A Comparative Study of Opinion Mining Utilizing Twitter Data for Depression Detection

A thesis

Submitted in partial fulfillment of the requirement for the Degree of

Master of Technology in Computer Technology

Of

Jadavpur University

By

Sanghita Bhaumik

Registration No: 149862 of 2019-2020

Examination Roll No: M6TCT22029

Under the Guidance of

Prof. Diganta Saha

Department of Computer Science and Engineering

Jadavpur University

2022

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

CERTIFICATE OF RECOMMENDATION

This is to certify that the dissertation titled **A Comparative Study of Opinion Mining Utilizing Twitter Data for Depression Detection** was completed by **Sanghita Bhaumik** , University Roll No: 001910504028, Examination Roll Number : M6TCT22029, University Registration No: 149862 of 2019-2020, under the guidance and supervision of **Prof. Diganta Saha,** Department of Computer Science and Technology, Jadavpur University. The findings of the research detailed in the thesis have not been incorporated into any other work submitted for the purpose of earning a degree at any other academic institution.

_____

Prof. Diganta Saha

Department of Computer Science & Engineering

Jadavpur University

COUNTERSIGNED BY

_____

Prof (Dr.) Anupam Sinha
Head of The Department
Department of Computer Science and Engineering
Jadavpur University

COUNTERSIGNED BY

_____

Prof. Chandan Mazumdar
Dean, FET
Department of Computer Science and Engineering
adavpur University

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

CERTIFICATE OF APPROVAL

This is to certify that the thesis entitled **A Comparative Study of Opinion Mining Utilizing Twitter Data for Depression Detection** is a bonafide record of work carried out by **Sanghita Bhaumik** in partial fulfillment of the requirements for the award of the degree **Master of Technology** in **Department of Computer Science and Engineering, Jadavpur University** during the period of **June 2021 to May 2022** (5^{th} & 6^{th} Semester). It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn there in but approve the thesis only for the purpose for which it has been submitted.

_____

Signature of Examiner

Date:

_____

Signature of Supervisor

Date:

DECLARATION

I certify that,

(a) The work **A Comparative Study of Opinion Mining Utilizing Twitter Data for Depression Detection** contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary

ACKNOWLEDGEMENT

It is with great pleasure and pride that I attribute the success of my project to the guidance and support of many of whom I am grateful.
I sincerely express my gratitude to our Head of the Department **Prof. (Dr) Anupam Sinha** and to all our professors who have taught us throughout the course and always supported us and guided us to the right direction.
I would want to specially thank my supervisor **Prof. Diganta Saha**, who has been helpful in every possible way with his expert guidance, valuable comments, encouragement and constructive criticism right from the beginning of my work. His supervision has been extremely beneficial for me at every stage of my work and has helped me to conclude this thesis in an appropriate way.
I would also want to thank my family and friends for their constant support, love and sacrifice. My sincere thanks to my friends **Shankha Banerjee**, **Rebhu Roy** for their immense support and guidance to complete the project work.
I am ever grateful to the almighty God for showering his blessings on me.

Abstract

The use of social media has a significant impact on human life. People use this platform for a variety of purposes, including entertainment, information retrieval, news, commerce, and many others. It is possible to identify whether data is good, negative, or neutral using Sentiment Analysis, which is a type of natural language processing technology. Sentiments are acquired from a variety of sources, including conversations, blogs, tweets, and other social media. The tremendous rise of social media has aided in the development of this topic as a research hotspot in recent years. Nowadays, every corporate organization seeks to understand the public's perception of their products by conducting a variety of surveys in order to improve their operations in the future. This type of platform allows ordinary people to express their emotions on many themes with multiple users through the usage of social media networks. Multiple audio, video, and text formats are available for us to share our daily thoughts, comments on current events, as well as our political views on a variety of topics. Although Social Media provides a variety of connection options, text mode remains the most effective method of communicating our thoughts to others. People can express their feelings about a variety of subjects through the use of text messages known as "tweets." As a result, data from the Twitter application has been selected for analysis. In this work, we will investigate a complete study of sentiments from Twitter, where people express their feelings in the form of tweets and provide helpful information gleaned from the data mining process. It is our intention to look at a comprehensive technique that has been implemented in Sentiment Analysis where Naïve Bayes Classification technique has been adopted to train and test the data to predict depression level of each user. Depression score of every users will be calculated and the final opinion will be delivered based on the score. This concept will be beneficial for medical ground.

Contents

List of Figures

List of Tables

Chapter -1

INTRODUCTION

"Sentiment" and "Opinion" – two important aspects of human nature. Our thoughts and emotions have a vital role in our daily life. The decisions we make are closely related to the feelings and attitudes of our society. Sentiment Analysis and Opinion Mining are two processes to classify and investigate the behavior and approach of the customers in regards to the brand, product events, and customer services. People want to share their experiences, thoughts, opinions, sentiments, and preferences according to their comprehension and observation about the services. Their point of view or perception may be good, negative or neutral.

With the major advancement of web-based technology, a good number of people are expressing their views through this web based platform. Ecommerce based websites are the examples of such a source which encourages people to express their reviews towards their products through some basic online rating manner as a survey. Several users can share their thoughts related to multiple products of different E-Commerce sectors which help any business-oriented organizations to understand the users view and liking and disliking about the product in terms of polling and text paragraph. Comments, reviews play an important role to determine the satisfaction level of a given population towards a product service. It helps to predict the sentiment of a wide variety of people on a particular event of interest like the review of a movie, their opinion on various topic roaming around the world which make the event more popular and most trending topic .

The feelings stated in the tweets provide an indication of the users' deeper emotions. As a result, feelings stated in tweets that have a negative connotation may suggest a negative emotion. The extraction of emotions and views from tweets made by users is known as sentiment analysis of Twitter data. Many people who suffer from stress find it simpler to express their sentiments on an internet platform since they can do so more easily. So, if they are warned ahead of time, they will be able to overcome their mental health issues and stress. One of the most well-known mental health illnesses and a key concern for medical and mental health professionals is depression. It has gained a lot of attention in recent years because of its importance in a variety of applications because it analyses such emotions and opinions analytically and provides a robust technology to a variety of problem domains ranging from business intelligence and fake news detection to analysing a user's emotional state, depression, or stress. The tweets contain precise keywords and sentences that are examined to identify people who are at a high risk of being a victim of cyber bullying. These data are essential for sentiment analysis. F.Neri, C.Aliprandi, F.Capeci, M.Cuadros, T.By, (2012)

1.1 Sentiment Analysis and Opinion Mining

Sentiment Analysis can be defined as the automatic process of extracting emotions from the user’s written thoughts by processing unstructured information and preparing model to extract knowledge from it. It is the process of categorizing positive, negative, or neutral opinions expressed in papers or statements on the web as positive, negative, or neutral. The goal of sentiment analysis is to extract the sentiment behind a user's remark and show the user's interest, preference, and ideas on a specific topic. Because the majority of user-generated messages on microblogging services are textual, detecting their sentiments has become a major concern. Sentiment classification, which handled the topic as a text classification problem, was the first

step in the field's research. The world's textual information can be divided into two primary categories: facts and views. Facts are objective statements regarding the world's entities and happenings. Opinions are personal declarations that represent people's feelings or perceptions about things and occurrences. Opinion mining and sentiment analysis are two terms that refer to the same topic of study. Sentiment analysis, also known as opinion mining, is the study of people's attitudes, sentiments, assessments, appraisals, and feelings about things like products, services, organisations, individuals, issues, events, and subjects, as well as their attributes. Despite the fact that the terms sentiment analysis and opinion mining were coined in the fields of linguistics and natural language processing, little research had been done prior to the year 2000. However, as a major study topic, researchers are now focusing on sentiment analysis in the domain of natural language processing, because it has a wide variety of commercial applications in practically every discipline. In recent years, a variety of methodologies have been described; some rely on machine learning approaches using supervised, unsupervised, or semi-supervised learning, while others may employ semantic-based approaches. Furthermore, a few hybrid approaches based on techniques from several fields may be applied. Subjectivity and sentiment categorization, sentiment lexicon generation, opinion spam detection, and review quality are all key tasks in sentiment analysis.

1.1.1 Why Sentiment Analysis

Several e-commerce sites' business is influenced indirectly by online opinions. Those websites promote their items, and web visitors read product reviews before making a purchase. Many businesses use opinion mining software to track customer feedback on things they sell online.

Opinion mining is a fantastic approach to stay on top of a variety of company trends such as deal administration, status management, and advertising. The opinion of customers is also used to forecast patterns.

1.2 Different Approaches for sentiment analysis

There are many approaches are adopted for sentiment analysis which works on linguistic data and those depends on the nature of the data and the platform where working for. Most of the researches based on Sentiment Analysis deploy the concept of lexicon based and machine learning techniques. Most of the researches based on Sentiment Analysis employ the concept of lexicon based and machine learning techniques.

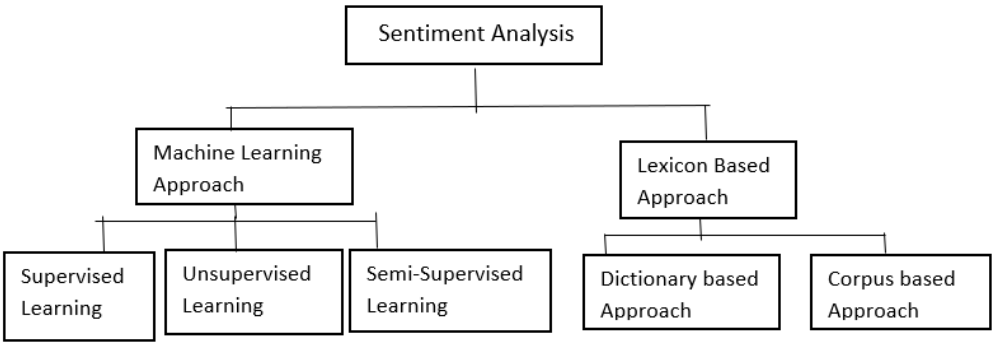


Fig 1: Sentiment Analysis Technique

There are many approaches are adopted for sentiment analysis which works on linguistic data and those depends on the nature of the data and the platform where working for. Machine learning techniques use machine learning algorithms to regulate data processing and classify linguistic data by and represent it in a vector form, Po-Wei Liang and colleagues (2013), whereas Lexicon based classifies linguistic data based on dictionary lookup database which computes sentence level polarity with the help of lexicon databases like WordNet, SentiWordnet.

1.2.1 Machine Learning Approach

In the literature on sentiment analysis, the use of machine learning is common. The words in the phrase are treated as vectors in this technique, and different machine learning algorithms such as Naïve Bayes analyse them. The data is then trained in this manner, and it can then be used in machine learning algorithms. In this research work Naïve Bayes algorithm has been used to implement Sentiment Analysis. The details will be discussed in Chapter 5.

1.2.2 Lexicon Based Approach

Using lexical datasets such as SentiWordNet and WordNet, the lexicon-based technique predicts feelings. It calculates a score for each word in the sentence or text and annotates it using the features available in the lexicon database. It calculates text polarity using a collection of words, each of which is assigned a weight, and extracts information that helps to determine the text's overall attitudes. Lexicon-based technique predicts feelings with the help of lexical datasets such as WordNet. It calculates a score for each word in the sentence or text and annotates it using the features available in the lexicon database. It calculates text polarity using a collection of words, each of which is assigned a weight, and extracts information that helps to determine the text's overall attitudes.

1.3 Different Levels of Sentiment Analysis

In general, three layers of sentiment analysis have been studied: level of the document Entity and aspect levels at the sentence level Level of Documentation At this level, the aim is to determine whether the entire opinion document expresses a favourable or negative mood. The system, for example, evaluates if a product review communicates an overall good or negative judgement about the product. Document-level sentiment classification is the term for this task. This level of analysis implies that each document conveys a single entity's viewpoint (e.g., a

single product). As a result, it cannot be used in documents that assess or compare several entities. According to current research, even in a negative document, there is more than 40% positive content. Level of Sentence At this level, the task is to the sentences, determining if each one reflected a good, negative, or neutral viewpoint. In most cases, neutral means that you don't have an opinion. Subjectivity classification, which distinguishes sentences that communicate objective information from sentences that express subjective thoughts and opinions, is closely related to this level of analysis. The document is nothing more than a collection of sentences put together, however sentence level sentiment analysis is far more accurate than document level sentiment analysis.

Levels of Entity and Aspect It produces substantially better results than sentiment analysis at the document and sentence levels. The document level and sentence level analysis both fail to reveal what people liked and disliked. Finer-grained analysis is performed at the aspect level. Aspect level examines the opinion itself rather than language constructs (documents, paragraphs, sentences, clauses, or phrases). It is founded on the notion that an opinion is made up of two parts: a feeling (positive or negative) and a target (of opinion).

1.4 Role of Social media in Sentiment Analysis

The Internet is a vast virtual area where people may express and share their personal views, influencing every part of life and having ramifications for marketing and communication. Consumer preferences are influenced by social media by modifying their attitudes and behavior. Monitoring consumer loyalty and attitude towards brands or products via social media is a useful technique to measure client loyalty and keep track of their mood. The next natural marketing domain is social media. Facebook currently reigns supreme in the digital marketing world, closely followed by Twitter. In order to discover the sentiment from a large scale of people social network plays essential role to collect data. Social network website is a recognized web-based

platform where user-generated opinions are data abounds. Social network websites usually contain a great scope of topics, mostly related to the big events.

In today's world Microblogging has gained major popularity among Internet users. Millions of messages are delivering daily through several microblogging websites. Authors of those messages share their thoughts related to several topics and discussed their opinion about current topics. The free format of messages as well as accessibility of microblogging platforms make this platform popular among users and they shifted from traditional communication to modern blogging services. Microblogging web-sites become valuable sources of people's opinions and sentiments because of a huge number of people's post related to social, political and personal posts. Determining what others think has always been an important component of our information gathering process. New opportunities have arisen as a result of the rising availability of opinion-rich sources such as social networking sites or personal blogs, as individuals may now diligently use the information to study and understand the perspectives of others.

1.4.1 Twitter as the corpus for Sentiment Analysis

Being a popular microblogging website a million of users share their thoughts over twitter as tweet. So, twitter is a good resource for corpus collection for Sentiment analysis. Twitter is the collection of short messages of personal thoughts and it vary from public comments. Twitter generates huge data for extracting information and need ingredients of automatic classification to handle those large data. Tweets are unambiguous text messages that are up to a maximum of 140 characters. A million of people can connect with each other by the use of Twitter through their mobile phones, laptop. Twitter interface allows user to post several short messages which are accessible by other twitter users and can share their thoughts on the topics. Twitter contains a variety of text posts and rapidly grows every day. According to the statistics, users that are active monthly range at about 316 million, and

on average 500 million tweets are sent daily. Twitter users are interested in. Pang et al.in(2008) outline for different cases for expressing their opinions via their tweets which is closely related to real world situations, like product/service reviews on restaurants, electronics goods, hotels. The unstructured text format of Twitter helps to put up challenges for the classifiers to analyze the data.
Twitter is considered as a renowned corpus for analysis purpose for the following reasons:
i) People shares their several thoughts through this platform so it is a strong source for data collection.

ii) Tweets are collections of texts along with several emoticons and special characters, so it provides a huge number of information about opinion analysis.

iii) Twitter is handled by several countries people so it is the collection of several languages which helps to analyze linguistic features.

iv) All the texts are intended to positive, negative or neutral objectives. In the era of rapid growth of technology, the sentiment analysis become a popular research field where Twitter data is the pillar of this study. Twitter has grown to be such a popular tool for expressing one's opinions that researchers have begun to use it as a valuable source of data when investigating mental health issues. As a result, sentiment analysis can benefit from the textual information shared on Twitter.
Twitter data is considered as a data set of sentiment analysis in my project because of its several features, popularity and unstructured format of text representation.

Sentiment Analysis is all about the concept of linguistic form of human's view, thought and emotions. Billions of people have their own style of communicative way which contains separate sentiments and it is tedious to interpret by machine. Here the concept of Natural Language Processing plays the vital role to process the millions of texts and make them machine understandable along with machine learning

techniques. Natural Language Processing process data with the help of Natural Language Toolkit. NLTK is a Python-based set of libraries and tools for symbolic and statistical natural language processing (NLP) for the English language.

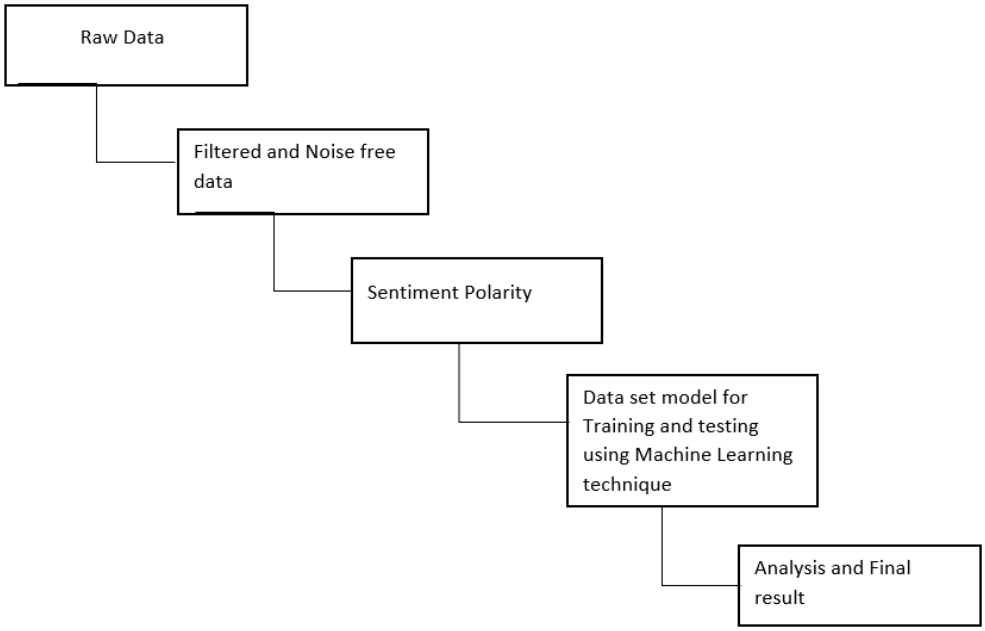


Fig 2 Structure of Sentiment Analysis

1.5 Thesis Goal

The purpose of this thesis is to suggest a method for detecting the level of depression among Twitter users. It combines Natural Language Processing and Naïve Bayes classification to create a system that can identify at-risk tweets and thus at-risk individuals diagnosed with a particular level of depression given a sequence of tweets from a user. To calculate the accuracy and depression score of individual user and the ultimate opinion to set based on the score of the community of twitter users.

Chapter 2

Review of Literature

Sentiment analysis, also known as opinion mining, is the study of people's attitudes, sentiments, assessments, appraisals, and feelings about things like products, services, organisations, individuals, issues, events, and subjects, as well as their attributes. Microblogging has grown in popularity among Internet users in recent years. Every day, millions of messages surface on prominent microblogging websites such as Twitter, Tumblr, and Facebook. The authors of those messages write about their lives, express their thoughts on many themes, and debate current events. Internet users are shifting from traditional communication tools (such as traditional blogs or mailing lists) to microblogging services due to the open structure of messages and easy accessibility to microblogging platforms.
Microblogging websites are becoming valuable sources of people's thoughts and sentiments as more users post about items and services they use, or express their political and religious beliefs. Such information can be useful in marketing and social studies.

2.1 Twitter as the corpus of Sentiment Analysis

Opinion mining and sentiment analysis have become popular study topics as the number of blogs and social networks has grown. In this paper, a fairly wide summary of existing work was presented (Pang and Lee, 2008). The authors describe existing methodologies and approaches for opinion-oriented information retrieval in their survey.

The authors of (Yang et al., 2007),use web-blogs to create a corpus for sentiment analysis, and they use emotion icons assigned to blog entries as indications of users' mood.

J. Read in (Read,2005) used emoticons like ":-)" and ":- (" to create a sentiment classification training set. The author gathered emoticon-containing texts from Usenet newsgroups for this project. "Positive" (texts with cheerful emoticons) and "negative" (texts with sad or angry emoticons) samples were taken from the dataset. On the test set, emoticon strained classifiers such as SVM and Naïve Bayes were able to achieve up to 70% accuracy.

The authors of (Go et al.,2009) research used Twitter to obtain training data before doing a sentiment search. The strategy is similar to that of J. Read in (Read,2005). The authors create corpora by extracting "positive" and "negative" samples from emoticons and then applying various classifiers. The Naïve Bayes classifier using a mutual information measure for feature selection produced the best results. On their test set, the authors were able to get up to 81 percent accuracy. With three classes ("negative," "positive," and "neutral"), however, the technique performed poorly.

(Bifet and Frank,2010) used Twitter streaming data supplied by the Firehouse API, which delivered all messages from all users that are publicly available in real-time. They tried out multi nominal naive Bayes, stochastic gradient descent, and the Hoeding tree. They concluded that the SGD-based model, when employed with an acceptable learning rate, outperformed the others.

(Pak and Paroubek,2010) suggested a technique for classifying tweets into objective, positive, and negative categories. They built a Twitter corpus by collecting tweets via the Twitter API and automatically annotating them with emoticons. Using that corpus, they created a sentiment classifier based on the multi nominal Naive Bayes approach, which employs characteristics such as N-grams and POS tags. The training set they utilized was inefficient because it only included tweets with emoticons.

Turney et al. (2002) employed the bag-of-words method for sentiment analysis, which ignores word relationships and represents a document as a collection of words. To determine the sentiment for the entire document, the sentiments of each word were identified and those values were combined with some aggregation methods.

Alexander Pak and Patrick Paroubek(2010) in their research paper have focused on leveraging Twitter, the most popular microblogging network, for sentiment analysis in this article. They demonstrate how to gather a corpus automatically for sentiment analysis and opinion mining. They conduct linguistic analysis of the corpus and provide explanations for the occurrences they observe. They use the corpus to create a sentiment classifier that can assess whether a document is good, negative, or neutral. They gathered 300,000 text postings from Twitter, which were automatically divided into three categories: happy emotions, negative emotions, and no feelings. They analyse the corpus using statistical linguistic analysis. They use the corpora to create a microblogging sentiment classification system. They gathered a corpus of text posts using the Twitter API.

The relationship between subjectivity detection and polarity classification is investigated by Bo Pang and Lillian Lee. Subjectivity identification allows reviews to be compressed into shorter extracts while retaining polarity information at a level comparable to the entire review. The subjective extract is found to be a more effective input than the originating document when applying the Naïve Bayes polarity classifier. The study demonstrates how the minimum-cut methodology leads to the construction of an efficient sentiment analysis system. Contextual information can lead to a statistically significant improvement in polarity classification accuracy using this methodology.

2.2 Sentiment Analysis using machine learning technique

In the research paper of Gokulakrishnan, Balakrishnan, et al (2012)The study is broken into two parts: Opinion Mining and Sentiment Analysis on a Twitter Data Stream and Opinion Mining and Sentiment Analysis on a Twitter Data Stream. Initially, the preprocessed tweets were classified as neutral, polar, or irrelevant. The polar tweets are then sorted into positive and negative categories. The classifiers utilised in this two-step division didn't have to worry about neutral data, and the accuracy was also increased. Naïve Bayes classification was adopted to show the better accuracy value of the analysis

In the research paper of G.Gautam, D.Yadav(2014) have suggested a set of machine-learning approaches based on semantic analysis to categorise sentences and reviews of various products using WordNet for improved accuracy based on Twitter data.

In the research paper of Goel, Ankur, Jyoti Gautam, and Sitesh Kumar, the authors have discussed different classifiers, such as Naive Bayes, Support Vector Machine, and others, are used to verify the correctness of the classification procedure with selected feature vectors in the area of electronic products.

The work is separated into two parts in Opinion Mining and Sentiment Analysis on a Twitter Data Stream. Initially, the preprocessed tweets were classified as neutral, polar, or irrelevant. The polar tweets are then sorted into positive and negative categories. The classifiers utilised in this two-step division didn't have to worry about neutral data, and the accuracy was also increased. Naive Bayes and Random Forest were the machine learning classifiers used in this study. Bayesian models had the highest accuracy, followed by Random Forest.

2.3 Sentiment Analysis for generating depression level

Other authors have discovered the hypothesis that sad persons are more self-reflective and tend to focus on themselves and talk about themselves frequently even during a discussion, as proposed by Ireland and Mehl in 2014. Singular first-person pronouns were analysed in text transcripts using this method, and it was discovered that there was a link between more frequent usage of singular first-person pronouns and depression. The most important finding was that this effect occurred across all demographics, including age and gender. However, the use of singular first-person pronouns could only indicate to a limited extent that the individual was depressed, i.e. the sentence would be negative polarity, but the amount of depression could not be determined.

Singh and Wang concentrated on predicting depression from a Twitter user's tweets. They constructed their own dataset by extracting tweets from various Twitter pages and then labelling them using the polarity score derived from the Text blob Python tool. The researchers then built a number of deep learning models, including RNN, CNN, and GRU, which were used to make predictions on the dataset. For each model, they looked at the effects of character-based vs. word-based models, as well as pre-trained vs. learnt embeddings. As a result, they discovered that the word-based GRU, which had 98 percent accuracy, and the word-based CNN, which had 97 percent accuracy, were the models that performed the best.

In a major amount of the research that has been done so far for Depression, one common theme has been that the researchers used material from limited sources, such as internet gatherings, and did not use every word that was provided in the content.

This research paper will explore a proper method to calculate the depression score of each user with the help of naïve bayes classifier along with the Natural Language Processing and NLTK tool to get more accurate and real value. Based on the depression value multiple users will be categorized and set of opinions will be suggested to them.

Chapter 3

Background Study

Sentiment Analysis is the concept of computationally recognising and categorising opinions expressed in a piece of text, especially to determine whether the writer's attitude toward a given topic, product, etc. is favorable, negative, or neutral.

Opinions have a crucial role in practically all human activities since they shape our actions. We seek out the opinions of others whenever we need to make a decision. Businesses and organisations in the real world are continuously looking for consumer or public feedback on their products and services. Individual customers also want to know what other people think about a product before buying it, and what other people think about political candidates before voting in a political election. When a person wanted advice in the past, he or she turned to friends and family. When a company or organisation sought public or customer feedback, they conducted surveys, polls, and focus groups. Obtaining public and customer feedback has long been a lucrative business.

Sentiment analysis, often known as opinion mining, is a natural language processing (NLP) technique for determining the emotional tone of a body of text. This is a common method for businesses to determine and categorise customer opinions about a product, service, or concept. It entails mining text for sentiment and subjective information using data mining, machine learning (ML), and artificial intelligence (AI).

Sentiment analysis as well as opinion mining, is the study of people's feelings, sentiments, assessments, attitudes, and emotions regarding things like products, services, organisations, individuals, situations, events, themes, and their attributes. It encompasses a significant issue area. Sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis,

emotion analysis, review mining, and more terms and tasks are used. They are currently all grouped together under the heading of sentiment analysis or opinion mining. While sentiment analysis is more generally used in industry, both sentiment analysis and opinion mining are regularly used in academia. Regardless, they all belong to the same academic discipline. Nasukawa and Yi(2003) may have coined the term sentiment analysis,and the term Opinion mining first appeared in Dave et al(2003)

Sentiment analysis tools assist businesses in extracting information from unstructured and unorganised material found on the internet, such as emails, blog posts, support tickets, web chats, social media channels, forums, and comments. Algorithms use rule-based, automatic, or hybrid methods to replace manual data processing. Automatic systems learn from rule-based systems and do sentiment analysis based on established, lexicon-based rules.

3.1 Types of Sentiment Analysis

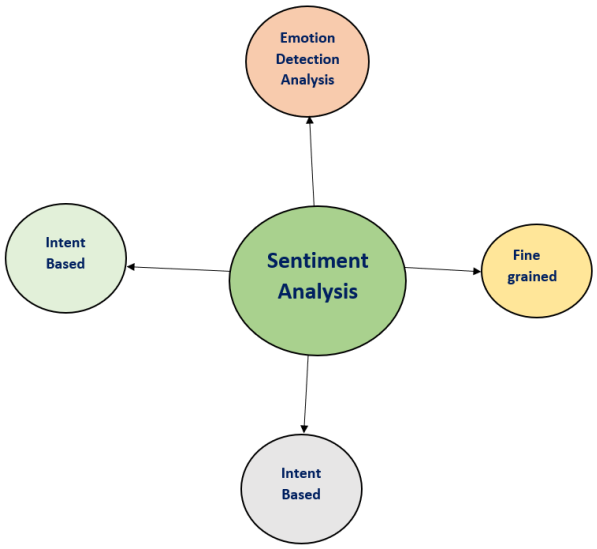


Fig 3: Types of Sentiment Analysis

1)Fine-grained sentiment analysis breaks down polarity into smaller groups, usually highly positive to very negative, to provide a more specific level of polarity. This can be compared to a 5-star rating system in terms of opinion.

2)Rather than identifying positive and negative emotions, emotion detection recognises specific feelings. Happiness, frustration, shock, anger, and grief are just a few examples.

3)Intent-based analysis distinguishes between acts and opinions in a text. An online comment indicating dissatisfaction with changing a battery, for example, can motivate customer service to contact you to remedy the problem.

4)Aspect-based analysis collects the exact component that is being referenced positively or negatively at various levels such as document, paragraph, sentence, and sub-sentence levels.

3.2 About Twitter

Twitter is a microblogging and social networking website based in the United States that allows users to send and receive messages known as "tweets." Jack Dorsey, Biz Stone, Noah Glass, and Even Williams founded Twitter in 2006 as a social networking or blogging site (Twitter, 2016). The concept was to build an SMS-based communication platform where a group of individuals could create an account, update their status, and send text messages. During a brainstorming session at the podcasting business Odeo, Jack first offered this idea to his companions Biz and Even. After more study, the platform Twitter, also known as 'twttr,' was developed, and on March 21, 2006, at 9:50 p.m., Jack posted the first message on Twitter by creating an account on the platform (MacArthur 2016). Today, Twitter is the most popular and successful social media platform. Twitter is a useful tool. Unregistered users can only read tweets that are publicly

visible, whereas registered users can write, like, and retweet tweets. Twitter is used by users using browser or mobile frontend applications, as well as programmatically through its APIs.

Statistics of Twitter user growth

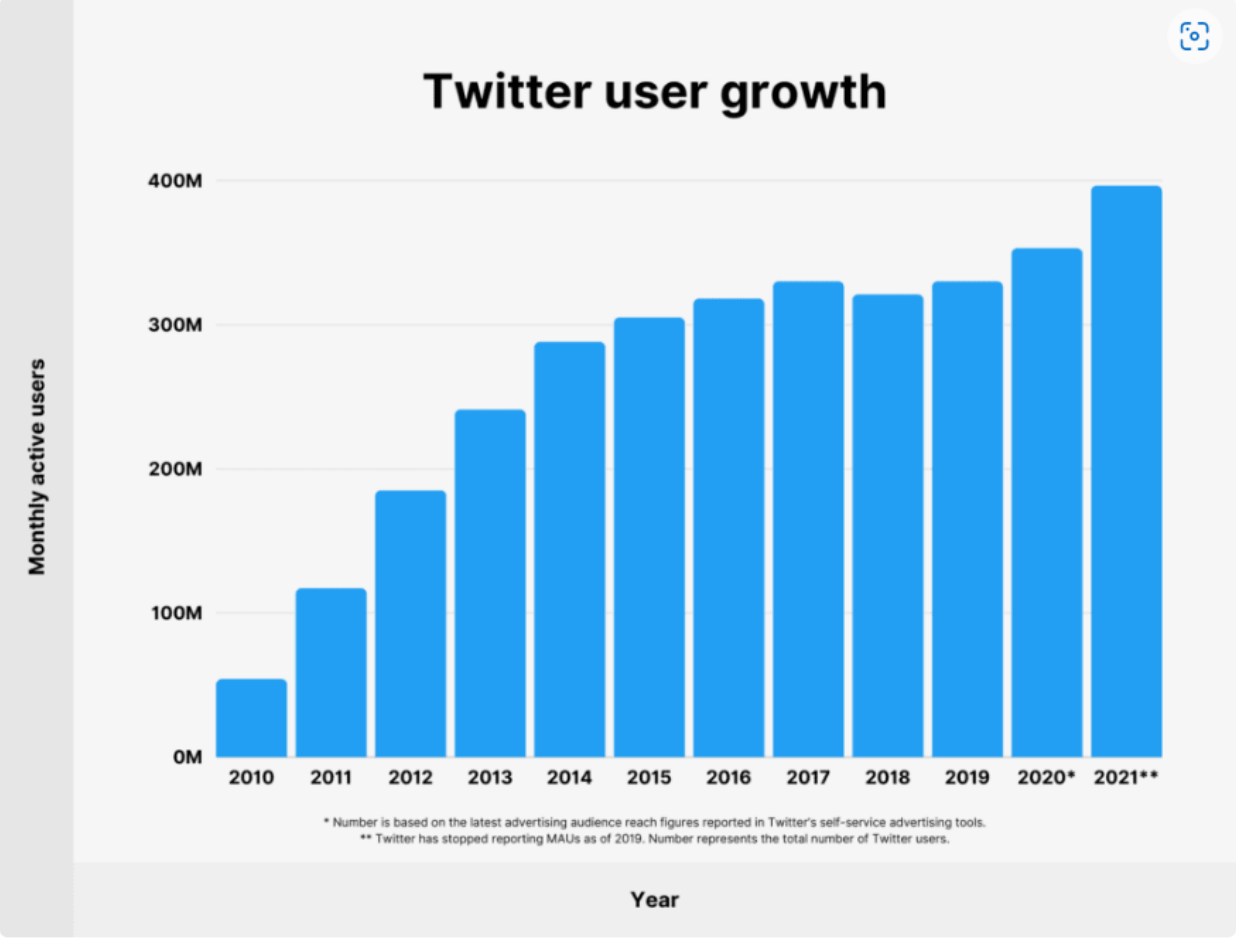


Fig 4: Graphical Representation of active Twitter users over a month

Year	MAU, millions
2010	**54**
2011	**117**
2012	**185**
2013	**241**
2014	**288**
2015	**305**
2016	**318**
2017	**330**
2018	**321**
2019	**330**
2020*	**353**
2021**	**396.5**

** Number is based on the latest advertising audience reach figures reported in Twitter's self-service advertising tools.*

*** Twitter has stopped reporting MAUs as of 2019. Number represents the total number of Twitter users.*

Sources : Statista, DataReportal, DataReportal

3.2.1 Characteristics of Twitter Data

The SMS-based Twitter platform is designed to express one's concept in a clear and succinct manner. As a result, tweets are limited to 140 characters in length; however, sharing movies, photos, and additional tweets is always an option within the tweet (MacArthur, 2016). This concise and accurate expression of one's feelings and thoughts can be shared (Hemalatha et al. 2012). Twitter data also includes '#Hashtags,' the platform's most essential and meaningful symbol. In every tweet on Twitter, this number sign, also known as a hashtag, is used to designate subjects, events, companies, or phrases. For instance, on Twitter, the hashtag '#DonaldTrump' displays all of the most recent or live material, such as news, images, and videos. Donald Trump, the freshly elected

President of the United States, is featured in a series of videos. This means that the symbol # is used to identify a person, corporation, sport, or any public event that occurs throughout the world and is discussed on Twitter. Another key Twitter property or symbol is @' followed by a word or name, which denotes the account's user id. In a Twitter comment, for example, '@narendramodi' is the username (narendramodi) of India's Prime Minister. Furthermore, one can view his or her followers, tweets, retweets, and likes, as well as reply to them using the username '@username'. For example, in the dataset available, '@username' indicates the user's name, as seen in the data set's 'text' element and the 'screen name' attribute.

3.2.2 Twitter API

The Twitter API is an interface that allows a website or app to communicate with Twitter. It gives users access to the platform's major functions, such as posting tweets, retweeting, and searching for tweets that contain a specific word via a website. This is made feasible through so-called endpoints, which are addresses that correspond to specific types of information. The Twitter API makes it simple to acquire tweets, replies, direct messages, advertisements, and publisher tools. To get the details of twitter account for research purpose one should have the Twitter API key to access the account details.

3.3 Sentiment Classification

Sentiment classification is an automated process of identifying opinions in text and labeling the polarity as positive, negative and neutral one based on the expressive emotions. Many popular methodologies and techniques are there to perform sentiment classification. Most of the approaches are based on document
level sentiment classification where a whole document is considered as an informative unit. The methods work on such informative construction can be categorized into three classes: lexicon-based, machine learning-based and rule based methods.

3.3.1 Lexicon Based method

The Lexicon-based technique assesses a document by summing the sentiment ratings of all the terms in the content using a pre-prepared sentiment lexicon. A word and its related sentiment score should be included in the pre-prepared sentiment lexicon. Individual entries for the negation form of vocabulary words should be added to the lexicon, and they should take precedence over the corresponding non negation terms. Negation terms can also be handled with simple rules.

Lexicon-based semantic analysis is one of the approaches or strategies used in the field. From the semantic orientation of lexicons, this technique estimates the sentiment orientations of the entire document or group of sentences. Positive, negative, or neutral semantic orientations exist.

Both manually and programmatically constructed lexicon dictionaries are possible. Many researchers make use of the WorldNet dictionary. To begin, lexicons are extracted from the entire document, and then WorldNet or another online thesaurus is utilised to find synonyms and antonyms to expand the vocabulary. Adjectives and adverbs are used in lexicon-based procedures to determine the text's semantic orientation. Adjective and adverb combinations are retrieved with their sentiment orientation value for computing any text orientation.

3.3.1.1 Dictionary based approach

In this method, a dictionary is built by starting with a few terms. Then, by including synonyms and antonyms of those words, an online dictionary, thesaurus, or WordNet can be utilised to build that dictionary. The dictionary is enlarged until there are no more terms that can be added to it. Manual inspection can help to improve the lexicon. To determine polarity, a dictionary-based technique employs a sentiment lexicon comprising opinion terms and matches them to the data.

3.3.1.2 Corpus based approach

The sentiment orientation of context-specific words is discovered using a corpus-based approach. Two methods of corpus based approach are:

A) Statistical Approach:

Positive polarity is defined as phrases that demonstrate chaotic behaviour in positive activity. They have negative polarity if they display negative recurrence in negative text. The term has neutral polarity if the frequency is the same in both positive and negative text.

B) Semantic Approach:

This method assigns emotion values to words and words that are semantically similar to those words by locating synonyms and antonyms for the phrase in question.

3.3.2 Machine Learning Approach

Sentiment Classification is a type of two way text categorization task. Text categorization classifies data into several pre-defined categories. Text categorization and Sentiment Analysis mainly comes under machine learning methodology(Supervised method, Unsupervised method and Rule based approach).

For sentiment analysis, machine learning methods have been frequently employed. For sentiment analysis, the bag-of-words representation is often utilised. The Bag of Words technique believes that words are self-contained and ignores the significance of semantic and subjective information in a text. All of the words in the text are

given equal weight. For sentiment analysis, the Bag of Words format is often utilised, resulting in a high dimensionality of the feature space. Machine learning algorithms use feature selection approaches to decrease the high-dimensional feature space by removing the noisy and irrelevant features and selecting only the most important characteristics. Machine learning-based sentiment analysis algorithms have recently gained traction in the industry.

3.3.2.1 Supervised Learning Method

Supervised learning is a sort of machine learning in which machines are trained using well-labeled training data and then predict the output based on that data. The labelled data indicates that some of the input data has already been tagged with the appropriate output. In supervised learning, the training data presented to the machines acts as a supervisor, instructing the machines on how to correctly predict the output. The process of supplying input data as well as proper output data to the machine learning model is known as supervised learning. A supervised learning algorithm's goal is to discover a mapping function that will map the input variable(x) to the output variable(y).

A. How Supervised Learning works:

Models are trained using a labelled dataset in supervised learning, where the model learns about each category of input. The model is tested using test data (a subset of the training set) when the training phase is completed, and it then predicts the output.

B. Types of Supervised Machine Learning Algorithm

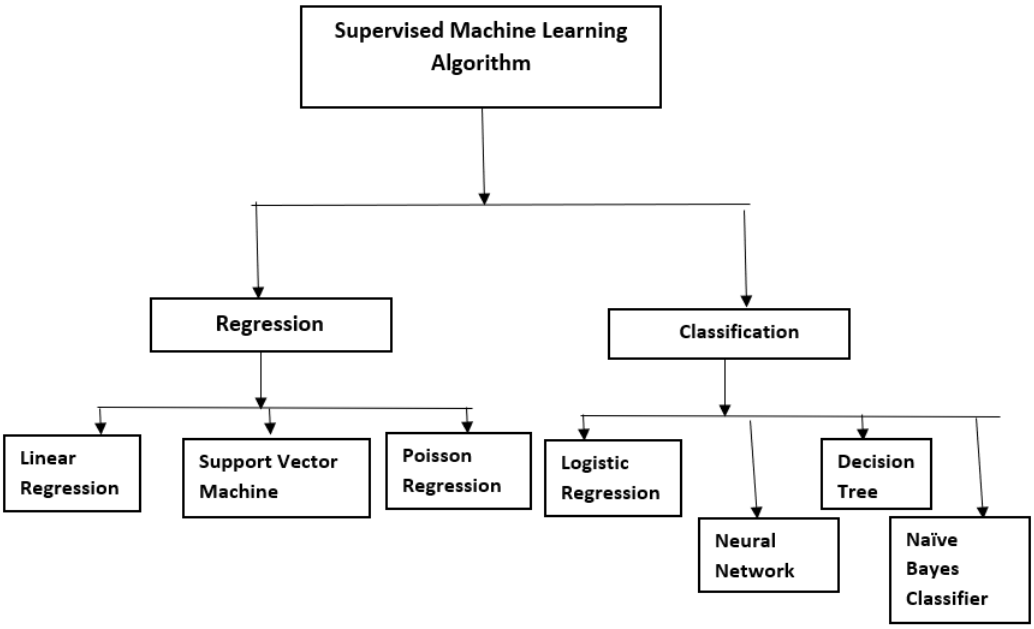


Fig 5: Types of Supervised Machine Learning

1. Regression:

Regression procedures are applied when there is a relationship between the input and output variable. Regression procedures are applied used to predict continuous variables like weather forecasting, market trends, and so on.

1.1. Naïve Bayes Classifier

The supervised learning algorithm known as the naive bayes algorithm, which is based on the Bayes theorem, is used to solve classification

problems. It primarily utilises a high-dimensional training dataset for text classification. The Naive Bayes Classifier is one of the most straightforward and efficient classification algorithms. It aids in the development of quick machine learning models capable of making accurate predictions. It is a probabilistic classifier, meaning it makes predictions based on the likelihood that an object exists.

The words Naïve and Bayes, which make up the Naïve Bayes algorithm, are as follows: The word Naïve presumes that the occurrence of one trait is unrelated to the occurrence of other features and the word Bayes is for the dependency on Bayes principle.[25]

1.1.1 Bayes Theorem

The Bayes theorem, commonly referred to as Bayes' Rule or Bayes' law, is used to calculate the likelihood of a hypothesis given some prior information. Because it is based on the Bayes' Theorem's basic principle, it depends on conditional probability.

The formula of Bayes' Theorem

$$\mathbf{P(A|B)} = \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) (Posterior Probability): Hypothesis A's likelihood of occurring in the observed occurrence B

P(B|A) (Likelihood Probability): Probability of the information provided that a hypothesis is likely to be correct.

P(A) (Prior Probability): hypothesis' likelihood before looking at the evidence.

P(B) (Marginal Probability): The Likelihood of Evidence

1. **Classification:**

   Classification algorithms are utilized when the output variable is categorical, meaning there are two classes, such as Yes-No, Male-Female, True-False, etc.

3.3.2.2 Unsupervised Learning Method

Unsupervised Learning technique is a type of machine learning algorithm which is used to analyze and cluster unlabeled datasets. Hidden patterns or grouping of data can be discovered without human intervention with the help of Unsupervised Learning Algorithms. It enables the model to function independently and find previously unnoticed patterns and information. It mostly addresses unlabeled data.

Unsupervised learning is considerably more like how humans learn to think via their own experiences, which brings it closer to actual artificial intelligence and Finding valuable insights from the data is made easier with the aid of unsupervised learning. It is more significant because it operates on unlabeled and uncategorized data and is necessary to address situations where we do not always have input data and the matching output.

A. Types of Unsupervised Learning Method

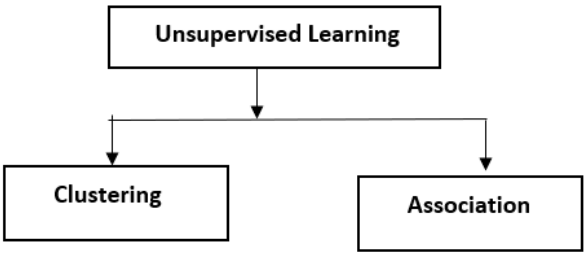


Fig 6: Types of Unsupervised Machine Learning

1. Clustering

Using the clustering technique, items are grouped into clusters so that those who share the most similarities stay in one group and share little to none with those in another. The data objects are classified based on the existence or lack of commonalities discovered by cluster analysis. When it comes to unsupervised learning, clustering is a crucial idea. In a set of uncategorized data, it primarily focuses on identifying a structure or pattern. Unsupervised Education If there are any natural clusters or groupings in your data, clustering algorithms will process them and locate them.

1.1 K-means Clustering Algorithm

It is an iterative clustering algorithm, denoted by the letter K, that aids in locating the highest value for each iteration. The desired number of clusters is initially chosen. You must divide the data points into k groups in order to use this clustering technique. In the same way, a bigger k results in smaller groups with more granularity. Less granularity and larger groups are the results of a lower k.

A collection of "labels" is the algorithm's output. It assigns data point to one of the k groups. Each group in k-means clustering is identified by the creation of a centroid for that group. The centroids act as the cluster's "heart," capturing and incorporating the nearby points into the whole.

2. Association

An unsupervised learning technique called an association rule is used to uncover the connections among the variables in a sizable database. It establishes the group of items that co-occur in the collection.

Associations can be created between data elements in sizable databases using association rules. Finding intriguing correlations between variables in huge databases is the goal of this unsupervised technique.

3.3.3 Rule Based Method

The rule based automatic text categorization systems relied heavily on knowledge engineering techniques , where a set of human-created logical rules would be applied. Basel et .al (2011) ,Computer scientists refer to any machine learning technique that finds, learns, or evolves "rules" to store, manipulate, or apply as rule-based machine learning (RBML). The defining property of a rule-based machine learner is the identification and application of a collection of relational rules that collectively represent the knowledge collected by the system. This contrasts with other machine learners, which often find a single model that can be consistently used to forecast each event. Building such an expert system is usually labor-intensive, time consuming and expensive. Some rule based algorithms are incorporated in lexicon-based- system to provide high level performance. VADER is a rule-based model with rich lexical features. It aims at sentiment analysis in micro-blog data and achieves effective and generalizable results compared to other state-of-the-art methods.

Summary

This chapter is about to the concern topics related to Sentiment Analysis and their discussion. The importance and concept of Sentiment Analysis and Opinion Mining have been discussed here. The rapid growth of Twitter as a social media has been shown here. Different approaches of Sentiment Analysis have been discussed.

Chapter 4

Proposed Work

This chapter is about the detailed work flow of Sentiment Analysis on Twitter data to show the comparative study to show the depression layer and the opinion built on that study. Proposed Work will be divided into multiple chapters. Chapter 4 for Data pre-processing technique along with twitter data fetching technique

4.1 Methodology

In this thesis Sentiment Analysis concept has been proposed on the concept of Machine Learning approach. The project work have been divided into four stages of structures. The whole process is done for 500 tweeter users

- Fetching of Twitter data
- Pre- Processing of Twitter data to build up training and testing data model
- Training the model with Naïve Bayes classifier and predict the value
- Finally Psychological Analysis will be done on multiple Twitter users and final opinion will be done based on result.

For i= 1 to 500 users

Stage 1

Step 1: Fetching of Twitter data with the help of Twitter API. Every time 100 tweets are fetched.

Stage 2

This layer consists of the Pre-processing stages of Twitter data
After fetching of twitter user's data following steps are performed in Pre-Processing stage

Step 2: After fetching of twitter data, tweets are processed under three steps –

i) Removal of White space , #tags,@
ii) Removal of stop words.
iii) Removal of noise.

Step 3: After filtering and cleaning of tweets' features are extracted with the help of NLTK.

i) After the Filtering process filtered tweets will be processed under Tokenization process to collect individual text as a token in the bag.

ii) After the tokenization process cleaned tweets are lemmatized with WordNet lemmatizer to collect meaningful words and arrange them as bag of words along with POS(Parts of Speech) tagging.

iii) Now each word is marked as positive, negative and neutral and sentiment polarity

iv) Now all the positive, negative and neutral data are merged together.

Stage 3:

In this stage the Data model is prepared for training and testing purpose with the help of Naïve Bayes Classification algorithm.

Step 4: Now the bag of words are randomly shuffled to build up the training model. During classification 60% data are considered for Training purpose and 40% of data are testing purpose.

Step5: After the classification process the tweets are collected under a data frame and marked under tweet column and followed by positive, negative and neutral percentage calculation.

Step 6: The predicted values are processed under stage 4.

Stage 4:

In this stage the psychological analysis is performed and opinion is built up.

Step 7: Users positive, negative and neutral thoughts are predicted.

Step 8: Depression score of each users are calculated.

Step 9: Based on the depression score the users are categorized within different depression level groups and the opinions are provided on that.

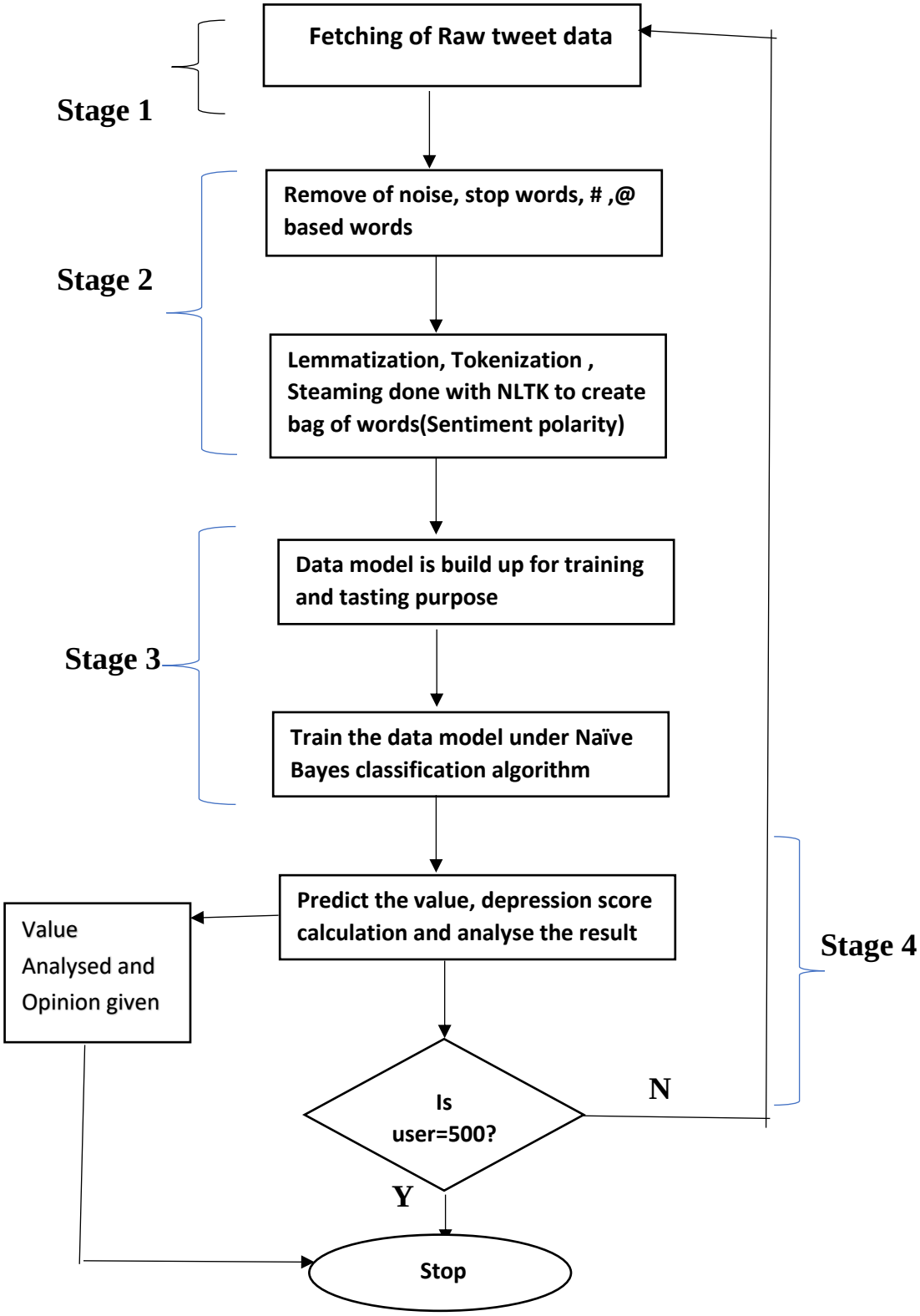


Fig 7: Structure of Sentiment Analysis

4.2 System Requirements

4.2.1 Python

A robust programming language is Python. Programmers can write more effective and efficient code with less lines of code thanks to the notions of data structures and object-oriented programming (Van Rossum et al. 2007). It is open source, and its interactive interpreter enables one to directly code one's programme in addition to granting access to many free standard libraries and online resources to meet your needs for application development. Due to its rich syntax and dynamic coupling, which is more fascinating for processing linguistic data, it is also the most suited language for scripting and application development (Bird et al. 2009). Dynamic name resolution,which permits method and variable name binding during programme execution, is a key feature (Van Rossum et al. 2007).

During this research work Python version 3.10.0 was used for process the linguistic feature with the help of Natural Language Tool Kit. This version is compatible for all these frameworks.

4.2.2 Interface

For programming interface Visual Studio Code
Version: 1.68.1 (user setup)
Commit: 30d9c6cd9483b2cc586687151bcbcd635f373630
Date: 2022-06-14T12:48:58.283Z
Electron: 17.4.7
Chromium: 98.0.4758.141
Node.js: 16.13.0
V8: 9.8.177.13-electron.0

OS: Windows NT x64 10.0.22000
This version has been used to compile, debug, implement and checking and investigating the code to make something better.

4.3 Pre-processing

Twitter is a social networking and micro blogging website. Twitter data has been considered as the corpus of analysis.

4.3.1 Twitter Data Fetching

Twitter API is used to fetch the twitter data of a particular user. Twitter data has been fetched with the help of Python library function.

Library functions:
import tweepy

Table 1: Library function to fetch tweeter data

Each time 100 tweets have been collected form each user. The account of each user has been accessed by @username. From (Date and Time) 22-06-07 03:45:08+00:00 To (Date and Time): 2022-06-22 17:43:14+00:00 (within this time period) 100 tweets are fetched for analysis.

4.4 Data Processing

This study used a semi-structured data set of Twitter data that was made available. The user-generated tweets, which may include noise and incomplete and unreliable linguistic data, are used for research in the data set's "text" field. As a result, in order to analyse data analysis from Twitter, it is important to clear and eliminate this irregular data in order

to account for the data's underlying meaning and attitudes (Hemalatha et al. 2012). Data preparation is useful in this situation. The following article discusses the algorithm used to filter and remove noise from the data, along with all of the preparation chores. The algorithm was built using the Python programming language.

4.4.1 Data Cleaning and Noise Reduction

Every tweets of individual users are very much unstructured and contains noise, unwanted spaces. This textual information is jam-packed with extraneous text, repeating special characters, and possibly even extraneous white space. Therefore, pre-processing and transforming this data so that machine learning algorithm analysis may be applied to it is the first stage in performing sentiment analysis on this data set. Therefore, this irregular data must be cleared and eliminated in order to correctly assess this data from tweets and account for the sentence's genuine meaning and sentiments.

According to Fernández-Gavilanes et al. (2016), preprocessing of data normalises linguistic data, removes noise, and streamlines the language used to analyse attitudes from tweets.

Basic structure of Data Processing is

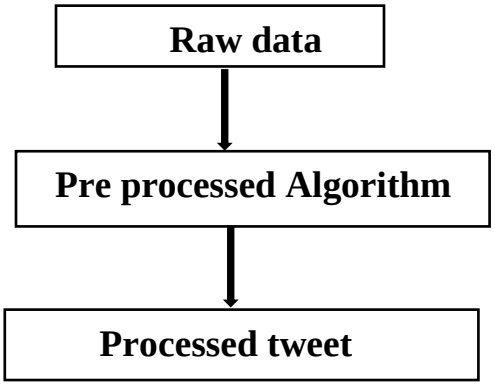


Fig 9: Structure of Pre-processed data

There are lots of research regarding the pre-processed of data and extract a meaningful tweet structure. In the article "Preprocessing the informal text for efficient sentiment analysis", Hemant et al (2012).

In their research work they have created a platform for sentiment analysis using a machine learning algorithm and published their results. Here, they demonstrate how to use a machine learning method to better analyse text input by extracting the qualified content from it within the context of natural language processing. Later, in the case study written by Hemalatha et al. (2014), it was indicated that in order to improve performance and outcomes, terms and expressions should be eliminated.

This structure is quite beneficial to build up a good data model and to get a good result of accuracy.

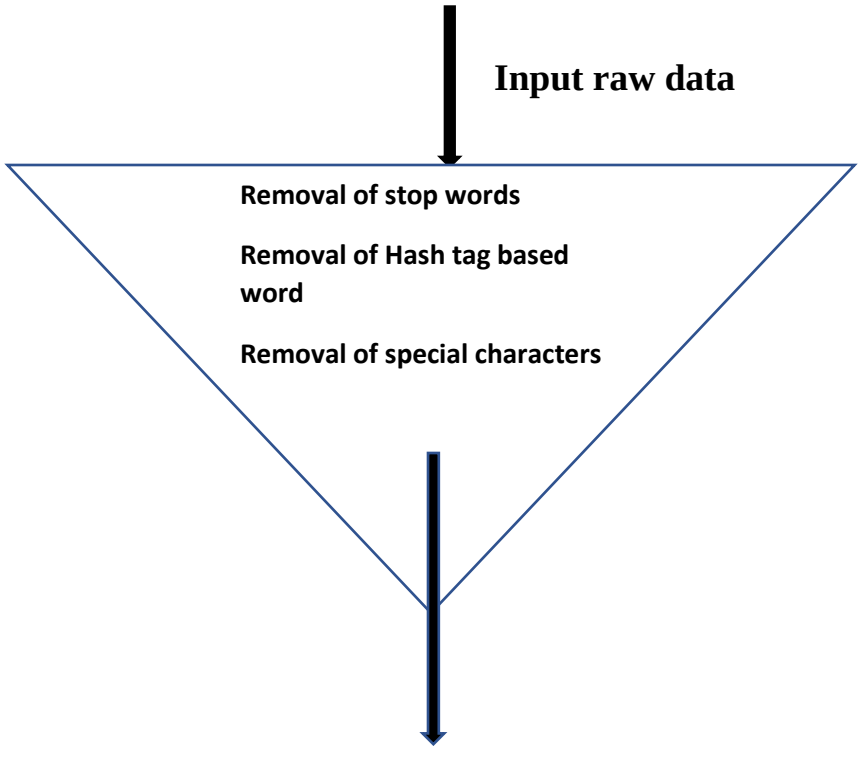


Fig 10 : Structure for preprocessing user data

Pre- Processing algorithm should built up in such a way that noise can remove properly and to get more accurate result.

The algorithm to process the data is:

For each user's Twitter data

Step 1: All the URLs are replaced with the help of model and the values are stored in processed data file.

Step 2: Replace all the @username and #taged words are replaced and stored in the process file.

Step 3: All the repetitive and stopped words are processed and stored in the process file.

Step 4: All the special characters 'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+#]|[!*(),]|' '(?:%[0-9a-fA-F][0-9a-fA-F])are removed and stored in twitter data process file.

Step 5: Return processed data.

i) URL are removed to increase the efficiency

ii) One can see that the event tags are the pieces of information that come after the "#Hashtags." These can occasionally give a word an emotive or sentimental meaning. Therefore, the third step is to merely delete the "#" or "hashtag" symbol from the tweet in order to maintain the term that is followed by the hashtag.

iii) The text data becomes more intelligible and each word provides us with some significance when URLs, retweets, usernames, and #Hashtags are removed. Once more, these were merely some fundamental procedures to carry out in order to prepare Twitter text data for analysis, which not only

eliminates noise from the data but also improves performance for subsequent data processing tasks.

iv) The user-generated content could also include repeating letters, special characters like punctuation, and superfluous whitespace at the start, middle, and end of tweets. First, using a built-in Python method, all excess white space was eliminated. In addition, all of the pointless and superfluous special characters from the tweets were removed. Hemalatha et al(2012)

v) After that special characters along with repetitive letters are removed to make the tweet more sensible for analysis purpose.

Input Data: The movie is good. It should get a good rating @Srijitmukherjee.
Output : The, movie, is, good, It , should, get a good rating Srijit Mukherjee

Table 1: Output Structure of noise removal

Now the tweet data are prepared for next step (Natural Language Processing)

4.5 Data Processing using NLP(Natural Language Processing

4.5.1 Why NLP (Natural Language Processing)

Natural language, which might be English, French, Hindi, or any other language, is the mode of communication that individuals use to interact with one another for specific objectives. Language-based communication is also known as linguistic communication (Bird et al. 2009). Both written and verbal forms of communication are acceptable.

Emails, social media blogs, letters, books, and any other written form, whether typed or handwritten, are examples of written communication. Voice over phone, lecture presentations, and anything auditory are examples of verbal means of communication. Additionally, every medium of communication—verbal or written—has its own lexicon, syntax, grammar, part-of-speech system, or all of these things combined. Consequently, there are two categories in which natural language processing can be divided

Natural language processing is a research area in computational linguistics or artificial intelligence in the discipline of computer science that focuses on the interaction between Computers and artificial natural language (Chowdhury, 2003).

During this POS (Part of Speech) tagging, parsing, named entity recognition, information extraction, word sense disambiguation, word stemming and lemmatization, stop word analysis, word tokenization, and many more techniques are used to handle natural language, depending on the study goal. The evaluation process includes careful observation of the punctuation between the lines and the identification

of idiomatic or colloquial expressions, which aids in understanding and clarifying the "negations" that change the polarisation of the word depending on the different types of parts of speech (nouns, prepositions, adverbs, pronouns, adjectives, interjections, conjunctions, and verbs).

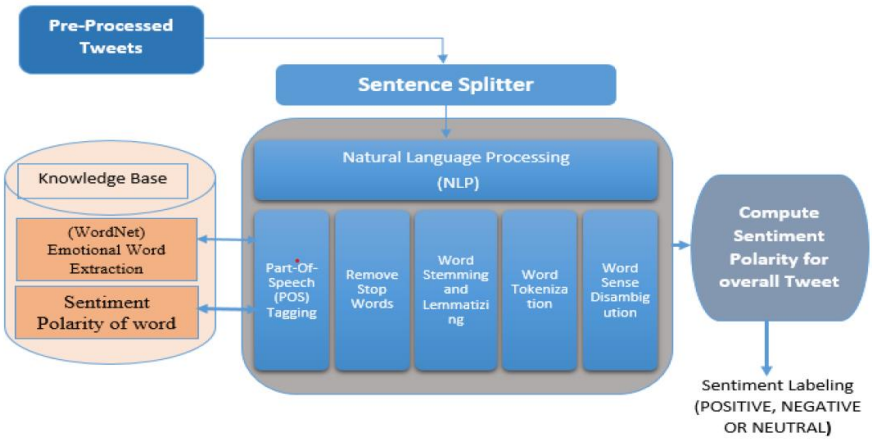


Fig 11. Sentiment Analysis Architecture using NLP

Source: R. Patel, "Sentiment Analysis on Twitter Data using Machine Learning", 2017. [Accessed 22 June 2022].

4.5.2 Natural Language Toolkit (NLTK)

Python uses relatively basic functions for language processing tasks when working with strings. Natural Language Toolkit (NLTK), which is available for Python, is used to provide a sophisticated functionality for processing linguistic data. The GPL open-source licensing for NLTK's set of modules and corpora makes it possible for us to study. Additionally, it provides the pre-processed and raw versions of the typical corpora used in NLP books and courses in addition to methods and packages for popular NLP tasks (Bird et al. 2009).and conduct NLP

research (Bird et al. 2006). For the analysis of text documents, it has more than 50 corpora and lexical resources, including WordNet, in addition to language processing libraries that work tokenization, categorization and stemming, tagging, parsing, and sematic rules (Bird et al. 2006). The academic community has complimented NLTK's self-contained nature, which is its main advantage (Bird et al. 2006)

NLTK imported packages, are used for Sentiment Analysis are-

```
from nltk import FreqDist

from nltk.corpus import stopwords
from nltk.corpus import twitter samples
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.tag import pos_tag
from nltk.tokenize import word_tokenize
from nltk import classify
from nltk import NaiveBayesClassifier
```

Table 2: List of NLTK packages

4.5.3 Tokenization

The sentences' raw words remained after the noise in that dataset was removed. Each of these words has a specific significance and may represent the user's feelings or emotions that were communicated in the tweet. Tokenization is the procedure or procedures used in natural language processing to break down sentences into words and punctuation. Tokenizing language into words is the process of splitting the string into a list of words (Perkins, 2010).

For word tokenization process NLTK package tokenize is imported. It works on the type of dataset. As our main concern is twitter data so the working mode is on English language.

Tokenization Algorithm:

Step 1: Input filtered tweet

Step 2: Process all words of filtered tweet

Step 3: Send the words to Tokenizer method for tokenization

Step 4 : After tokenization process the words are added to tokenized sentence

Step 5: Return Tokenized sentence

Input Data(Processed Tweet) : "Today I am feeling Good"
Output (Tokenized data) : "Today" , "I" , "am" , "Feeling" , "Good"

Table 3: Tokenization sample output

4.5.4 Word Lemmatization

Word lemmatization is a crucial task in natural language processing. By filtering the affixation or replacing the vowel in the word, this strategy converts a word's structural form to its basic or dictionary form. Lemma is the result that the word produces (Liu et al. 2012). Lemmas are words that have the same meaning as the requested term at their root, and lemmatized words serve as the entrance to the WordNet (Bhattacharyya et al. 2014). Thus, by employing an algorithm to lemmatize the word, a lemma will be produced, which will then be sent

on to WordNet, which will extract the word's sense and sense number, with the goal of improving the word's sentiment score.

Here, the "WordNetLemmatizer" class, accessible through the "wordnet" class of the stem package in Python NLTK, is used to lemmatize words by matching characters one at a time. It is a wise idea to make use of in order to generate vocabulary and effective lemmas from the text (Bird et al. 2009)

The Lemmatizing Algorithm is:

Step1: Input Tokenized words

Step 2: For every word in Tokenized word

Step 3: Method call Lemmatizer word using WordNet Lemmatize method

Step 4: Return Lemmatized words

Input Data(Tokenized word): " The" , "Children", "Women", "Life"
Output (Lemmatized word): "The" ," child" ,"Woman" , "Life"

Table 4: Lemmatized Word

4.5.5 Removing Stop Word

When processing natural language, stop words like "and," "the," "am," and "is" that have a high frequency in the document but low emotional significance do not alter the sentiment score when applied to lexical resources.

As a result, it has been standard practise for many academics to remove stop words from documents when evaluating sentiment. In their experiment, (Saif et al. 2012) examined the findings of retaining stop words in the text as well as those produced by filtering stop words from the text, and they found that the results obtained had a high accuracy in sentiment classification for keeping stop words in the document as is.

Input Data(Unfiltered Word Token): "I" , "am" , "children", " school"
Output(Stop word removal): : " child" , "school"

Table 5 : Removal of Stop Word

4.5.6 Bag of Words

Mainly Bag Of Words is a phrase frequency matrix. Bag of words does not yield a defined number of features, in contrast to polarity words, depression words, and pronouns, which do.(Z.Zamil,2022) .

4.5.7 Parts-of-Speech (POS) Tagging

This technique provides the basic form of word meaning disambiguation by annotating the parts of speech (such as Noun, Adverb, Adjective, Subjects, and Objects) to the words and evaluating the sentence structure (Pang et al. 2008). It is the final stage of natural language processing to determine the sentiment of the text, according to Kouloumpis et al. (2011). One can obtain featured words that indicate the sentence structure and the meaning of the words in the domain they belong to in the phrase by carrying out this phase. The POS tagger class from the NLTK package has been used to construct

an algorithm to extract word sense for only English language tags from the phrase in order to accomplish annotated part-of-speech in the method utilised. It examines the syntactic structure of the most basic sentence. In POS tagging if the tagged value starts with ‘NN’ then the POS of that value will be considered as ‘Noun’, If it starts with ‘VB’ then it will be considered as ‘Verb’ and rest of the tagged value will be considered as ‘Adjective;

4.6 WordNet

The functional and intricate WordNet databases enable information retrieval in the area of linguistic data processing (Lam et al. 2014). Emotional words and "sematic links" between words can be found in one of the most well-liked and polite tools for analysing natural language (Ohana et al. 2009). The term "Synsets," "Synonyms set," or "group of synonyms" refers to this connection between the semantic and lexical relationships for the words and their meaning. In the year 2006, there were 207,000 pairs of wordsenses in each of the nearly 115,000 synsets that make up the WordNet database, which, according to Wawer et al. (2010), has 150,000 words. The WordNet lexical collection is reported to contain 155, 287 words and 117,659 synset (similar meaning) in the book (Bird et al. 2009).

After the Pos tagging and all filtering polarity of each word within the BOW will be labeled. Every positive one will be (+1) , negative one will be(-1) and neutral one will be 0.

4.7 Text Blob

A Python (2 and 3) package called TextBlob is used to process textual data. It offers a straightforward API for getting started with typical natural language processing (NLP) activities like part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and others

Summary

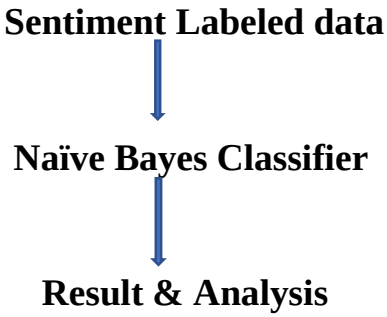
This chapter deals with the proposed methodology of the research work followed by the stages of data pre-processing. The detailed discussion of tweet fetching with Twitter API to several stages of noise cleaning of tweets have been mentioned. The concept of NLP and the data processing tool kit NLTK have been discussed here. How NLP along with NLTK is performing Lemmatization, Tokenization and Stop Words removal have been discussed here. The importance and POS tagging is mentioned along with the polarity checking of individual token within the bag. The concept of Text Blob is mentioned here. Here the detailed discussion of pre-processing of data before going through the Naïve Bayes classification process to provide final result.

Chapter 5

Machine Learning technique for Sentiment Analysis

The data output produced by the algorithm suggested in Chapters 4 and 5—which filters data and analyses linguistic data using Natural Language Processing techniques—is shown in the following paragraphs (NLP).

The total positive score, the negative score in the tweets, and the sentiment labelling ('POSITIVE', "NEGATIVE" and "NEUTRAL") have all been included to this data. The accuracy, performance, and dependability of the results from machine learning-based sentiment analysis are evaluated using these data sets that have been labelled with the sentiment of the tweets. The training data set is ready to execute sentiment analysis using machine learning algorithms like Nave Bayes, and the sentiment labelled data with the total positive and total negative score for the terms in the tweet has been produced.



5.1 Training and Testing Data Set

We separated the 100Users dataset into a training set of 60 users and a test set of 40 users. The division was made at random. In the case of tweet-level classification, the tweets from 60 users were included in the training set, while the tweets from the remaining 40 users were included in the test set.

5.1.1 Importance of Training and test the data set

The objective of machine learning is to create a model that is capable of making precise predictions for a particular task. Making predictions about data it has not yet seen as well as data it has seen is a useful feature of machine learning.

Using sample data gathered from the domain, we construct our estimate of the ideal discriminant function. It is a sample or subset of all potentially available data; it is not all possible data. Predictions wouldn't be necessary if all the data were labelled because the solutions could be easily sought up.

The structure of the data that we use to create our rough model corresponds to the structure of the ideal discriminant function. To best disclose this structure to the modelling programme is the aim of data preparation. Additionally, the data includes elements unrelated to the discriminant function, such as biases resulting from data selection and random noise that disturbs and obscures the structure. These challenges must be solved by the chosen model to approximate the function.

It is quite likely that a model will have lesser accuracy on unobserved data if it is chosen for its accuracy on the training dataset rather than its accuracy on an unobserved test dataset. The model's limitations are the cause of this. It has become customised to the training dataset's structure. Overtting describes this (Z.Zamil,2022)

5.2 Naïve Bayes Classifier

The most used classifier in Natural Language Processing is Naïve Bayes.

The Naive Bayesian Classification is both a statistical and supervised learning method for classifying data. It is a probabilistic model, allowing us to establish probabilities to logically express model uncertainty. It aids in the resolution of diagnostic and predicative issues. In honour of Thomas Bayes, who developed the Bayes Theorem for calculating probability, this classification is called Naive Bayes. By combining prior knowledge with observed data, Bayesian classification offers helpful learning methods. It aids in offering a practical viewpoint for comprehending and also assessing several learning algorithms. This aids in calculating precise probability for hypotheses and is resistant to input data noise.(Praveen, Huma;Pandey,Sikha(2016))

Naïve Bayes algorithm and various machine learning algorithms for sentiment analysis were compared by Pang et al. (2008), who classified the data set 90 percent accurately. This classifier's key benefit is its simplicity and ability to anticipate the appropriate class for a new instance (Murphy 2006). It merely categorises as a class any feature values that have been retrieved from each instance of the class (for example, POSITIVE, NEGATIVE, and NEUTRAL). Each labelled emotion tag (instance) is given equal weight in relation to the other tokens in the data set and contributes to the final classification result. This classifier will be used in machine learning to categorise the sentiment labelled.

Naïve Bayes Classifiers considers every attribute of instances work independently in the class. In order to compute Bayesian probability, it will therefore multiply each member of the feature vector in the provided class set of data. The sentiment labelled will be categorised using this classifier in machine learning, and other qualities won't be taken into account any longer. In order to classify data using the Naive

Bayes classifier, only one attribute—"Tweet Sentiment"—will be taken into consideration.

Naïve Bayes Classification algorithm

Step 0: Start

Step 1: For every attribute of B

Step 2: At every testing node traverse attribute list for B

Step 3: Calculating the probability using value of B to be in that class.

Step 4: Updating the B attribute class.

Step 5: Checking for all the values that B have or not.

Step 6: If yes go to step 7 else go to step 3

Step 7: Checking for next available attribute if yes go to step 2
Else go to step 8

Step 8: END

$$\mathbf{P(A|B)} = \frac{P(B|A)P(A)}{P(B)}$$

Where,

A: Is instance of class (sentiment labels for each tweet)
B: Is sentiment class (POSITIVE, NEGATIVE or NEUTRAL)
P(A|B): instance occurred in particular class for each value of B (class-conditional density)
P(A) : prior probability of class

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Prediction}{Total\ number\ of\ prediction}$$

$$= \frac{TP+TN}{TP+TN+FP+FN}$$

TP = True Positive
TN= True Negative
FP= False Positive
FN= False Negative

Based on the Test data and Trained data the Accuracy value of the data model is calculated.

Using the above formula 500 users accuracy values have been calculated.

Summary

In this chapter the Naïve Bayes Algorithm Concept is discussed to calculate the Accuracy of data model.

Chapter 6

Result & Analysis

This chapter is about the result work of the whole experimental data. This thesis work is about the computational study of Sentiment Analysis where Twitter data is the main corpus to analysis the depression level of every twitter user.

Data Set:

100 Twitter data set of each user has been collected. Then the unstructured twitter data has gone through several steps to prepare itself for training and testing model.

6.1 Accuracy Calculation

During the experiment Naïve Bayes Algorithm has been improvised to calculate the accuracy every user.

Accuracy= (test data, trained data)

Applying the Naïve Bayes classifier the accuracy of the this method is more than 90%

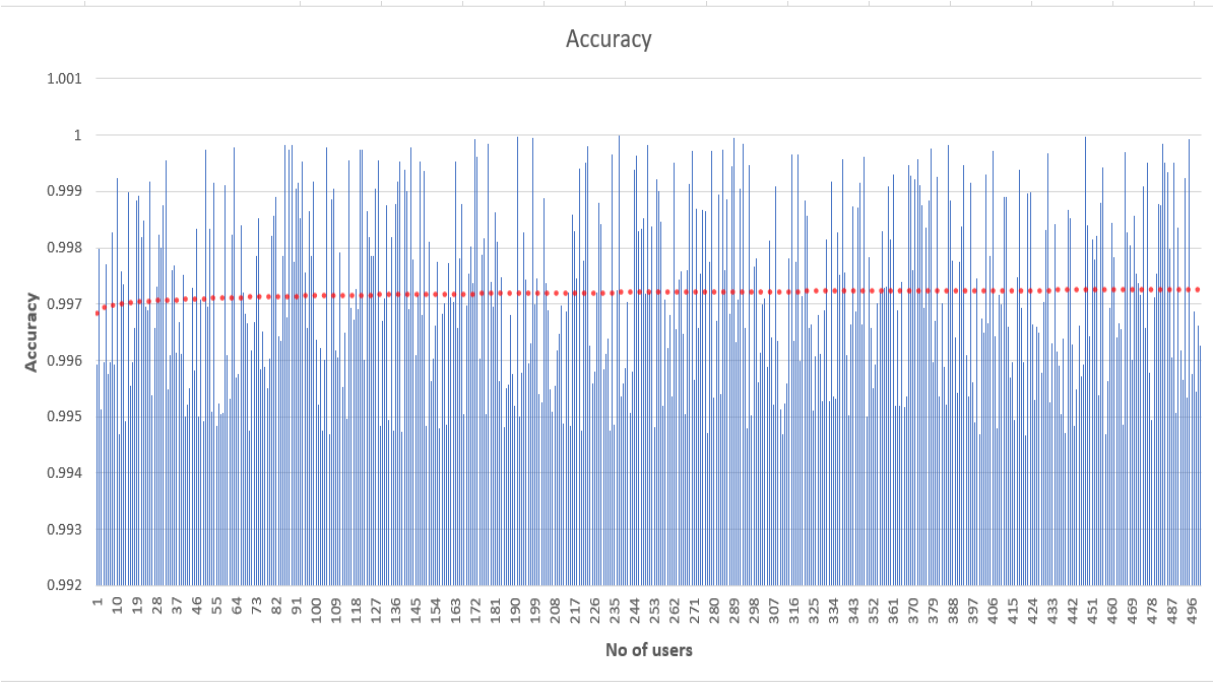


Fig 12 : Graphical representation of the accuracy values of 500 users

B	C
User name	**Accuracy**
Kamaal R khan	0.994666667
Sharukh Khan	0.990555556
Roddur Roy	0.991777778
Narendra Modi	0.990555556
Mamata Banerjee	0.991555556
Madan Mitra	0.990777778
Amitabh Bachan	0.994111111
Shraddha Kapoor	0.994444444
Katrina Kaif	0.991
Srijit Mukherjee	0.992111111
Vicky Kaushal	0.993111111
Rupankar Bagchi	0.994

Table 6: Accuracy values of Twitter users

6.2 Depression score analysis

Depression Score = $\frac{Negative\ Polarity}{No\ of\ tweets}$

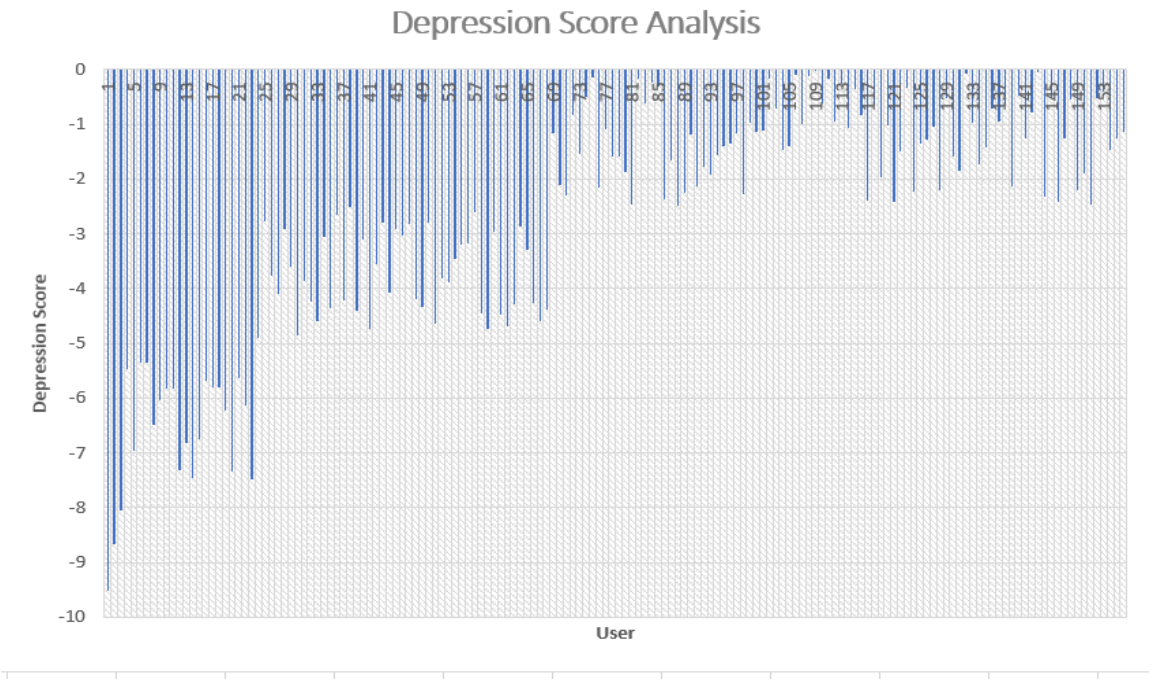


Fig 13: Depression range vs users

As the depression value is negative so the graph is inverted one.

As per the sampling of every 10 users twitter account it has been noted that every 10 users there is 29% users are belonging within a depressed zone where as 71% of them are having the positive polarity compare to negative one.

Here my observation is near about 500 users. So based on every 10 users calculation approx. 147 twitter users are there where depression

scores are under observation. Rest of 343 users are belonging to the zone of positive polarity.

Now the scenario of observation is as per the collected twitter corpus is real time based so when 100 tweets are being fetched very few users are there who are belonging from a danger zone.

As per the observation, based on the depression score the whole twitter users' scores are divided within 4 groups. Where it is noted that high risk depression scores are appearing within the range of -10 to -7.5. and there after the ranges are categorized towards the positive zone as per the approx. ranges of -7.5 to -5.1, followed by -5.0 to -2.5 and the almost positive score is -2.5 to 0.

Now the opinion set is declared based the zone of depression score.

Twitter users' group having the score of -2.5 to 0 are overcoming their negative zone and getting back to a good life. As per the research work 56% of having depression scores are belonging from this zone.

Twitter users having the score of the zone -7.5 to -5.1 will be advised to think positive and they are also tending to a positive life. Only 23% of people are belonging from this zone. As per their twitter analysis it has been noted that their recent tweets are tends to very positive one. Very few old tweets are there for negative score

The zone of depression scores within -5.0 to -2.5 users are advised to think positive in their life. They yet have not overcome all the negativities. In this zone only 13% of user's analysis have been observed. Their recent tweets are gradually healing towards to positive pole.

The final zone of the depression category is within -10 to -7.5. There is very few users are having this. As per the analysis only 2% of users are there who are really struggling with their life. So, the opinion for this zone is to consult with doctor immediately.

Summary:

This chapter is about the result and the analysis of accuracy and depression score calculation. And the analysis portion contains the opinion of depression score zone as per their score and the time zone of their posted tweets.

Chapter 7

Conclusion

In this research I have collected the tweeter corpus data of near about 500 people to analyse their depression score and make a group to category them within a range. I have tried to grab those users account who are currently on a trend to spread their negative thoughts on Twitter .

The whole research work is done with the implementation of Naïve Bayes Classification method. The main advantages of this classifier are its simplicity and capability to predict the right class for a new instance (Murphy 2006). Any feature values that have been received from each instance of the class are simply classified as members of that class (for example, POSITIVE, NEGATIVE, and NEUTRAL). In comparison to the other tokens in the data set, each labelled emotion tag (instance) is given equal weight and contributes to the final classification outcome. Because of its advantages this classification method has been chosen to predict the output.

Moreover this research work will work efficiently to predict the depression level of any twitter user.

Limitations: i) As twitter is negative comments restricted media so it is tough to generate depression score. So now it is little bit tedious to identify proper depressed user.

ii) As we are collecting data as per real time based so every time the accuracy and depression score will be changed.

Example: During the testing @kamaalrkhan's depression score was -6.5 but after few days the depression score became -2.5 as in between the positive polarity of 100 tweets got changed. So graph and value calculation gets difficult to predict.

Future Work

For the future work the Sentiment Analysis can be done based on real time polarity tagging of twitter data. The whole structure of the proposed work can be reimplement with the help of different machine learning technique to get more accurate predicted value.

It is also expected to implement the concept of this research work for different social media corpus to predict depression score of users so in future the target of this project can overcome the limitations and can work with more success.

Bibliography

[1] F.Neri, C.Aliprandi, F.Capeci, M.Cuadros, T.By, “Sentiment Analysis on Social Media", IEEE/ACM International Conference on Advances in Social Networks Analysis andMining (ASONAM), 2012, pp. 919 – 926

[2] Po-Wei Liang, Bi-Ru Dai, “Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy,June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-http://doi.ieeecomputersociety.org/10.1109/MDM.2013.

[3] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf.* Retr., 2(1-2):1–135.

[4] Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Emotion classification using web blog corpora. In *WI ’07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 275–278, Washington, DC, USA. IEEE Computer Society

[5] Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *ACL*. The Association for Computer Linguistics.

[6] Alec Go, Lei Huang, and Richa Bhayani. 2009. Twitter sentiment analysis. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group

[7] Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", In *Proceedings of the 13th International Conference on Discovery Science,* Berlin, Germany: Springer,2010, pp. 1-15.

[8] Pak and P. Paroubek. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining", *In Proceedings of the Seventh Conference on International Language Resources and Evaluation*, 2010.

[9] P. D. Turney,”Thumbs up or thumbs down?: semantic orientation appliedto unsupervised classification of reviews,” in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417-424, Association for Computational Linguistics, 2002.

[10] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

[11] Gokulakrishnan, Balakrishnan, et al. "Opinion mining and sentiment analysis on a twitter data stream." Advances in ICT for emerging regions (ICTer), 2012 International Conference on. IEEE, 2012.

[12] G. Gautam and D. Yadav, “Sentiment analysis of twitter data using machine learning approaches and semantic analysis”, in 7th Int. Conf. on Contemporary Computing, 2014, pp. 437-442.

[13] Goel, Ankur, Jyoti Gautam, and Sitesh Kumar. "Real time sentiment analysis of tweets using Naive Bayes." Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on. IEEE, 2016

[14] Yair Neuman, Yohai Cohen, Dan Assaf, Gabbi Kedma, “Proactive screening for depression through metaphorical and automatic text analysis,” *Artificial Intelligence in Medicine*, Vol. 56, No. 1, pp. 19-25, 2012.

[15] Diveesh Singh and Alineen Wang, “Detecting Depression Through Tweets” Standford University CA 9430, ;pp.1-9

[16] "Company | About." *Twitter*. Twitter, 30 June 2016. Web. 04 Mar. 2017.

[17] Hemalatha, I., Dr GP Saradhi Varma, and A. Govardhan. "Preprocessing the informal text for efficient sentiment analysis." *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)* 1.2 (2012): 58-61.

[18] MacArthur, Amanda. "The Real History of Twitter, In Brief - How the micro-messaging wars were won." *lifewire*, 3 Oct. 2016 https://www.lifewire.com/history-of-Twitter-3288854.Accessed1 December 2016.

[19] Nasukawa, Tetsuya and Jeonghee Yi. *Sentiment analysis: Capturing favorability using natural language processing*. In *Proceedings of the K-CAP-03, 2nd International Conference on Knowledge Capture*. 2003.

[20] Dave, Kushal, Steve Lawrence, and David M. Pennock. *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. In *Proceedings of International Conference on World Wide Web (WWW-2003)*. 2003.

[21] How Many People Use Twitter in 2022? [New Twitter Stats] (backlinko.com)

[22] Liu B (2010) Sentiment analysis and subjectivity. In: Indurkhya N, Damerau FJ (eds) Handbook of natural language processing, 2nd edn. Chapman & Hall/CRC, Boca Raton, pp 627–666

[23] Lin Y, Zhang J, Wang X, Zhou A (2012) An information theoretic approach to sentiment polarity classification. In: Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality. Lyon, France, pp 35–40

[24] Naive Bayes Classifier in Machine Learning - Javatpoint

[25] Bassel, G.W., Glaab, E., Marquez, J., Holdsworth, M. J., & Bacardit,J.(2011). *Functional Network Construction inArabidopsisUsing Rule-Based Machine Learning on Large-Scale Data Sets. The Plant Cell, 23(9), 3101–3116.* doi:10.1105/tpc.111.088153

[26] Van Rossum, Guido. "Python Programming Language." *USENIX Annual TechnicalConference*. Vol. 41. 2007.

[27] Bird, Steven, Ewan Klein, and Edward Loper. *Natural language processing with Python*.“O’Reilly Media, Inc.", 2009.

[28] Hemalatha, I., Dr GP Saradhi Varma, and A. Govardhan. "Sentiment analysis tool usingmachine learning algorithms." *International Journal of Emerging Trends &Technology in Computer Science (IJETTCS)* 2.2 (2013): 105-109

[29] Fernández-Gavilanes, Milagros, et al. "Unsupervised method for sentiment analysis in online texts." *Expert Systems with Applications* 58 (2016): 57-75.
Firmino.

[30] Hemalatha, I., Dr GP Saradhi Varma, and A. Govardhan. "Case Study on Online Reviews Sentiment Analysis Using Machine Learning Algorithms." *International Journal of Innovative Research in Computer and Communication Engineering* 2.2(2014):3182-3188.

[31] Chowdhury, Gobinda G. "Natural language processing." *Annual review of information science and technology* 37.1 (2003): 51-89.

[32] Image Source: R. Patel, "Sentiment Analysis on Twitter Data using Machine Learning", 2017. [Accessed 22 June 2022].

[33] Bird, Steven. "NLTK: the natural language toolkit." *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, 2006.

[34] Bird, Steven, Ewan Klein, and Edward Loper. *Natural language processing with Python*.“O’Reilly Media, Inc.", 2009.

[35] Perkins, Jacob. *Python text processing with NLTK 2.0 cookbook*. Packt Publishing Ltd, 2010.

[36] Liu, Haibin, et al. "BioLemmatizer: a lemmatization tool for morphological processing of biomedical text." *Journal of biomedical semantics* 3.1 (2012): 1.

[37] Bhattacharyya, Pushpak, et al. "Facilitating multi-lingual sense annotation: Human mediated lemmatizer." *Global WordNet Conference*. 2014.

[38] Saif, Hassan, Yulan He, and Harith Alani. "Semantic sentiment analysis of Twitter." *International Semantic Web Conference*. Springer Berlin Heidelberg, 2012.

[39] Z. Zamil, "MONITORING TWEETS FOR DEPRESSION TO DETECT AT-RISK USERS", 2017. [Accessed 23 June 2022].

[40] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." *Icwsm* 11 (2011): 538-541.

[41] Lam, Khang Nhut, Feras Al Tarouti, and Jugal Kalita. "Automatically constructing Wordnet Synsets." *ACL (2)*. 2014.

[42] Ohana, Bruno, and Brendan Tierney. "Sentiment classification of reviews using SentiWordNet." *9th. IT & T Conference*. 2009

[43] Parveen, Huma; Pandey, Shikha (2016). *[IEEE 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) - Bangalore, India (2016.7.21-2016.7.23)] 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) - Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. , (), 416–419.* doi:10.1109/ICATCCT.2016.7912034