset of tweets, and the results show that the approach outperforms several baselines and achieves state-of-the-art performance on this task.

McKeown, et al. (2009) [13] This paper proposes a novel approach to sentiment analysis that performs phrase-level polarity classification based on a combination of lexical affect scoring and syntactic n-grams. The proposed approach is evaluated on a dataset of movie reviews, and the results show that it outperforms several baselines, including both lexicon-based and machine learning approaches. The authors also show that their approach can be adapted to other domains, such as product reviews and political blogs, with only minor modifications.

Kouloumpis, et al. (2011) [14] This paper presents an evaluation of several different approaches to sentiment analysis of tweets, including both lexicon-based and machine learning approaches. The authors evaluate these approaches on a dataset of tweets related to the 2010 World Cup, and the results show that machine learning approaches outperform lexicon-based approaches on this task. The authors also analyze the errors made by the different approaches and provide insights into the strengths and weaknesses of each approach.

Mohammad, S. and Turney, P., (2010) [15] Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (pp. 26-34). This paper describes a crowdsourcing approach to building an emotion lexicon, which consists of words and phrases that evoke specific emotions (such as joy, sadness, anger, and fear). The authors use the Amazon Mechanical Turk platform to collect ratings of the emotional content of a large number of words and phrases.

Bifet, A. and Frank, E., (2010) [16] The paper presents a framework for mining sentiment information in real-time from Twitter data streams. The proposed framework is based on the concept of incremental learning, which allows the model to adapt to changes in the data stream over time. The authors also present an evaluation of their approach on a large-scale dataset of Twitter messages, demonstrating the effectiveness of their approach in terms of accuracy and efficiency.

Bermingham, A. and Smeaton, A.F., (2010) [17] This paper investigates whether the brevity of microblog messages, such as those on Twitter, is an advantage or disadvantage in sentiment analysis. The authors conducted experiments using various machine learning algorithms and feature sets to classify sentiment in microblog messages. They found that, despite the brevity of microblogs, they can still be used effectively for sentiment analysis and that certain features, such as emoticons and hashtags, can improve the accuracy of the classification.

Barbosa, L. and Feng, J., (2010) [18] This paper addresses the problem of sentiment analysis on Twitter data that is noisy and biased. The authors propose a method that utilizes distant supervision and semi-supervised learning to improve the robustness of sentiment detection on Twitter. They also introduce a new dataset, called the Twitter Sentiment Corpus, which consists of manually labeled tweets that have been annotated for sentiment. The evaluation of their approach on this dataset shows that their method is effective in handling noisy and biased data.

Chakraborty, et al. (2020) [19] This paper proposes a method for improving sentiment analysis on Twitter by incorporating contextual semantic analysis. The authors utilize a combination of a pre-trained sentiment lexicon and a word embedding model to capture the context of the tweet. They also present an evaluation of their approach on a dataset of tweets related to the 2012 US Presidential election, demonstrating the effectiveness of their method in improving the accuracy of sentiment analysis.

Kharde, V. and Sonawane, P., (2016) [20] This paper provides a comprehensive survey of techniques used for sentiment analysis on Twitter data. The authors discuss the challenges associated with sentiment analysis on Twitter data, such as the brevity and informality of the messages, and present a taxonomy of the various techniques used for sentiment analysis. They also provide an analysis of the strengths and weaknesses of each technique and highlight some of the future directions in sentiment analysis research.

Wang et al. (2012) [21] developed a system for real-time sentiment analysis of Twitter data during the 2012 US presidential election cycle. The system used a combination of natural language processing techniques and machine learning algorithms to analyze tweets and determine their sentiment. The authors used a set of labeled tweets to train their models and evaluated the system's performance on a separate test set. They found that their system was able to accurately classify the sentiment of tweets in real-time and provided valuable insights into public opinion during the election.

4 Practical Part

There are multiple steps required to get desired results in performing sentiment analysis. This starts from collecting the relevant data using twitter open API. After collecting the data, we will process it by methods shown in chapter 4.2, i.e., cleaning it to make the data available for the model. This includes removing the stop words, special characters, punctuations, etc. After this step, we assume that the data is available for training the model. In the training part, we will feed the data to multiple algorithms in order to check the correctness of the algorithm. The algorithms include, Logistic Regression, Random Forest classifier and some more to get the desired results. Based upon the outcome of the models shown in chapter 5, we will define the algorithm suitable for the subsequent modelling.

4.1 Datasets

Gathering the required data and perparing is the first and most cruicial part of this process. For the Model training and accurecy testing, I prepared 2 datasets,

1. **Train.csv** – It is a labelled dataset of 1.6 Mil Tweets. The dataset is provided in the form of a csv file with each line storing a tweet id, its label, and the tweet.
2. **Test_tweets.csv** - The test data file contains only tweet ids and the tweet text with each tweet in a new line.

Train.csv is used as a training dataset. Training set helps the model to identify patterns, correlations, and features that characterize the relationship between inputs and desired outcomes during the training stage. The training set forms the foundation on which most machine learning models sharpen their cognitive abilities. This is a carefully selected set or collection of data instances that are meant to train the model on the nuances involved in a certain task. In essence, this set is like a course that presents illustrations featuring well-thought results to boost the model's training procedure. The model fine tunes itself by making repetitive revisions on its inner settings so as to sharpen the predictions.

Test_tweets.csv dataset was utilized as the test set. Test set contains random samples of data instances which have never been presented to the system in both training and validation stages. The model is then evaluated with regard to its ability to generalise based on unseen data and predict reliably in unknown contexts. Putting the training and test set apart prevents a purely inductive behavior of the model by not memorizing the particularities of training instances only.

Significantly, the test set serves as a neutral judge, whose views into the model's ability on unlearned tasks are taken. Success on the test set convinces that model can predict true values in practice when it will be given new tasks. Ultimately, the test set serves as the final proof of the workability of a machine learning model outside the regulated boundaries of training and verification in real life.

4.2 Data Preparation

A data mining approach involves a variety of categorization techniques that assist in identifying spam in user-sent tweets. Various characteristics are defined by compiling data from all of the user's tweets that were sent and received, Categorized using this method. The features in the twitter dataset, such as "id," "no of followers," "text," "created at," and "statuses count," among others, contain information that is both repetitive and valuable. The "curse of dimensionality" is the term used to describe how decrease in conducting categorization increases due to the enormous space in irrelevant and redundant. Relevant features are helpful for categorization, and feature extraction works well. The redundant and irrelevant features are removed using the sum extraction technique to speed up the classification process and cut down on time.

The technique of extracting specific features that are useful for processing categorization is known as feature extraction. Even though numerous feature extraction strategies have been put forth, the accuracy of the classification process still falls short when applying machine learning classifiers and optimization techniques. Since most datasets are explicit, binary, and continuous, the real-world data are gathered in a continuous manner. Consequently, the effectiveness of data mining classification algorithms and features with continuous format suffer greatly.

1. Data loading
2. Feature engineering
3. Data preprocessing
    - Text normalization
    - Vectorization
    - smote
4. ML modeling
5. Hyperparameter tuning
6. Prediction submission

7.	**Id**	**Lebel**	**Tweet**
0	1	0	@ user when a father id dysfunctional and is…
1	2	0	@user @when thanks for #lyft credit I can’t us
2	3	0	Birthday your majesty
3	4	0	#model I love u take with u all time in
4	5	0	Facts guide society now # motivation

Table 1 Tweets and assigned Labels.

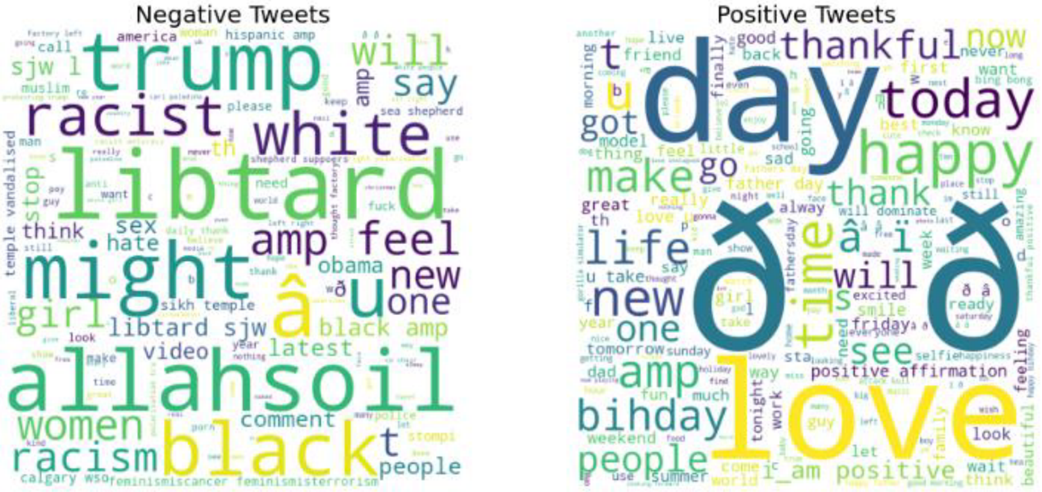
	Id	**Tweet**
0	31963	#studio life #a selfie #requires #passion #decic…
1	31964	@user #white #supremaciSTS want everyone to a
2	31965	Safe ways to heal you #acne!!
3	31966	Is the hp and the cussed child book up for res…
4	31967	3rd #birthday to my amazing. Hilarious #nephew….

Table 2 Tweets and assigned IDs.

Among the many jobs that sentiment analysis may perform is identifying the sentiments inside an input and its categorization into various classes according to the sentiment present in the text. When there are multiple classes of emotions, sentiment categorization positions the input dividing texts into categories like cheerful, sad, angry, etc.

As we can see in Figure 1, the tweets can be classified into positive and negative based on the some specific keywords. Words Like Racism and Vandalised can be categorised into negative while Birthday and thankful will fall under positive category.

The second analytical scenario focuses on determining the sentiment's polarity, whether it is "positive, negative or indifferent this process is called polarity detection. The scope of this study is modest to the identification of polarity in twitter data. The process of determining the sentiment polarity of a given text is known as polarity detection. It aids in classifying the texts that are presented into many categories as either positive or negative. State-of-the-art offers a variety of strategies for detecting polarity. They primarily use a lexicon-based machine learning method. Deep learning approach, for example. There are also combinations of the methods used to increase polarity detection's precision.



Negative Tweets

Positive Tweets

Figure 1 Example of Negative and Positive words

- **Feature engineering**

To categorise text into classes, a machine learning-based approach uses classification techniques. There are primarily two kinds of machine learning approaches.

For machine learning algorithms to work, the main characteristics of text or documents must be represented. These important traits are defined as feature vectors, which are then employed in the classification job. Some examples of features include:

- **Negation**: Negation is an essential yet difficult to comprehend characteristic. The existence of a negative generally shifts the opinion's polarity.
- **Keywords and Its Frequencies**: Unigrams, bigrams, and n-gram models, along with their statistical analysis, are regarded as attributes.
- **Opinion Words and Phrases**: In addition to particular words, phrases and idioms that communicate emotions can be utilised as features.

In Table 3, we extracted a few characteristics of the tweets to help us separate negative tweets from positive ones.

	Id	**Lebel**	**Tweet**	**Tweet_lenght**	**Num_hashtag**	**Num_questions_marks**	**Total_tags**	**Num_puntuations**	**Num_words**
0	1	0	@ user when a father id dysfunctional and is…	102	1	0	1	1	18
1	2	0	@user @when thanks for #lyft credit I can't us	122	3	0	2	1	19
2	3	0	Birthday your majesty	21	0	0	0	0	3
3	4	0	#model I love u take with u all time in	86	1	3	0	1	14
4	5	0	Facts guide society now # motivation	39	1	2	0	1	13

Table 3 Dataset

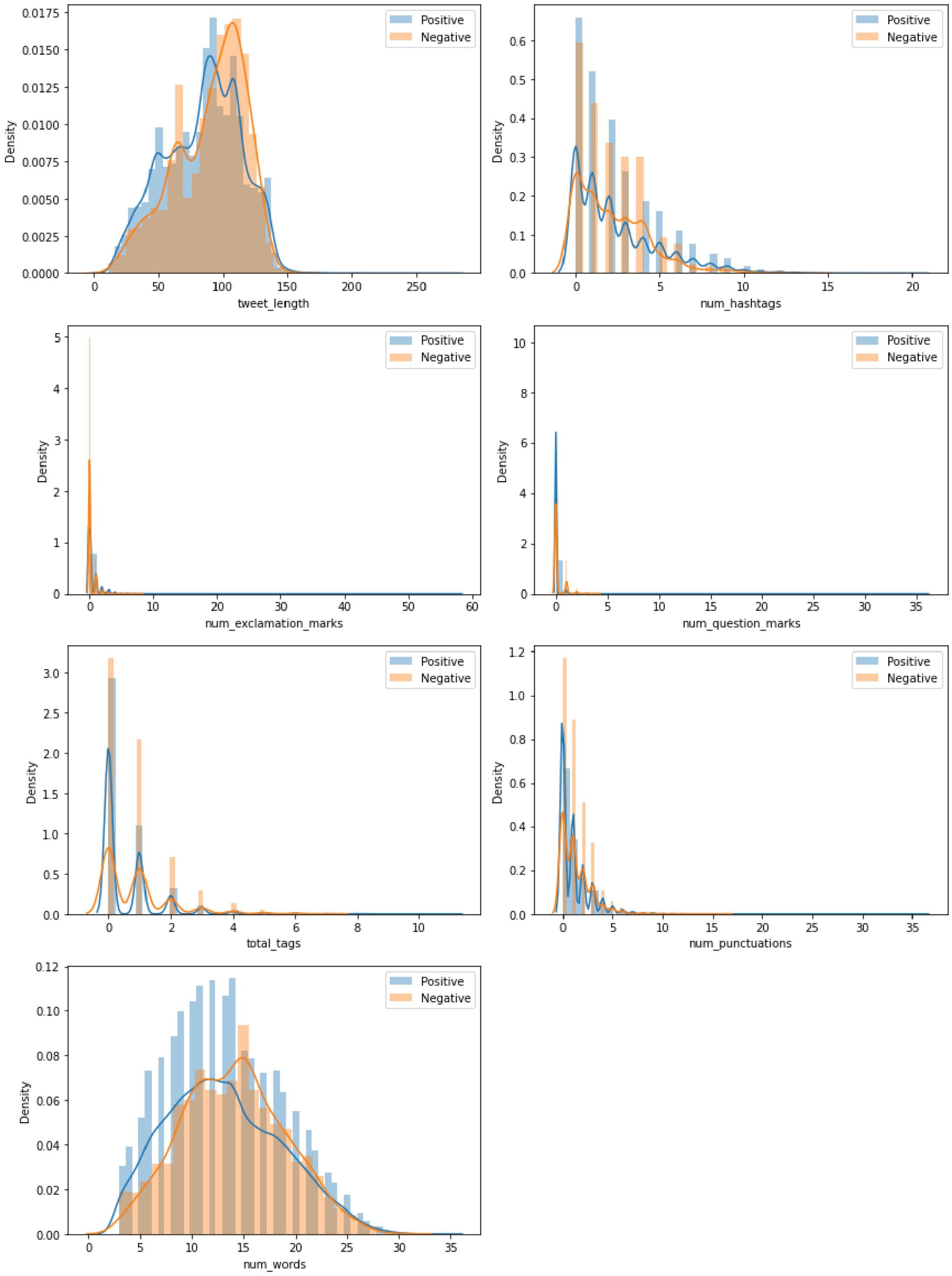


Figure 2 Polarity of tweets

The Figure 2 show that it is difficult for the engineered features to distinguish significantly between good and negative tweets. I tried to include these variables in our final model for predicting tweet sentiments, however it actually made the performance worse. I decided not to add these features in this notebook as a result.

4.3 Data pre-processing

4.3.1 Text normalization

Text normalization is the process of converting text input into a canonical or standardized form so that machine learning algorithms can examine it more efficiently. The format, structure, and meaning of text data can be standardized through the implementation of several strategies in this process.

Following are some common techniques of text normalization used in the process:

1. Remove any URLs (for example, www.xyz.com), hash tags (for example, #subject), and targets (for example, @username).
2. Correct the spellings; repeating characters must be handled in a certain order.
3. Change the sentiments of all the emoticons.
4. Remove any punctuation, symbols, and numbers from the text.
5. Eliminate Stop Words.
6. Acronyms Should Be Expanded (we can use an acronym dictionary).
7. Tweets that aren't in English should be removed.

4.3.2 Vectorization

In vectorization, we converted text data into numerical vectors that can be processed by machine learning algorithms. In this step each tweet is converted to a numerical vector using TF-IDF (term frequency - inverse document frequency) vectorization. In the case of TF-IDF vectors, the process of vectorization involves the following steps:

1. Tokenization: The text data is split into tokens, or single words.
2. Counting: Each token's frequency is measured over all the dataset's documents.
3. The term frequency (tf) and inverse document frequency (idf) scores are calculated for each token. idf is the logarithmically scaled inverse fraction of the documents that include the token, and Tf is the number of times a token appears in a document.
4. Vectorization: Each document in the dataset is then given a numerical vector based on the tf-idf scores. The tf-idf scores for each token in the document are contained in the vector.

I tried using n-grams, but it did not improve the model performance. Moreover, TF-IDF outperforms bow (bag-of-words) vectorization; therefore, I am further including only TF-IDF model results.

4.3.3 Smote

Synthetic minority class samples are produced via the synthetic minority over-sampling technique (smote), an oversampling strategy. It is frequently used and might perform better than straight forward oversampling.

Smote has been used, for instance, to identify network intrusions, speech sentence boundaries, species dispersion predictions, and breast cancer. Smote is also utilized in bioinformatics for histopathological annotation, the prediction of miRNA genes, the identification of the binding specificity of regulatory proteins, and the identification of photoreceptor-enriched genes based on expression data.

4.3.4 Polarity

Three categories, Positive, Negative and Neutral, are used to categorise a dataset. SentiWordNet 3.0.0 dictionary is used to categorise tweets. The 117659 word is included in the SentiWordNet 3.0.0 vocabulary. Positive and negative polarity are included in each word. The term is considered neutral if its positive and negative polarity is equal to zero. SentiWordNet 3.0.0 is used to determine the polarity of the words in a single tweet once it has been broken into words. Positive and negative polarity for each word is added once the polarity has been determined. After that, the comparison of positive and negative word polarity is completed. Classify a tweet as having a positive polarity if it contains more sentences with that polarity. Negative and neutral polarity are equivalent. A training dataset did not take duplicate tweets into account.

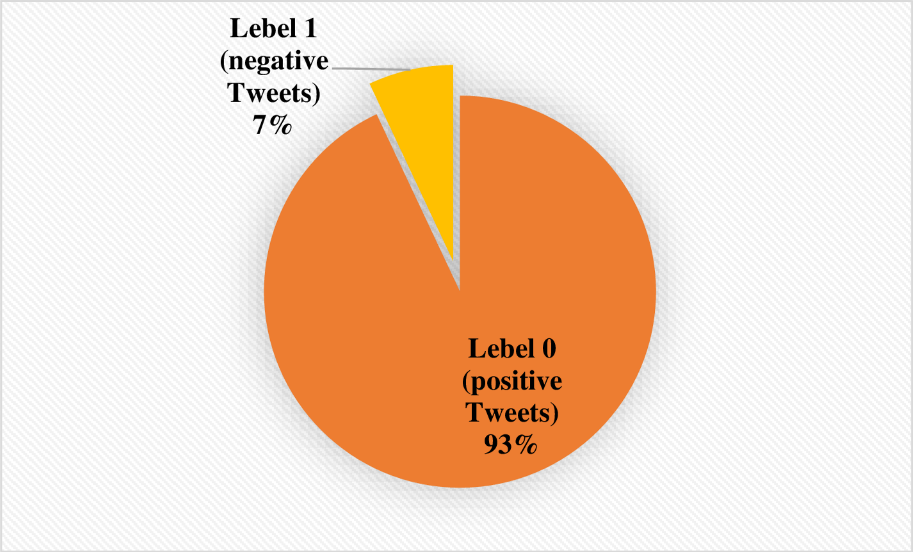


Figure 3 Pie chart of positive and negative label

Lebel 0 (positive tweets)	Lebel 1 (negative tweets)
93.0%	7.0%

Table 4 Label 1 and label 0

In Figure 3 and Table 4, we can see that there is a high-class imbalance in the dataset. We'll use smote (synthetic minority over-sampling technique) to balance the class. We tried using the random oversampling technique, but smote generated better predictions, so the same is used ahead. And the results in Figure 4 and Table 5 shows that our dataset is very balanced.

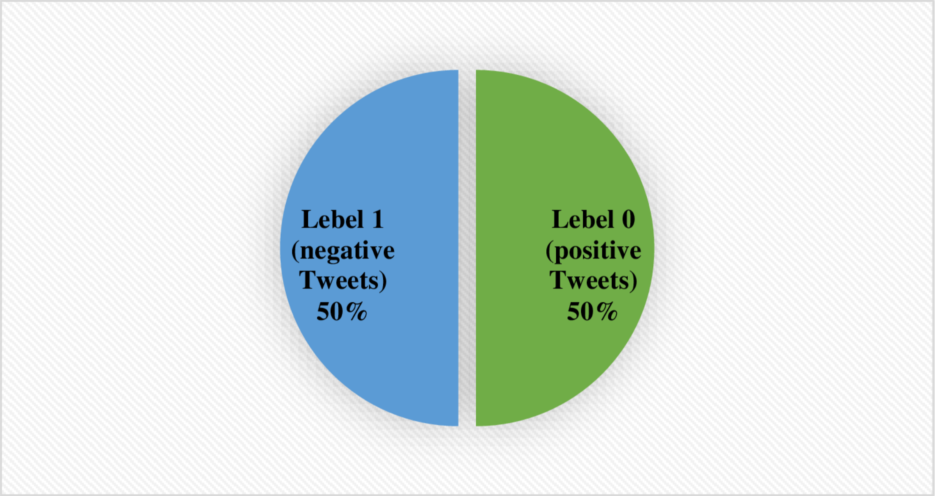


Figure 4 Positive and negative tweets

Lebel 0 (positive tweets)	Lebel 1 (negative tweets)
50.0%	50.0%

Table 5 Positives and negative tweets

4.4 ML Modelling

The evaluation metric used in the competition is F1-score; therefore, we will try to maximize the same.

4.4.1 Logistic Regression

An illustration of supervised learning is logistic regression. It is used to determine or forecast the likelihood that a binary (yes/no) event will occur. One use of machine learning to identify whether a person is likely to have the COVID-19 virus or not is an example of logistic regression. Binary classification refers to the fact that there are only two answers to this question: either they are infected, or they are not.

4.4.2 Naïve Bayes Classifier

A probabilistic classifier called Naive Bayes returns the likelihood that a test point belongs to a class rather than the test point's label. Although it is one of the most fundamental Bayesian network models, by using kernel density estimation, it may achieve higher levels of accuracy. Unlike many other ML algorithms, which often perform both Regression and Classification tasks, this algorithm is exclusively appropriate for Classification jobs.

Naive Because the assumptions made by the Bayes algorithm are so improbable to be supported by empirical data, the algorithm is regarded as naive. To determine the sum of the individual probabilities of the components, conditional probability is used. This indicates that, given the class variable, the algorithm assumes that the presence or absence of a particular feature of a class is independent of the presence or absence of any other feature (absolute independence of features).

4.4.3 Random Forest Classier

A supervised learning approach called random forest is employed for both classification and regression. But it is primarily employed for classification issues. As is common knowledge, a forest is made up of trees, and a forest with more trees will be sturdier. Similar to this, the random forest algorithm builds decision trees on data samples, obtains predictions from each one, and then uses voting to determine the optimal option. Because it averages the results, the ensemble method—which is superior to a single decision tree—reduces over-fitting.

4.4.4 Extreme gradient Boosting Classifier

A class of machine learning techniques known as gradient boosting classifiers combines a number of weak learning models to produce a powerful predicting model. Gradient boosting frequently makes use of decision trees. Due to their success in categorizing large datasets, gradient boosting models are gaining popularity and have lately been successful in numerous Kaggle data science challenges.

The Scikit-Learn machine learning toolkit for Python offers a variety of XGBoost implementations of gradient boosting classifiers.

5 Results and Discussion

We will analysis performance of our Models that we used to classify and perform our sentiment analysis. The data is used for these Machine Learning models are already Balanced by our SMOTE algorithm. This performance analysis is done using Confusion matrix. To test the results and assess the performance of the model, we will use Training and Validation score. The training score indicated how well our model is learning from the training dataset and serves as an initial indicator of its performance. While the validation score provides an independent measure of data that the model did not encounter during training. In addition, a validation set is used to adjust the model's hyperparameter during training in order to avoid over-fitting.

We used Various methods such as training and Validation Scores, Accuracy F1 Score and Other Matrix. The outcomes are measured in between the score of 0 to 1, in which closer to 0 represents weaker while closer to 1 is considered positive.

Logistic regression

Training scores:	**Accuracy=0.979**	**F1-score=0.98**
Validation scores:	**Accuracy=0.931**	**F1-score=0.624**

Table 6 Training scores of Logistic regressions

The training scores in Table 6 and confusion matrix in Figure 5 indicate that the model has achieved an accuracy of 0.979 and an F1-score of 0.98 on the training dataset. This means that the model is able to correctly classify 97.9% of the training data points and has a good balance between precision and recall.

In same (Table 6 and Figure 5), the validation scores indicate that the model has achieved an accuracy of 0.931 and an F1-score of 0.624 on the validation dataset. This suggests that the model may be overfitting to the training data, as it is not performing as well on the validation data. The F1-score is particularly low, indicating that the model may be struggling with precision or recall on the validation set. Overall, the training scores suggest that the model has been trained successfully, but the lower validation scores suggest that the model may need further optimization or regularization to improve its performance on unseen data.

Confusion matrix

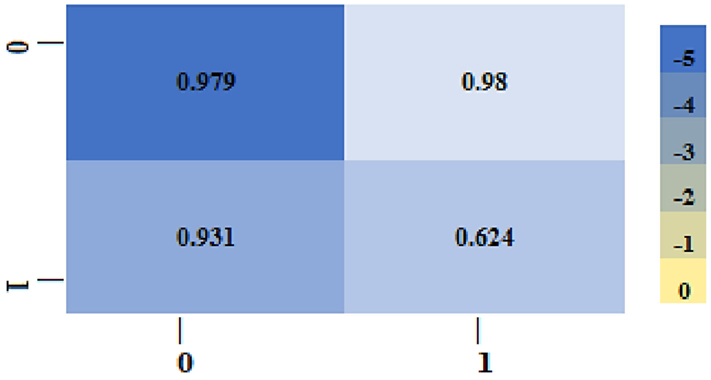


Figure 5 Logistic regression

Naive Bayes classifier

Training scores:	**Accuracy=0.967**	**F1-score=0.967**
Validation scores:	**Accuracy=0.925**	**F1-score=0.615**

Table 7 Training Scores of Naive Bayes Classifier

The information provided in Table 7 is about the performance of a Naive Bayes classifier on a training and validation dataset. The training scores indicate that the classifier has achieved an accuracy of 0.967 and an F1-score of 0.967 on the training dataset, which means that the classifier is able to accurately classify 96.7% of the training data points and has a good balance between precision and recall.

The validation scores indicate that the classifier has achieved an accuracy of 0.925 and an F1-score of 0.615 on the validation dataset, which means that the classifier is not performing as well on the validation data compared to the training data. This suggests that the model may be overfitting to the training data and may need further regularization or tuning to improve its performance on unseen data.

Overall, the Naive Bayes classifier results in Figure 6 seems to be performing reasonably well, but its performance on the validation dataset may need to be improved to ensure better generalization to unseen data.

Confusion matrix

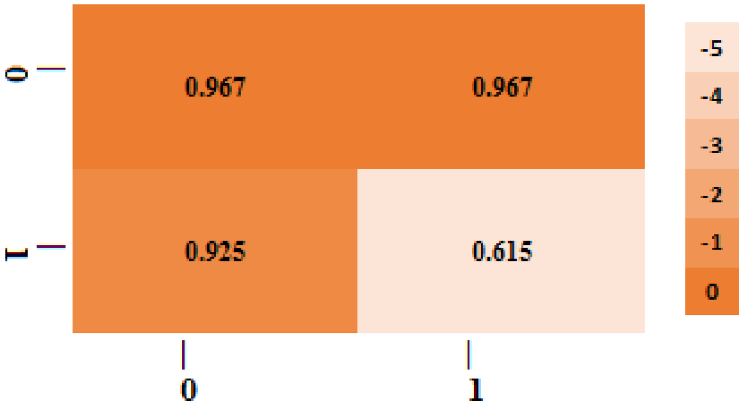


Figure 6 Naive Bayes classifier

Random forest classifier

Training scores:	**Accuracy=1.0**	**F1-score=1.0**
Validation scores:	**Accuracy=0.964**	**F1-score=0.707**

Table 8 Training Scores of Random Forest classifier.

The Table 8 shows the performance of a Random Forest classifier on a training and validation dataset. The training scores indicate that the classifier has achieved a perfect accuracy of 1.0 and a perfect F1-score of 1.0 on the training dataset, which means that the classifier is able to perfectly classify all the training data points and has a perfect balance between precision and recall. The validation scores indicate that the classifier has achieved an accuracy of 0.964 and an F1-score of 0.707 on the validation dataset, which means that the classifier is not able to perfectly classify the validation data points but is still able to achieve a relatively high accuracy of 96.4% and a moderate F1-score of 0.707. This suggests that the model may be overfitting slightly to the training data but is still able to generalize well to the validation data.

Overall, Figure 7 shows that the Random Forest classifier appears to be a good model for the given classification task, as it is able to achieve high accuracy and F1-score on both the training and validation datasets. However, further optimization or tuning may be required to improve its performance on unseen data or to prevent overfitting.

Confusion matrix

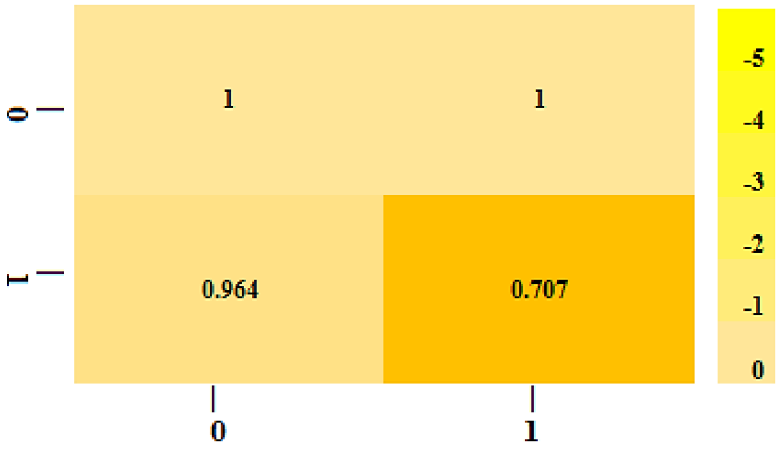


Figure 7 Random Forest classifier

Extreme gradient boosting classifier

Training scores:	**Accuracy=0.943**	**F1-score=0.941**
Validation scores:	**Accuracy=0.952**	**F1-score=0.639**

Table 9 Training Scores of Extreme gradient boosting classifier

The Table 9 provides the information about the performance of an Extreme Gradient Boosting (XGBoost) classifier on a training and validation dataset. The training scores indicate that the classifier has achieved an accuracy of 0.943 and an F1-score of 0.941 on the training dataset, which means that the classifier is able to accurately classify 94.3% of the training data points and has a good balance between precision and recall. The validation scores indicate that the classifier has achieved an accuracy of 0.952 and an F1-score of 0.639 on the validation dataset, which means that the classifier is performing relatively well on the validation data but not as well as on the training data. This suggests that the model may be overfitting to the training data and may need further regularization or tuning to improve its performance on unseen data.

Overall, according toresults in Figure 8, the XGBoost classifier appears to be a promising model, as it is able to achieve high accuracy and F1-score on the training dataset, and relatively good performance on the validation dataset. However, further optimization may be required to improve its generalization performance and prevent overfitting.

Confusion matrix

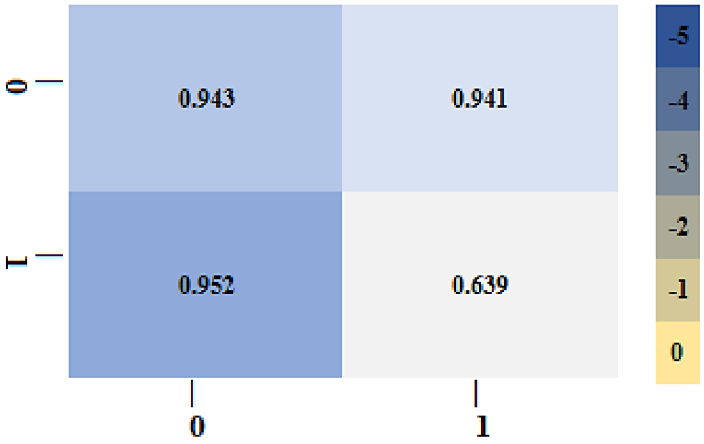


Figure 8 Extreme gradient boosting classifier

5.1 Hyperparameter Tuning

Among above 4 classifiers, random forest classifier delivered exceptionally good results. Xgboost stands as the second-best model. As we can notice, random forest classifier is suffering overfitting while xgboost classifier is suffering underfitting. With taking this into account, I applied various combination of hyperparameters for random forest and xgboost classifiers to reach to the optimal solution. Hyperparameter tuning is the process of selecting the optimal set of hyperparameters for a machine learning model in sentiment analysis. Hyperparameters are parameters that are not learned by the model during training but are instead set before training begins. These parameters can have a significant impact on the performance of the model and tuning them can help improve the model's accuracy and generalization performance. In the context of sentiment analysis, hyperparameters can include the number of layers and neurons in a neural network, the type of activation function used, the learning rate, the regularization strength, and many other parameters that affect how the model learns and makes predictions. Hyperparameter tuning typically involves testing different combinations of hyperparameters using a validation dataset to evaluate the performance of the model. This can be done manually or through automated methods such as grid search or random search. The goal is to find the optimal set of hyperparameters that results in the highest accuracy and F1-score on the test dataset, while avoiding overfitting to the training data.

5.1.1 Random Forest classifier

Training scores:	**Accuracy=0.999**	**F1-score=0.999**
Validation scores:	**Accuracy=0.962**	**F1-score=0.713**

Table 10 Training Scores of Random Forest classifier.

The results provided in Table 10 shows the performance of a Random Forest classifier on a training and validation dataset. The training scores indicate that the classifier has achieved a high accuracy of 0.999 and an F1-score of 0.999 on the training dataset, which means that the classifier is able to accurately classify almost all of the training data points with a high balance between precision and recall. The validation scores indicate that the classifier has achieved an accuracy of 0.962 and an F1-score of 0.713 on the validation dataset, which means that the classifier is not able to perfectly classify the validation data points but is still able to achieve a relatively high accuracy of 96.2% and a moderate F1-score of 0.713. This suggests that the model may be overfitting slightly to the training data but is still able to generalize well to the validation data.

Overall, the results shown in Figure 9 tells that the Random Forest classifier appears to be a good model for the given classification task, as it is able to achieve high accuracy and F1-score on both the training and validation datasets. However, further optimization or tuning may be required to improve its performance on unseen data or to prevent overfitting.

Confusion matrix

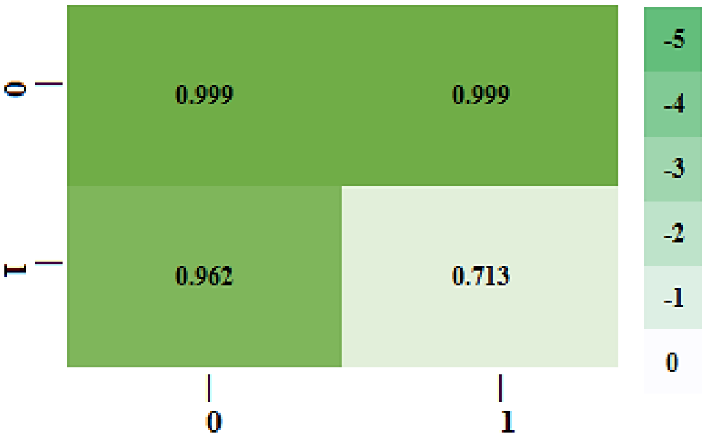


Figure 9 Random Forest classifier

5.1.2 Extreme gradient boosting classifier

Training scores:	**Accuracy=0.999**	**F1-score=0.999**
Validation scores:	**Accuracy=0.962**	**F1-score=0.701**

Table 11 Training Scores of Extreme gradient boosting classifier

The Table 11 provided is about the performance of an Extreme Gradient Boosting (XGBoost) classifier on a training and validation dataset. The training scores indicate that the classifier has achieved a very high accuracy of 0.999 and an F1-score of 0.999 on the training dataset, which means that the classifier is able to accurately classify almost all of the training data points with a high balance between precision and recall.

The validation scores in Figure 10 indicate that the classifier has achieved an accuracy of 0.962 and an F1-score of 0.701 on the validation dataset, which means that the classifier is not able to perfectly classify the validation data points but is still able to achieve a relatively high accuracy of 96.2% and a moderate F1-score of 0.701. This suggests that the model may be overfitting slightly to the training data but is still able to generalize well to the validation data.

Confusion matrix

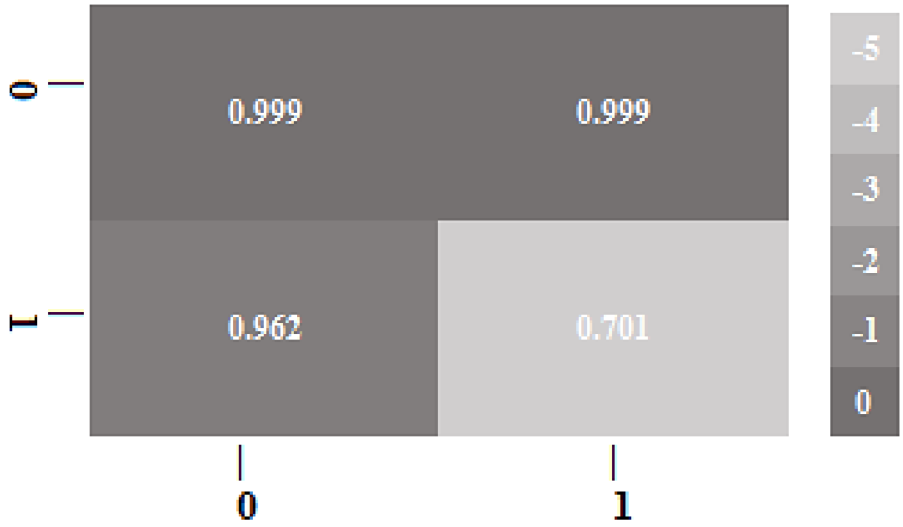


Figure 10 Result Matrix

5.2 Practical Implementations of Sentiment Analysis

Sentiment analysis is a type of natural language processing (NLP) that is used to detect and analyse the sentiment of a text phrase. It is a powerful tool that can be applied to gain valuable insights into the opinions, beliefs, and emotions of the readers of a text document. It can be used to identify customer sentiment toward a product or service, and to evaluate customer feedback. Moreover, it can also be used to detect customer trends and to help in decision-making processes. The application of sentiment analysis in machine learning involves the use of algorithms to detect and analyse the sentiment of a text document. This is done by training a machine learning model to recognize and classify the sentiment of a text document. During this process, features are extracted from the document, such as the words used and their context. After the features are extracted, the model is then trained to recognize patterns in the data and to classify the sentiment of a text document.

In addition to sentiment analysis, machine learning can be applied to other aspects of a text document, such as its grammar and syntax. For instance, machine learning can be used to detect spelling and grammar errors in a text document. This can prove to be extremely helpful, as it can improve the quality of the text document, making it more accurate and easier to understand. Furthermore, machine learning can also be used to identify and analyse the sentiment of a text document, allowing marketers to understand customer attitudes, opinions, and beliefs. This knowledge can then be used to inform marketing strategies, customer service approaches, and product development, thus improving the customer experience. All in all, sentiment analysis is an invaluable tool that can be used in machine learning to gain insights into customer attitudes, opinions, and beliefs. This knowledge can then be used to inform product and marketing initiatives, allowing businesses to create customer-centric products and services that are tailored to customer needs. Additionally, sentiment analysis can be used to gain an understanding of how customers perceive a product or service, and to determine which aspects of the product or service are most valued by customers. This knowledge can then be used to inform product development and marketing initiatives, thus improving customer satisfaction.

I believe that my work is different to the existing systems based on sentiment analysis as I can focus our work on small business. Currently this technique is being utilized by big corporations such as Apple, Zara, Amazon to get customer insights that they use to optimize their product and get a precise opinion of the customer about the product. But we can create a system that can be used by small business such as drop shipping, local brands to get a perception of the peoples in the specific region so they can tailor their products and services according to the trends. For example, if there is a small tour and travel company in a country, then the business owners can get the data of what are the desired locations, most frequent visited places by the people in the region. One more example can be the advertising companies as they can get the general perception of a product in the market and then plan their marketing strategy accordingly. This can be applied to almost any type of business. This will not only help the small business to grow but also will help the country in overall business revenue. Moreover, insights can be used to refine customer service strategies, such as by responding to customer feedback in an appropriate manner.

For future work on sentiment analysis, it is required to assign Twitter data sentiment polarity in real time. To achieve this, an implementation plan for the same data processing technique on the cloud that improves sentiment analysis performance will be developed utilizing principles of Natural Language Processing. This is accomplished by creating nodes.
Hadoop is a cloud data platform that allows us to store data on the cloud using HDFS (Hadoop File System) and the Map-reduce idea to spread the technique for data processing on the cloud to load and process big data sets and to do real-time sentiment analysis for the linguistic data. This will contribute to cloud-based real-time sentiment analysis environment and would enable corporate users to retrieve real-time sentiment information for their target market linguistic data.

6 Conclusion

In conclusion, this thesis tried to delved into the intricate relationship between technology and user sentiments on social media platforms, specially Twitter, with the goal of interpreting the sentiments concealed inside millions of tweets posted everyday. By analysing sentiments of twitter users, we can understand a lot of patterns about the user perspective about literally anything. The usecases of this technology is so vast that it can be used in almost any industry in different ways. Marketing companies can use it to target a product to a specific user group, while product based companies can use sentiment analysis to get user opinion about thier product. These are very few examples of usecases of this technology.

Understanding sentiments of hundreds of thousand individuals to gain a valuable insight is not an easy task so we decided to design and develop a sentiment analyser for twitter dataset with data mining approach. Our first objective was to collect the required data to begin with. So we used Twitter APIs to gather the tweets(data). As mentioned in Chapter 4.1, the collected data was divided into 2 sets for the purpose of training and testing. The data processing technique, as mentioned in chapter 4.2, the data was cleaned to make it available for model training. This step included removing any stop words, punctuations, symbols, fixing the spelling mistakes, converting imojies to text and much more. The final data was ready to be processed with different classifiers such as Naive Bayes and Random forest classifiers. We utlized multiple classifiers for as much accuracy as possible. The details of the result outcome were discussed in Chapter 5. Here we analysed the results on various factors such as Acccurecy and F1 score for both Training and Validation scores. According to the results, Random Forest and Extreme Gradient Boosting Classifier outperformed their counterparts in terms of classification accuracy, giving the exceptionally well accuracy scores in all paramteres. So we decided to use the results based on these 2 classifiers.

While the results came out to be very acceptable, it would be an interesting area of future research to increase the stability and accuracy of the model. The use cases of Sentiment Analysis goes beyond our current thinking and its always a learning process to make this technology better than before for companies and consumers.

7 References

1. Pak, A. and Paroubek, P., 2010, May. Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).
2. Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12), p.2009.
3. Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R.J., 2011, June. Sentiment analysis of twitter data. In Proceedings of the workshop on language in social media (LSM 2011) (pp. 30-38).
4. Liu, B., 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), pp.1-167.
5. Saif, H., He, Y. and Alani, H., 2012. Semantic sentiment analysis of twitter. In The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11 (pp. 508-524). Springer Berlin Heidelberg.
6. Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M., 2011. Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), pp.267-307.
7. Duan, L., Gao, F., & Li, J. (2012). Sentiment analysis with big data. IEEE transactions on knowledge and data engineering, 24(4), 617-627.
8. Pak, A., & Paroubek, P. (2011). Cross-lingual sentiment analysis: A case study on english and french tweets. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (pp. 56-64).
9. Mohammad, S.M. and Turney, P.D., 2013. Crowdsourcing a word–emotion association lexicon. Computational intelligence, 29(3), pp.436-465.
10. Tumasjan, A., Sprenger, T., Sandner, P. and Welpe, I., 2010, May. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the international AAAI conference on web and social media (Vol. 4, No. 1, pp. 178-185).
11. Read, J. and Carroll, J., 2009, November. Weakly supervised techniques for domain-independent sentiment classification. In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (pp. 45-52).
12. Gao, J., Zheng, Y., & Huang, Y. (2012). Microblog sentiment classification using a coupled generative/discriminative approach. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 169-173).

13. McKeown, K., Agarwal, A. and Biadsy, F., 2009. Contextual Phrase-Level Polarity Analysis using Lexical Affect Scoring and Syntactic N-grams.
14. Kouloumpis, E., Wilson, T. and Moore, J., 2011. Twitter sentiment analysis: The good the bad and the omg!. In Proceedings of the international AAAI conference on web and social media (Vol. 5, No. 1, pp. 538-541).
15. Mohammad, S. and Turney, P., 2010, June. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text (pp. 26-34).
16. Bifet, A. and Frank, E., 2010. Sentiment knowledge discovery in twitter streaming data. In Discovery Science: 13th International Conference, DS 2010, Canberra, Australia, October 6-8, 2010. Proceedings 13 (pp. 1-15). Springer Berlin Heidelberg.
17. Bermingham, A. and Smeaton, A.F., 2010, October. Classifying sentiment in microblogs: is brevity an advantage?. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 1833-1836).
18. Barbosa, L. and Feng, J., 2010, August. Robust sentiment detection on twitter from biased and noisy data. In Coling 2010: Posters (pp. 36-44).
19. Chakraborty, K., Bhattacharyya, S. and Bag, R., 2020. A survey of sentiment analysis from social media data. IEEE Transactions on Computational Social Systems, 7(2), pp.450-464.
20. Kharde, V. and Sonawane, P., 2016. Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.
21. Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S., 2012, July. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 system demonstrations (pp. 115-120).
22. Edgar, T. and Manz, D., 2017. Research methods for cyber security. Syngress.
23. Parveen, H. and Pandey, S., 2016, July. Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. In 2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT) (pp. 416-419). IEEE.
24. Sharma, P. and Moh, T.S., 2016, December. Prediction of Indian election using sentiment analysis on Hindi Twitter. In 2016 IEEE international conference on big data (big data) (pp. 1966-1971). IEEE.

25. Jain, A.P. and Katkar, V.D., 2015, December. Sentiments analysis of Twitter data using data mining. In 2015 International Conference on Information Processing (ICIP) (pp. 807-810). IEEE.
26. Kharde, V. and Sonawane, P., 2016. Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.
27. Hovy, E. H. (2014). What are Sentiment, Affect, and Emotion? Applying the Methodology of Michael Zock to Sentiment Analysis. Text, Speech and Language Technology, 13–24. doi:10.1007/978-3-319-08043-7_2
28. Pozzi, F.A., Fersini, E., Messina, E. and Liu, B., 2016. Sentiment analysis in social networks. Morgan Kaufmann.
29. Wankhade, M., Rao, A.C.S. and Kulkarni, C., 2022. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), pp.5731-5780.
30. Sudhir, P. and Suresh, V.D., 2021. Comparative study of various approaches, applications and classifiers for sentiment analysis. Global Transitions Proceedings, 2(2), pp.205-211.
31. Kraaijeveld, O. and De Smedt, J., 2020. The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. Journal of International Financial Markets, Institutions and Money, 65, p.101188.
32. Younis, E., 2015. Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study. International Journal of Computer Applications, 112, pp. 44-48.
33. Stieglitz, S. and Dang-Xuan, L., 2013. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. Journal of management information systems, 29(4), pp.217-248.
34. Ren, Q., Cheng, H. and Han, H., 2017, March. Research on machine learning framework based on random forest algorithm. In AIP conference proceedings (Vol. 1820, No. 1). AIP Publishing.
35. Rambocas, M. and Gama, J., 2013. Marketing research: The role of sentiment analysis (No. 489). Universidade do Porto, Faculdade de Economia do Porto.
36. Gladence, L.M., Karthi, M. and Anu, V.M., 2015. A statistical comparison of logistic regression and different Bayes classification methods for machine learning. ARPN Journal of Engineering and Applied Sciences, 10(14), pp.5947-5953.

37. Mahesh, B., 2020. Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), pp.381-386.
38. Wang, S., Jiang, L. and Li, C., 2015. Adapting naive Bayes tree for text classification. Knowledge and Information Systems, 44, pp.77-89.
39. Singh, A., Thakur, N. and Sharma, A., 2016, March. A review of supervised machine learning algorithms. In 2016 3rd international conference on computing for sustainable global development (INDIACom) (pp. 1310-1315). Ieee.
40. Dos Santos, C. and Gatti, M., 2014, August. Deep convolutional neural networks for sentiment analysis of short texts. In Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers (pp. 69-78).

8 List of pictures, tables, graphs, and abbreviations

8.1 List of pictures

8.2 List of tables

Appendix

List of Supplements…