SENTIMENT ANALYSIS AND CLASSIFICATION OF TWEETS BASED ON MACHINE LEARNING

2022
MASTER THESIS
COMPUTER ENGINEERING

Firas Fadhil SHIHAB

Thesis Advisor
Assist.Prof.Dr. Dursun EKMEKCI

SENTIMENT ANALYSIS AND CLASSIFICATION OF TWEETS BASED ON MACHINE LEARNING

Firas Fadhil SHIHAB

T.C.
Karabuk University
Institute of Graduate Programs
Department of Computer Engineering
Prepared as
Master Thesis

Thesis Advisor
Assist.Prof.Dr. Dursun EKMEKCI

KARABUK
Jun 2022

I certify that in my opinion, the thesis submitted by Firas Fadhil SHIHAB titled "SENTIMENT ANALYSIS AND CLASSIFICATION OF TWEETS BASED ON MACHINE LEARNING" is fully adequate in scope and in quality as a thesis for the degree of Master of Science.

Assist.Prof.Dr. Dursun EKMEKCI .....
Thesis Advisor, Department of Computer Engineering

This thesis is accepted by the examining committee with a unanimous vote in the Department of Computer Engineering as a Master of Science thesis. Jun 22, 2022

Examining Committee Members (Institutions)	Signature
Chairman : Assist.Prof.Dr Adnan Saher ALAJEELI (KBU)	.....
Member : Assist.Prof.Dr. Dursun EKMEKCI (KBU)	.....
Member : Assist.Prof.Dr. Veli BAYSAL (BARU)	.....

The degree of Master of Science by the thesis submitted is approved by the Administrative Board of the Institute of Graduate Programs, Karabuk University.

Prof. Dr. Hasan SOLMAZ .....
Director of the Institute of Graduate Programs

"I declare that all the information within this thesis has been gathered and presented in accordance with academic regulations and ethical principles and I have according to the requirements of these regulations and principles cited all those which do not originate in this work as well."

Firas Fadhil SHIHAB

ABSTRACT

M. Sc. Thesis

SENTIMENT ANALYSIS AND CLASSIFICATION OF TWEETS BASED ON MACHINE LEARNING

Firas Fadhil SHIHAB

Karabük University
Institute of Graduate Programs
The Department of Computer Engineering

Thesis Advisor:
Assist. Prof. Dr. Dursun EKMEKCI
Jun 2022, 68 pages

Sentiment analysis is a technique for mining online forums like Twitter for information about people's thoughts, feelings, and attitudes. It has grown in popularity as a source of study. Conventional sentiment analysis focuses mostly on textual data. Twitter is the most well-known micro-blogging social networking service, where users send out short messages (called "tweets") on a variety of subjects. In recent years, Twitter data has been utilized to improve political campaigns, product quality, and sentiment analysis. This study proposes the use of a machine learning classifier to assist in sentiment analysis for these organizations. Based on the content and tone of the tweets, tweets were classified into three categories: positive, negative, and neutral. Extracted Twitter data has been preprocessed in 11 stages in order to ensure classification accuracy when using feature extraction algorithms such as Term Frequencies and Inverse Document Frequencies (TF-IDF). According to these results, ensemble classifiers outperform non-ensemble classifiers. According to tests, machine learning

Classifiers may be improved by using TF-IDF as a feature extraction method. The Word to Vector (W2V) feature extraction process is less efficient than the TF-IDF feature extraction process. TF-IDF and the Bag of Words (BoW) were then picked as lexicon-based techniques deployed. Based on the results five machine learning models have been used to illustrate the best-categorized methods for region-based Twitter sentiment analysis. As it turned out, the Extra Trees classifier outperformed the BoW and linear classifiers for the TF-IDF feature in terms of performance. Using logistic regression, the provided classifiers outperformed their counterparts (LR). The results evaluation performance has been the F1 score of 0.6133 and an accuracy of 0.9616.

Key Words: Text classification, feature extraction, sentiment analysis, TF-IDF, machine learning, BoW, natural language processing.

Science Code : 92431

ÖZET

Yüksek Lisans Tezi

MAKİNE ÖĞRENİMİNE GÖRE TWEETLERİN DUYGU ANALİZİ VE SINIFLANDIRILMASI

Firas Fadhil SHIHAB

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr.Öğr.Üyesi. Dursun EKMEKCI
Haziran 2022 , 68 sayfa

Duygu analizi, insanların düşünceleri, duyguları ve tutumları hakkında bilgi almak için Twitter gibi çevrimiçi forumlarda madencilik yapmak için kullanılan bir tekniktir. Bir çalışma kaynağı olarak popülerlik kazanmıştır. Geleneksel duygu analizi, çoğunlukla metinsel verilere odaklanır. Twitter, kullanıcıların çeşitli konularda kısa mesajlar ("tweetler" olarak adlandırılır) gönderdiği en iyi bilinen mikro blog sosyal ağ hizmetidir. Son yıllarda, siyasi kampanyaları, ürün kalitesini ve duygu analizini iyileştirmek için Twitter verileri kullanıldı. Bu çalışma, bu kuruluşlar için duygu analizine yardımcı olması için bir makine öğrenimi sınıflandırıcısının kullanımını önermektedir. Tweetlerin içeriğine ve tonuna göre, tweetler olumlu, olumsuz ve nötr olmak üzere üç kategoriye ayrıldı. Çıkarılan Twitter verileri, Terim Frekansları ve Ters Belge Frekansları (TF-IDF) gibi özellik çıkarma algoritmaları kullanılırken sınıflandırma doğruluğunu sağlamak için 11 aşamada ön işleme tabi tutulmuştur. Bu sonuçlara göre, topluluk sınıflandırıcıları, topluluk olmayan sınıflandırıcılardan daha

iyi performans göstermektedir. Testlere göre, makine öğrenmesi öznitelik çıkarma yöntemi olarak TF-IDF kullanılarak sınıflandırıcılar geliştirilebilir. Word'den Vektöre (W2V) özellik çıkarma işlemi, TF-IDF özellik çıkarma işleminden daha az verimlidir. TF-IDF ve The Bag of Words (BoW) daha sonra konuşlandırılan sözlük tabanlı teknikler olarak seçildi. Sonuçlara dayalı olarak, bölgeye dayalı Twitter duygu analizi için en iyi kategorize edilmiş yöntemleri göstermek için beş makine öğrenimi modeli kullanılmıştır. Sonuç olarak, Ekstra Ağaçlar sınıflandırıcısı, performans açısından TF-IDF özelliği için BoW ve doğrusal sınıflandırıcılardan daha iyi performans gösterdi. Lojistik regresyon kullanarak, sağlanan sınıflandırıcılar benzerlerinden (LR) daha iyi performans gösterdi. Sonuç değerlendirme performansı, 0,6133 F1 puanı ve 0,9616 doğruluk olmuştur.

Anahtar Kelimeler : Metin sınıflandırma, özellik çıkarma, duygu analizi, TF-IDF, makine öğrenimi, BoW, doğa dili işleme.

Bilim Kod : 92431

ACKNOWLEDGMENT

I owe thanks and praise first and foremost to Allah the Almighty for this success and facilitation.

I would like to give thanks to my advisor, Assist.Prof.Dr. Dursun EKMEKCI, for his great interest and assistance in the preparation of this thesis, who spared no effort in providing unlimited advice and guidance until the completion of this thesis to the fullest image, I also extend my thanks and gratitude to the University of Karabuk, including the wonderful professors, doctors, and colleagues who accompanied us throughout our academic journey.

I dedicate this thesis to my beloved country, Iraq. And to beautiful Turkey, which embraced this scientific experiment and contributed to providing all possibilities to graduate in this distinguished way.

I bow to my beloved parents. My dear father gave me the most valuable things. My beloved mother is good at engineering my heart with her prayers. To my family in which I grow up and its extension that gives me pride and honor, my loved ones, my friends, and everyone who supported me on this journey, to achieve this dream, Insha'Allah I will continue to work and study to reach higher degrees.

CONTENTS

Page

Page

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

PART 1

INTRODUCTION

Due to the vast quantity of data available on the internet, numerous corporations have been interested in extracting this data, which may be quite beneficial to them. Sentiment Analysis is a completely new and extensive topic of research as a result of this. Cluster analysis, sentiment extraction, and other terms have been used to describe this field. However, there is a subtle distinction in meaning between these phrases. Longitudinal survey methodologies were incredibly biased before fully automated sentiment analysis as they were taken in isolation by users. As a result, the need for an automated process that can deal specifically with tens of people of opinions hidden in users' posts in comments, blog posts, and other forms of social media arose [1].

Sentiment analysis has a wide range of applications, including product evaluations, film reviews, economics, politics, and recommender systems. A company can make adjustments based on customer feedback on a product or different parts of a product. Similarly, public policies can be changed based on public opinion regarding a certain political party. In this context, deep learning and linguistic sentiment analysis are the two basic methodologies employed.

Sentimental Analysis [2] is an information extraction approach for extracting information about people's thoughts, attitudes, and feelings concerning commonplace events. And everyone has a different perspective on the same issue. Technically, the sentiment analysis task is more difficult, but it is more practical. For example, businessmen are constantly interested in hearing what the general public thinks about their products and getting feedback from different consumers. Customers also want to know how other consumers who have purchased the product have rated it, and marketers like sentiment analysis since they want to know who their target customers [3][4].

Most websites have a geo-tag feature that allows users to add a geographic characteristic the location to their post as an add-on. For this type of supplementary information, Goodchild coined the term "Volunteered Geographical Information (VGI)". This type of geo-tagged material is referred a "check-in record" by Liu. Several researchers have employed VGI and check-in information [5] [6]

1.1. THE CONSUMERS PERSPECTIVE

Consumers may use sentiment analysis comments to assist them in making decisions. Previously, while reviewing a product, consumers would seek the advice of intimate friends and relatives. However, with the advent of social media, individuals are now expressing themselves on the internet [7]. This material is valuable for consumers looking for product reviews to assist them in making a purchasing choice. These decisions are often binary in nature, i.e., the customer either buys it or does not.

1.2. PROBLEM STATEMENT

As previously stated, academics have investigated a range of methods for detecting social events using social media. However, just a few academics have attempted to identify real-life occurrences using sentiment as a significant component and indication of society's status [8]. The goal is to extract statements of opinion characterizing a target feature and categorize it as positive or negative from a batch of tweets including various features and varying viewpoints.

1.3. CONTRIBUTION OF THE THESIS

- Data is collected, cleaned, and analyzed for meaningful perspectives to be gleaned through the research process. The collecting of Twitter data from January 2021 to April 2022 was part of this project's activity. Approximately 49,000 tweets are represented in the form of rows. The user's Twitter ID and date were the only data elements used in this analysis (represented as text).

- To reduce the amount of data to fewer rows, special characters, missing values, and stop words should be removed.

- The data were subjected to feature selection using a machine learning technique in order to use only those characteristics (columns) that provide a more accurate interpretation of the data. Data analysis does not make use of every aspect of the data. It is possible to get skewed results if you include characteristics that are not important.

- To construct a model that can be used to assess the rest of the data, machine learning classifiers were conducted on a subset of the data, known as the train data. The models were created using (5) different classifiers, and the accuracies of each model were examined to see which model performed the best. Support Vector Machine, Decision Tree Classifier, Random Forest Classifier, Logistic Regression, and XGB Classifier are some of the classifier's sed.

1.4. RESEARCH OBJECTIVES AND QUESTIONS

Use several Machine Learning approaches to classify each tweet as positive or negative sentiment, and then see which classifier performs the best. Sentiment analysis and opinion mining [9] [10] is an open research field with a wide range of practical applications. Humans utilize blogs, forums, Twitter, Facebook, and other online sites to communicate their viewpoints [11].

Social media has brought individuals from all over the world closer together; communication is now just a click away. Prior to social media, mobile service providers offered a costly short messaging service (SMS) with both domestic & global rates.

This thesis aims to create a mechanism for detecting events based on demographic sentiment using geotagged social media data, which will lead to an understanding of the populous sentiment in time and place.

Three sub-objectives and associated research problems make up the primary objective:

1) Find a sentiment indicator that accurately represents the population's sentiment.
2) Create a methodology for spatial-temporal analysis of population sentiment to discover sentiment shifts and anomalies.
3) Identify and analyze the event in the spatial-temporal dimensions using appropriate approaches.

Here is the sub-question of the Research:

RQ 1) Which sentiment classification approach for the next from social media postings is best for determining the best indicator for expressing the population's sentiment orientation?
RQ 2) what should a survey for community sentiments be designed? What are the precise approaches or algorithms that can be used? What are the aspects of the analysis?
RQ 3) after geographic properties have been defined, how may events be detected? What are the methods best for interpreting the event?

1.5. THE ORGANIZATION OF THE THESIS

As mentioned earlier, that is dissertation reviews and analyzes the machine learning for sentiment analysis Tweets, the terminology we'll use, as well as the natural language processing methods used, before moving on to a discussion of opinion mining research are presented in Part 1. Part 2 delves into the state-of-the-art of sentiment analysis classification and Twitter data pretreatment. Part 3: We use many training datasets of tweets and orientation using a machine learning technique. Part 4: We use unit-gram and bigram features to examine the Twitter dataset and compare them to confirmed discoveries found in the dataset. Finally, we discover which classier the best is. Part 5: We conclude the thesis and discuss our future intentions.

PART 2

SENTIMENT ANALYSIS

Sentiment analysis is a method for determining how people feel about a subject, a product, a company, or even politics. Natural language processing and machine learning methods were used to investigate sentiment analysis. Formerly used journal article surveys gauge client attitudes but tracking and gathering all customer feedback is now impracticable. With the rise of social media, sifting through all client data and evaluating their thoughts as good or negative has become simpler and more accessible.

Sentiment Analysis [12] is a kind of computational analysis that examines people's sentiments, emotions, views, and judgments depending on their written words. Because of the internet and social media, people's views are becoming more significant in decision-making. People use social networking services like Facebook, Twitter, and blogs to express themselves in their original language. South Africa plays a huge role in the film business, politics, and marketing. Tweets are communications with a character restriction of 280 characters. As a result, sentence-level analysis is the best strategy for Twitter SA.

Social media sites like Twitter, Instagram, and Facebook have stormy interaction settings, therefore it is vital to express sensitive information about people's ideas, feelings, and opinions regarding any commodity, concept, or policy through these platforms [38]. This information is useful to both customers and providers. Consumers often examine other people's comments on a product when purchasing online.

The manufacturer may learn about the merits and downsides of its products depending on the customer's feedback. Although these views may benefit both businesses and people, the sheer volume of these opinions on text data can be overwhelming for consumers. It is a highly intriguing field for scholars to investigate and summarize the views expressed in this extensive opinion text material.

Sentiment Analysis or Opinion Mining [39] is another name for this current field of research. Every data collection has its own patterns. Examining hidden trends and patterns might assist foresee and averting possible issues. A machine-learning system takes a certain sort of data and exploits the patterns concealed in it to answer additional questions. Many businesses that deal with big amounts of data are increasingly aware of the value of machine learning. Furthermore, low-cost computer data and managing storage choices have enabled the development of models that evaluate massive amounts of complex data fast and accurately. Businesses must understand exactly how to match the suitable algorithm with a specific learning process or resources of Machine Learning & its applications in order to get the most value from big data (India is an excellent place to outsource) [40] [41].

2.1. OVERVIEW OF SENTIMENTAL ANALYSIS

Sentimental analysis [15] gathers primary data from various and un-oriented textual materials from different social media and online resources, such as conversing on social networking sites like Twitter, WhatsApp, Facebook, live blogs, or feedback.
Studying how people express themselves in ordinary language in reaction to a certain incident is known as "sentiment analysis." Because Twitter's profile information allows for collecting subjective data, they are used most often [12] [16].

This shows that sentiment analysis has reached a position where it can categorize positive/negative sentiment and handle the complete field behavior/sentiment for several networks and themes, as shown by current events." In the field of sentiment analysis, there has been a great quantity of studies done utilizing various approaches to anticipate social sentiments. According to the work in [12] [42], an appraisal might be favorable or negative based on the fraction of lovely comments to total words.

In the last years, the inventor created a technique for filtering tweet results according to the phrases mentioned in the tweet [17]. More study on Twitter sentiment analysis was conducted by Go et al. [18], who referred to the topic as multiple classification tasks, i.e., categorizing tweets into positively and negatively classed categories.

A Hadoop-based system was suggested by M. Trupthi, S.Pabboju, and G.Narasimha. Social Networking Sites (SNS) providers that leverage Twitter's streaming API are used to extract the data. The tweets are fed into Hadoop, which previously processed them using map-reduce algorithms. They used a single-word naive Bayes classification technique [19].

The study [20] looks at how SA is used in corporate applications. In addition, this article demonstrates how to use text analysis to audit popular beliefs about a brand and uncover hidden information that may be used for decision-making once the text analysis is completed. The qualitative study described in the article [21] was conducted in four phases. It takes up to a given number of real-time tweets, tokenizes each one as part of the pre-processing, matching them to a list of phrases, and classifying them as positive or negative.

2.2. SENTIMENT ANALYSIS RESEARCH FIELDS

In sentiment analysis, significant areas of research include subjectivity identification, attitude interpretation, aspect-based emotion summary, text overview, comparison point overview, specific product extraction, and the detection of opinion spam bots [16] [17].

Subjectivity Identification: This check determines whether the text is expressed. Predicting whether the text is favorable or bad is known as sentiment prediction. Sentiment Succinct summation Based on Aspects: It gives an opinion outline of sentiments in the form of high ratings or credits for product qualities.

Text Summarization: This feature generates a few words summarizing a product's reviews. Contrastive Perspective Summarization: It summarizes the Contrastive Viewpoint, highlighting opposing viewpoints.

Extraction of Product Features: It is essentially a task that stems from evaluating product attributes.

Detecting opinion spam: It allows for detecting false or incorrect opinions from feedback that necessitates spam thought detection.

2.3. LITERATURE REVIEW MODEL EVALUATION

Sentiment analysis is the careful examination of how one's feelings and points of view are linked to one's ideas and attitude as expressed in ordinary language in response to a certain occurrence. The primary rationale for employing Twitter profile information is that it permits the collecting of subjective data [18].

Deep learning is now showing promise in Computational Mathematics. Here are some examples of DNN-based projects. Cambria investigated the impact of emotions on sentiment categorization in 2016 [19], using a hybrid polarity detection approach that used Sentic-Net and Deep Learning algorithms. Wang et al. recommended capsule model-based back-propagation neural networks for trend analysis [20]. For classifying aspect feelings, the authors used an attention-based sentiment reasoner. They employed attention processes to figure out how much different words in a sentence were worth. Four different English and Chinese datasets were used to test the AS Dissatisfactory model [21].

Tweets are imported into Hadoop and pre-processed using map-reduce techniques. They used unit words in a Naive Bayes classification technique [22] [43]. The study [23] looks at how SA is used in commercial applications. This article also demonstrates the text analysis approach for auditing consumers' popular perceptions of a certain brand and reveals hidden data that can be used for decision-making when the text analysis is completed [44].

The results of [20] should have been utilized to identify which users were the most engaged and helpful. Identifying the retweet percentage and network architecture of active users, on the other hand, would help determine their effect and resolve the debate over the link between someone being active and being influential. Furthermore, neither the number of tweets nor even the date of participation is an indicator of effect, Therefore, the tactics in [24] lack the theoretical behavior of influence. Moreover, past achievement does not imply future success. Text communications are simpler to spread than visual information, according to ref. [19].

People are more interested in information sharing than in interacting with other users, as seen by this. Individuals were also much more willing to argue and share information and perspectives on a specific subject than to participate in conversations in response to top news messages, demonstrating that individuals are more willing to argue and share information and perspectives on a specific subject than to participate in conversations. Consequently, the number of people participating in breaking news events has risen.

SA has been produced in different cities in India among other Indian languages. Because Malayalam is a strongly fusional language, it is more difficult to preprocess than other languages. The absence of the labeled dataset is a serious concern in Malayalam SA. For the SA of Malayalam movie reviews, [24] employed both linear SVM and CRF methods. Their work did not include hyper-parameter adjustment. And in [25], we used several DNN models to conduct SA of Malayalam tweets. For the feature matrix, they considered all of the words in the corpus.

Finally, they demonstrated that GRU outperformed other DNN models. A brief description of the many types of algorithms utilized in sentiment analysis is provided. Sentimental analysis is described as the examination of a text's ideas, thoughts, feelings, and subjectivity. The relevance of several domains such as transfer learning, sentiments detection, and creating resources is highlighted, recently proposed algorithms and emotional analysis tools. The major purpose of this research is to classify current publications, and 54 of the most recent publications have been classified and summarized using content analysis [28].

SA accomplished this with the use of machine learning methods such as Convolutional Neural Networks (CNN) and Extended Short Attention span (ESAS) [26]. When building the similarity matrix for the input data, they considered all of the individual words in the text. The characteristic matrix was huge since they didn't eliminate extraneous terms. The driving drivers behind big data include advanced ways for data processing with massive and high-dimensional information, dramatically increased storage space, and intricate data formats.

Big Data in this domain need cutting-edge Technology and/or strategies to address the various computational times in order to capture meaningful data without compromising sensitive data. A new and rapidly expanding field of research has recently been proposed: machine learning to overcome these problems. Master learning algorithms have generally been thought to learn from large data volumes and find practical and valuable information [27].

2.4.DATA PRE-PROCESSING

Following data purification, knowledge pre-processing is the next phase. It's a huge advancement in the field of learning algorithms. It is the process of converting or encoding data into a machine-understandable format. In layman's words, the algorithms can quickly comprehend the dataset's features. The component is a measurable attribute of the thing being observed. A person's height, age, and gender, for example, are all characteristics. In an unstructured style, a Twitter stream pulls all related tweets from Twitter.

Cleaning is also accomplished by utilizing python regular expressions to remove non-letters or graphics. To process, all of the tweets must be in the same case; as a result, it will shift to lowercase, and each word will be separated depending on space. After that, collect all top words and organize them into a single set before removing them. Finally, return a list of significant terms [29].

Before extracting the characteristics, a pre-processing step is conducted to filter away slang terms and misspellings [30]. The steps below can assist with data pre-processing:

- Evaluation of data quality. It would be absurd to expect the data to be flawless because it is derived from numerous sources. While pre-processing data, the first step must be to assess its quality.

- Values that are inconsistent. At times, data might be inconclusive for example, enter a phone number in the "address" box. As a result, the evaluation should be done appropriately, including determining the data type of the characteristics.

- Aggregation of features. As the name implies, characteristics are combined to improve performance. When compared to individual data items, aggregated features behave substantially better.

- The following are some examples of features. It's a way to choose a subset of the original (first) dataset. The essential feature of sampling is that the selection should have properties nearly equal to the original dataset.

  The considerations will be carried out in the proposed model:
  - Changing the case of tweets.
  - Substitute spaces for at least two dots.
  - Remove any superfluous spaces and replace them with a single space.
  - Remove spaces and quotations from the end of your tweets.
  -

This pre-processing phase aims to get text data ready for future processing. Feature Vector Construction and Feature Selection Text data cannot be processed immediately by a computer, which is an inherent difficulty. Text data must also be understood mathematically.

Term definitions are frequently employed to describe the text's features. This adds a lot of depth to the text rendering. Must filtered features to reduce dimensionality and remove noise to increase classification performance and processing efficiency.

Text Analytics Classification Methods: The Multilayer Perceptron Naive Bayes Algorithm and the K-Nearest Neighbor Algorithm are two notable and extensively used classification techniques for assessing the sentiment polarity of users' views based on opinion data given. Support Vector Machines' Mechanism

1. Linear Regression: A test for normality is defined as the estimation of the value of the dependent or dependent variable using different statistical techniques. A relationship is defined as the translation of a variable down a straight line, as indicated by the equation Y= a*X, where Y is the dependent variable, X is the independent component, is the slope, and is the intercept.

2. Logistic Regression: This approach creates a discrete response variable from a set of independent factors. Logistic regression provides the coefficients for estimating a Probability logistic transformation. A clustering algorithm with a tree-like architecture may be utilized for regression and classification. The best feature of a dataset is included using a decision tree building approach, following which the training data set is divided into subsets. The decision tree technique is used to build an effective model for predicting the objective variable's category or value.

3. Binary Classifier (BC): A Probabilistic Classifier (BC) is a Support Vector Machine (SVM). Row data were derived on the n-dimensional point. A hyperplane is drawn to separate the data sets. The training examples margin is increased as a result of the improved separation.

4. Naive-Bayes: This approach is based on the Bayes theorem, employed in more advanced classification systems. It is a method of categorization. It discovers how an entity with specific qualities belonging to a specific category or class may exist.

2.5. MACHINE LEARNING BASED ON SENTIMENT ANALYSIS

2.5.1. Machine Learning Classifiers

Machine Learning (ML) is an important part of this process. Sentiment Analysis (SA), a machine learning approach, aids the system in detecting the sentiment of a particular comment. Multiple deep learning algorithms underpin the system, which can identify the kind of sentiment or a set of feelings. ML algorithms beat knowledge-based and English oxford methods in identifying polarity in studies [13] [14].

Machine learning refers to the process of allowing computers to understand and solve problems on their own by recognizing patterns in big data sets. Examining underlying trends and patterns may help predict and avoid potential problems. A computer system analyzes a certain kind of data and uses the hidden patterns to answer further queries. Machine learning is becoming more important to many firms that deal with large volumes of data. Furthermore, low-cost computer processing and data storage options have allowed the creation of models that quickly and accurately analyze large volumes of complex data. To get the most out of big data, companies must understand how to match the right algorithm to the right learning process or resources [14].

To measure the sentiment of English tweets, three machine learning algorithms were used, including NB, SVM, and RF. The most difficult procedure for successful data forecasting is the selection of hyper parameters. NB uses the Multilayer perceptron NB classifier to assess if the sentiment of the test results is positive or negative. Various strategies are used to determine the polarity of analytical results. The most common and efficient sentiment analysis technology is machine learning. The polarity in analysis data and the most successful algorithm are computed [31].

The gathering of information. Data sets can be utilized for any type of text classification job that is size-specific in terms of the number of words. After a little preparation, such as case folding and term elimination, these data sets were utilized for sentiment analysis.

2.5.2. Supervised Learning

Based on one or more independent factors, machine learning may identify variables. A bank client's status (a variable called status) is the dependent variable in this example, while the loan amount, length, and demographics of the customer are the independent variables. This is an example of a dependent model. Both the measurement items (X) and the dependent variables (Y) are referred to as "input" and "output" variables, respectively. Using lions, horses, dogs, and cats as an example from a dataset is instructive. Figure 2.1 shows the architected of supervised learning.

The machine learning technique is taught to comprehend a subset of data that has been correctly tagged [44] [47]. After that, the model is given the remaining data (test data) to sort. The model can properly identify which animal is which based on test results since it already recognizes the attributes of the many animals. With the help of supervised machine learning, it is possible to predict the dependent variable from the independent factors.

Supervised learning is often a classification issue when the goal variable (also the dependent variable) is a categorical data type. It is possible to develop supervised learning models using the Logistic Regression, Random Forest, Decision Tree KNN, Linear Support Vector Machine, and Non-Linear Support Vector Machine algorithms.

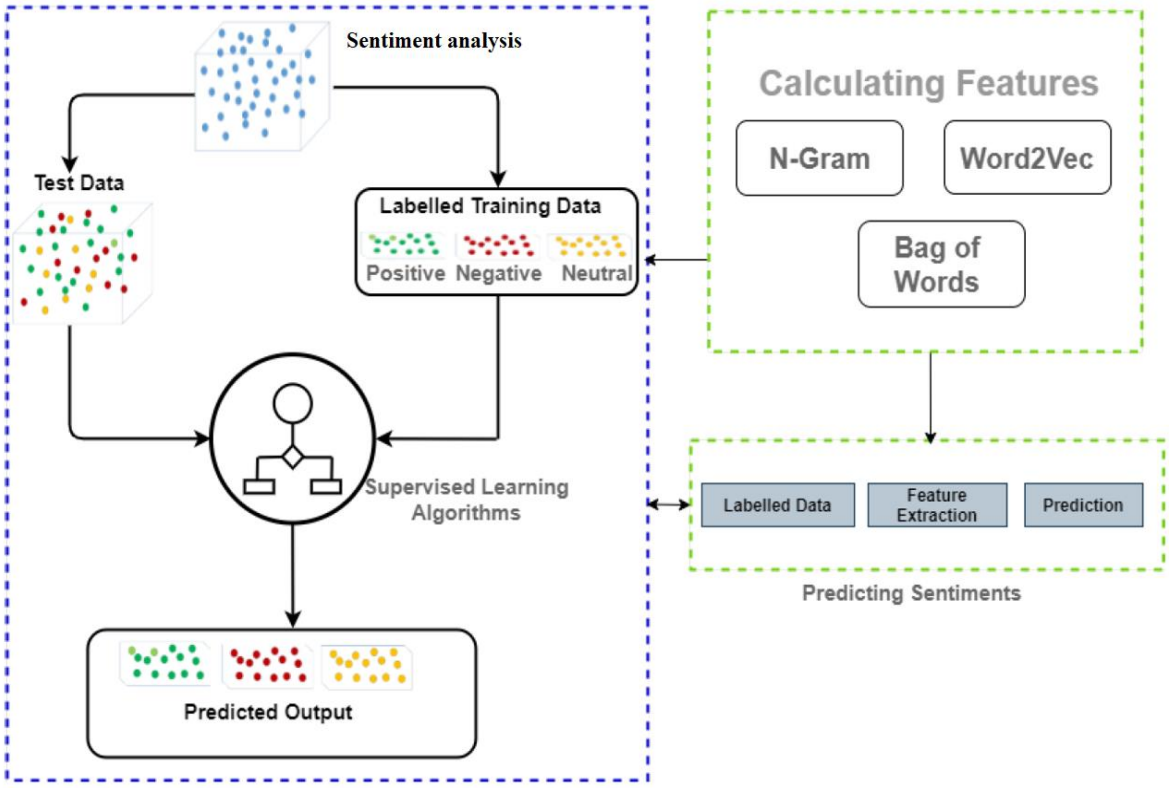

Figure 2. 1 Supervised Learning

2.5.3. Semi-Supervised Learning

Semi-Supervised Unlabeled data (input data) and recognized data (output data) are the two main types of machine learning in this form of learning algorithms (output data). The model is trained using this information. Between unsupervised and supervised learning, this is a sort of machine learning that may be used. This is the case for the vast majority of data collected in the real world. With pseudo labeling, a variety of neural network models and training approaches may be combined [33].

Just as in supervised learning, a limited subset of the data is used to train the model until a satisfactory result is produced. In order to forecast outputs, we utilize the unlabeled training data (also known as pseudo labels). Perhaps it's incorrect. There is a connection between real and fake labels in the labeled training data. Both the labeled training data and the unlabeled data are connected in their inputs, as is their output [44] [47]. The model is re-trained in the same way as before to reduce mistakes and increase the model's accuracy. Figure 2.2 displays the Semi-Supervised learning example.

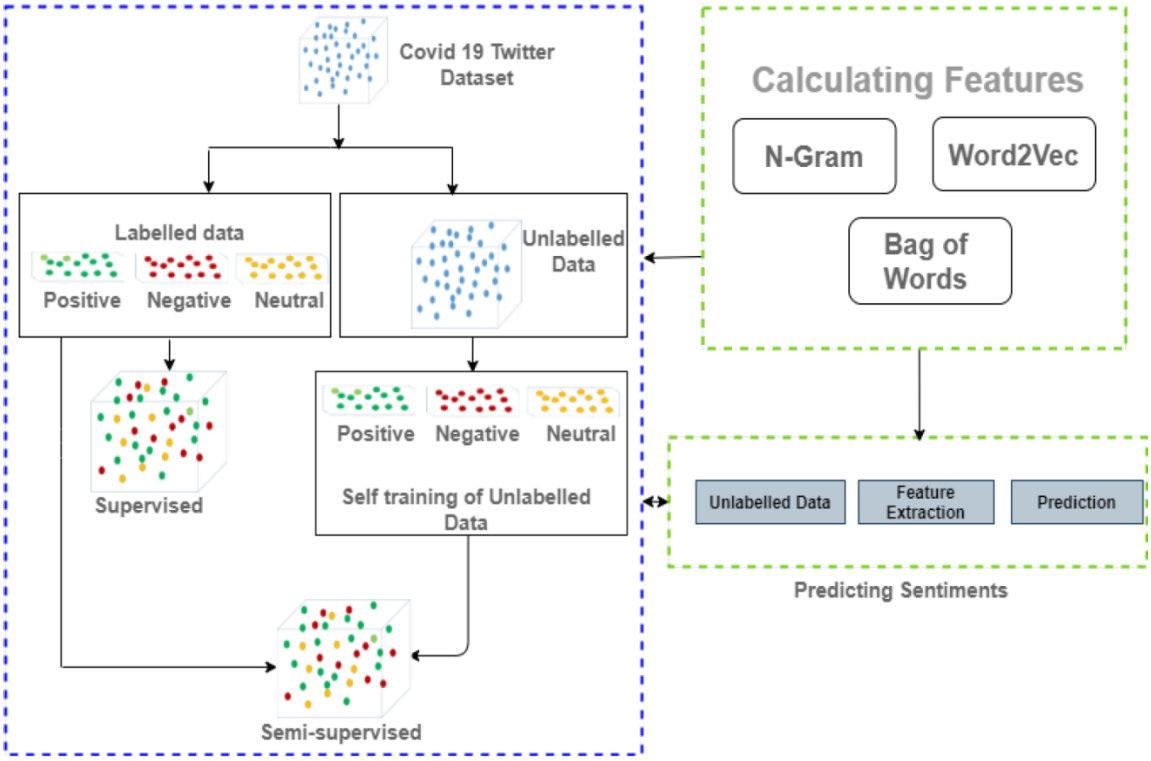

Figure 2. 2 Semi-Supervised Learning

2.6. NATURAL LANGUAGE PROCESSING APPROACH

Natural Language Processing (NLP) techniques are used to analyze Twitter sentiment. According to researcher [32], a vast amount of data from people's ideas on Twitter is required for opinion mining. Several ways in natural language processing aid in the direct retrieval of tweets from Twitter. The tweets are not organized. To accomplish opinion mining and data formats, it must analyze and sanitize tweets.

Until the research, all links, hashtags, and capitalized terms, repetitive phrases, and appropriate statistical, spelling errors, special symbols, twitter characters, and residual content are removed from the data. Text from tweets is removed as part of the data removal procedure. It only contains the text of tweets that have been processed and cleaned. This word for tweets is one by one and uses the Vader lexicon tool to get the word meaning from WordNet. The value of a growing word is measured and recorded as a tweet emotion score. A machine learning classifier labels each tweet as good, average, or bad if consensus is obtained.

Multinomial Naive Bayesian and Logistic Regression were used to analyze Twitter sentiment. Twitter sentiment studies are debatable. Certain of the difficulties include I the fact that some tweets are widely known in vulgar languages and that even a few brief sentences only convey minor indicators of emotion. (ii) Hashtags, URLs, acronyms, emoticons, and acronyms are also often used on Twitter.

The accuracy of various machine-learning techniques is used in concepts such as tweets before transmission, extraction approaches, and table design. With the train data on the test outcomes, machine learning approaches are used to exercise the algorithms, which include Multinomial Naive Bay and the logistic regression algorithm. The author looked at the airline sentiment set of data as well as the analytical information. When using the Count Vectorizer feature, people of all sorts get excellent results in machine learning. According to the author, the test set is derived from the Logistic Regression with Counting Vectorizer features [33].

2.7. APPLICATIONS OF SENTIMENTAL ANALYSIS USING ML TECHNIQUES

Machine Learning Techniques for Twitter Sentiment Analysis. Multi-layer Perceptron (MLP), Naive Bayes, Fuzzy Identification, Decision Trees, and Support Vector Machines are among the machine learning approaches used to classify tweets (SVM.). Such strategies aid in examining numerous component vectors with a doled-out class, allowing the assessment and each element's connection dependency to be identified.

The Twitter dataset examines performance metrics such as accuracy, duration, alert, and F calculation. These methods are put to the test. For classification techniques, accuracy ranges from 73.66 to 93.34 percent, accuracy from 73.16 to 90.12 percent, recall is 74.81 percent to 95.34 percent, and F-measurement is 73.4 to 93.3 percent. SVM outperformed all other methods in the Twitter Sentiment analysis [34, 45].

Using a machine learning method, sentiment analysis for the Indian Premier League. [35] Describes an opinion mining method on social media using a unique machine learning strategy. After the algorithm depends on what the hashtag (# IPTEAM) is, the tweet list of 2016 Indian premier league tweets that are seen using Twitter's API (Application Programming Interface) services. The Random Forest methods' performance is compared to the accuracy, precision, and sensitivity of directed machine learning techniques already deployed.

Sentimental Analysis in Multilingual Web Texts Using Machine Learning [36] conducted research to define nostalgic customer remarks in blog words and sentences. Some of the articles are available in French, Dutch, or English. The primary goal was to divide the exams into "positive" and "neutral" emotive categories. They took distinct techniques to solve the mission's issues in this situation.

They were able to attain the most outstanding results using natural language processing and machine learning methods. A small number of linguistic characteristics are added to the features. The percentages for English, Dutch, and French web information are 83 percent for English, 70 percent for Dutch, and 68 percent for French web information [46].

To categorize sentiment assessments, the n-gram pattern recognition approach was applied. [37] Employed supervised methods such as Naive Bayes, Maximum Entropy, Stochastic Gradient Descent, and Support Vector Machines to assess the video. However, it has been proved that transforming texts to a quantitative matrix using template matching, bigram, word embedding, and mixtures of these, as well as a combination of TF-IDF and Scores Victories, provided more precise classification results. However, the fact that the Twitter message cannot be reviewed in small quantities or in cases where phrases or symbols express the feeling and repetition of the last letter numerous times is a drawback. Both flaws can be used to improve the recognition of emotions in the context of future study.

The approaches for machine learning utilized in the study of emotions in recent times are summarized in this work. Industry, politics, public behavior, and finance are among the several application areas of sentiment analysis that are investigated.

The influence of performing data transformations on the accomplishment of classification algorithms is discussed in this work. However, the type of transformation depends on the dataset and its language. Machine learning algorithms appear to consistently provide identical outcomes, depending on the form of those outputs.

This review anticipates that sentiment analysis applications will continue to expand in the future, and that sentiment analytical approaches will be standardized across diverse systems and services.

The future study will concentrate on three distinct features that will be used to analyze diverse datasets using a combination of logistic regression and SVM methods. Through this effort, it can uncover unfair good and negative evaluations, reputation difficulties, as well as cooperation and control. In Table 2.1 summarized the machine learning models of state of arts for tweets SA.

Table 2.1: Summarizes studies that employed machine learning methods for tweets SA.

No	Reference	Method	Features	Database	Description
1	Abd El-Jawad [11]	hybrid model	without	1 million tweets	compares the performance of various machine learning and deep learning algorithms, as well as introducing a new hybrid system for sentiment classification that uses text mining and neural networks.
2	Yadav [13]	KNN	TF-IDF	6000 tweets	The Twitter sentiment analysis performed determines what percentage of public opinion towards the Agriculture Ministry is positive, negative, and neutral.
3	Hasan [18]	SVM	SentiWordNet	100000 tweets	Provides a comparison of techniques of sentiment analysis in the analysis of political views by applying supervised machine-learning algorithms such as Naïve Bayes and support vector machines (SVM).
4	Trupthi [22]	NLP + uni-word naïve bayes	Hadoop and MapReduce	20,00,000 tweets	Performed real time sentimental analysis on the tweets that are extracted from the twitter and provide time-based analytics to the user.
5	Soumya, S. [24]	1-D CNN, DNN	without	5468 Malayalam tweets	In this work the models are used to classify Malayalam tweets as positive and negative.
6	Kumar, S. [25]	LSTM, CNN	handcrafted	12922 Malayalam tweets	This work is first in its attempt to perform sentiment analysis of tweets in Malayalam language

PART 3

METHODOLOGY

This study is not an exception to the rule that procedures are required to analyze data to obtain useful insights. It is easier for consumers and businesses or institutions to communicate on Twitter.

Using Twitter as a free form of social interaction allows people to voice opinions on anything and everything [50][51]. This feedback, which may be favorable, unfavorable, or neutral, is based on the user's experience with the product or service in question. Identifying customer complaints about a product or service from user comments posted on the Twitter platform is critical to the success of the company's goods and services [62].

Consequently, analyzing comments from social media users is critical. May analyze microblog data to determine the polarity (negative, neutral, or positive) of user perceptions of a service or product using Opining Mining.

Analyzing comments left by users on various social media platforms serves as the basis for the sub-operations shown in Figure 3.1. Because of this set of sub-operations, we can determine the polarity of the social network text we are analyzing (positive, negative, or neutral). May use machine learning methods to uncover hidden information in the daily stream of social media posts [52] [54].

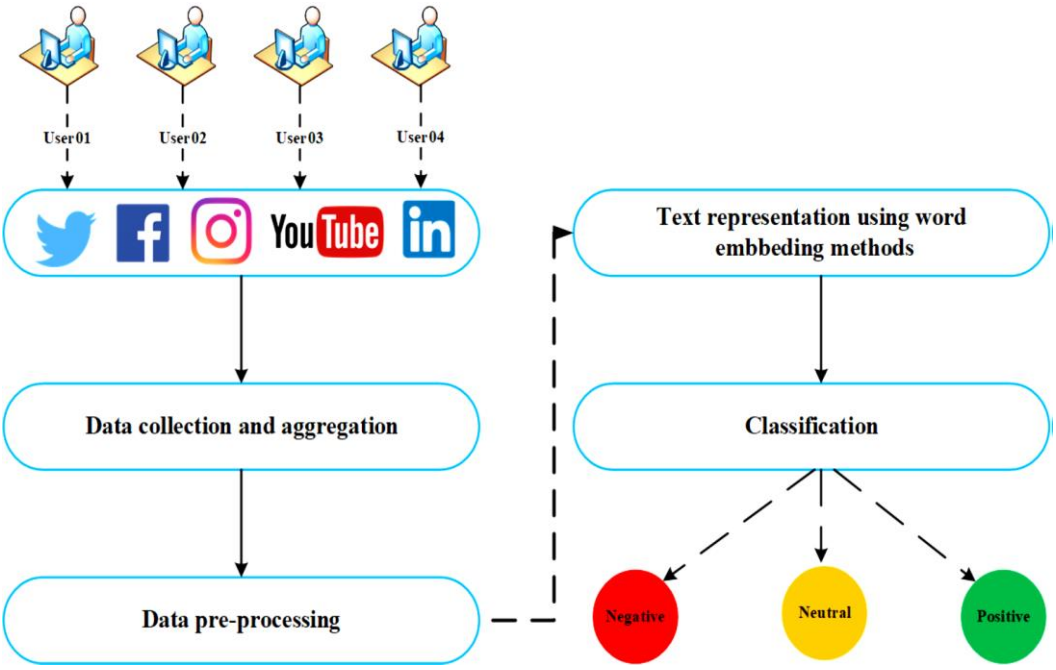

Figure 3.1 Basic steps of opinion mining on social network platforms.

Machine learning techniques all have their strengths and weaknesses. Superior performance is ensured by checking that your algorithm matches the assumptions and criteria. Regardless of the circumstances, no algorithm can be used. For instance:

Using a categorical dependent variable in linear regression is something may want to experiment with. Do not even bother trying! Getting will not recognize low values of adjusted statistics. SVM, Random Forest, and other algorithms such as Logistic Regression and Decision Trees should be used instead in these types of circumstances Read Essentials of Machine Learning Methods (Logistic Regression (LR)) to gain a basic understanding of these algorithms.

The approaches used may have a substantial effect on the result or performance of the project. Figure 3.2 illustrates the proposed methodology for sentiment analysis of tweets using machine learning.

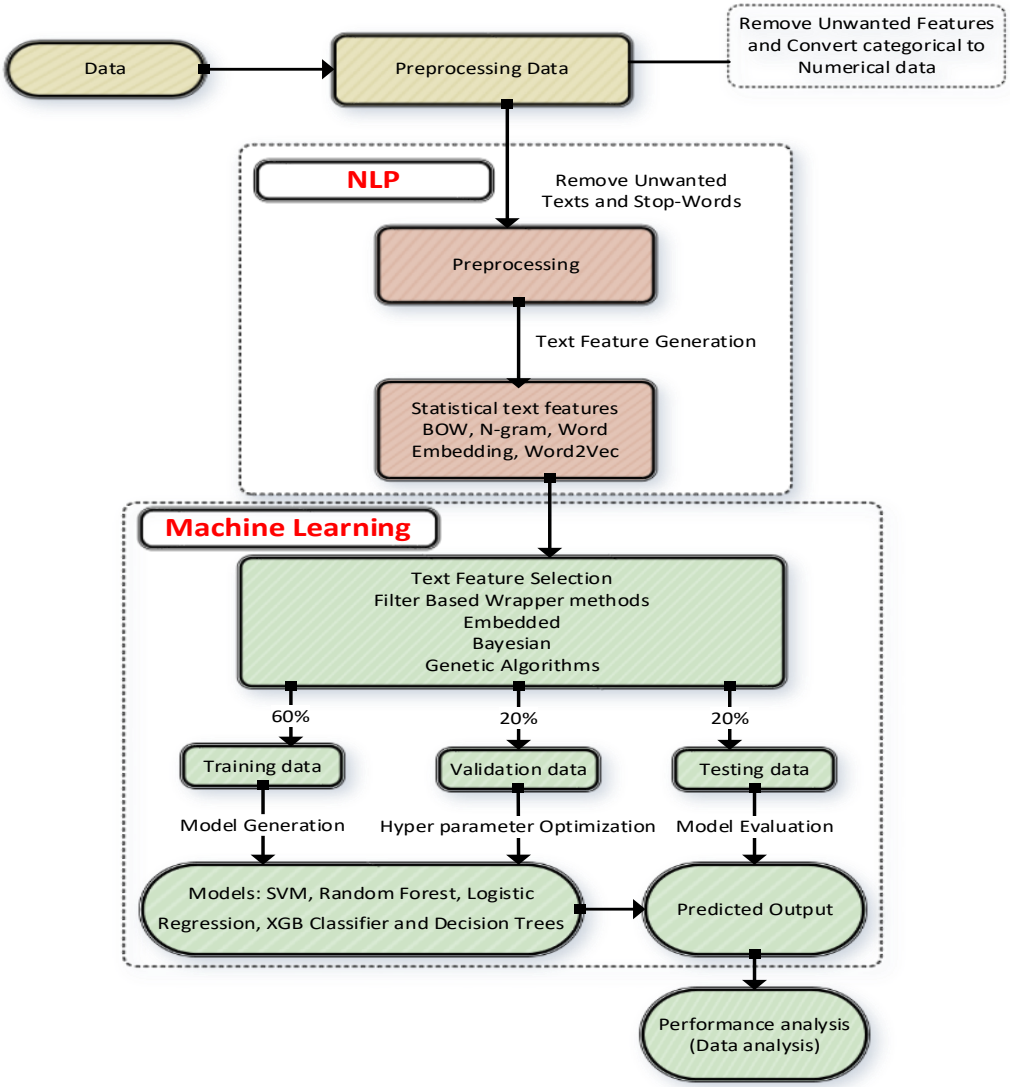

Figure 3. 2 Proposed methodology for sentiment analysis of tweets using Machine Learning

An N-gram is a succession of words in a phrase. N-gram is perhaps the simplest idea in machine learning. There are several types of N-gram utility. It may be used for automatic word correction, auto-spell check, and grammatical checks [48] [60].

Checking the link between words is also useful, particularly when attempting to predict what someone is going to say in order to assess the feelings or thoughts conveyed by the word. N-grams are the word combinations that are used together.

Unigrams are N-grams when N = 1. These are referred to as bigrams for N = 2 and trigrams for N = 3. N-grams capture the structure of the language and assist identify which word is likely to follow a word.

Using a variety of feature extraction approaches, we reached a similar conclusion. There are various methods for extracting feature information from text, but the Term Frequency (TF) and its Inverse Document Frequency (IDF), as well as word2vec and doc2vec, are among the most prominent. The authors discovered that using TF, IDF, and TF-IDF with linear classifiers like SVM, LR, and perception increased the accuracy of a native language recognition system. Ten different languages are used in cross-validation trials.

The TF-IDF is used to tags n-gram words, characters, and parts-of-speech tags. The TF-IDF weighting on features is better than other techniques when dealing with unigrams and bigrams of words. Similarly, the authors of [49], [61], and [67] looked at the performance of a neural network combined with three feature extraction algorithms for text analysis. TF-IDF and its two derivatives, Latent Semantic Analysis (LSA) and linear discriminant analysis, are used to assess the performance of each feature analysis technique (LDA). The findings show that the model's accuracy increases when using a large dataset with TF-IDF. For smaller datasets, combining TF-IDF with LSA gives equivalent accuracy.

Given a set of independent parameters, the LR approach may be used to determine a binary outcome (1 / 0, True / False, True / False). Dummy variables are employed to indicate binary/categorical outcomes [63] [64]. It may alternatively think of logistic regression as a kind of linear regression in which the outcome variable is categorical, and the dependent variable is the log of changes. In other words, data is fitted to a logit function to assess the probability of an event happening [65] [66].

3.1.DERIVATION OF LOGISTIC REGRESSION EQUATION

The Generalized Linear Model (GLM) includes Regression Techniques as a subclass (GLM). It was Nelder and Wedder burn in 1972 that developed this model as to apply

Linear regression to issues that were not naturally suited to it [55]. Logistic regression was included as instance in a class of models that includes other models (such as linear regression, ANOVA, and Poisson Regression). There are two parts to the linear regression model [56], [57].

$$G\,(E(y)) \;=\; \alpha \;+\; \beta x1 \;+\; \gamma x2 \tag{3.1}$$

Here, g () is the link function, E (y) is the expectation of target variable, and α + βx1 + γx2 is the linear predictor (α, β, γ to be predicted). The role of the link function is to 'link' the expectation of y to the linear predictor.

Important Points,Study variables aren't assumed to be linearly related in GLM. In the logit model, the link function and independent variables are assumed to have a linear relationship. The regression model does not have to be distributed in a typical. For parameter estimation, it does not employ OLS (Ordinary Least Squares). Maximum Likelihood Estimation is used in its place (MLE) must not spread errors must typically [68] [69].

This is a basic linear regression equation with the dependent variable wrapped in the link function, to begin with, the logistic regression.

$$G\,(y) \;=\; \beta o \;+\; \beta(Age) \tag{3.2}$$

For ease of understanding, we considered 'Age' as the independent variable.
It is all about probabilities in logistic regression (success or failure). g () is the link function, as mentioned above. The probability of success (p) and the probability of failure (f) are used to calculate this function $(1-p).p$ must satisfy the following requirements:

Since p, $>=$ 0, it must always be positive.

In other words, it can never be more than1, since p is smaller than 1.

We can get to the heart of logistic regression by meeting these two requirements. We begin by denoting the link function as g () with 'p' and finally derive it. The exponential

Version of the linear equation is used since probability must always be positive [70] [72]. The exponent of this equation will never be negative for any combination of slope and dependent variable.

$$p = exp(\beta o + \beta(Age)) = e\text{^}(\beta o + \beta(Age)) \quad (3.3)$$

To make the probability less than 1, we must divide p by a number greater than p. This can simply be done by:

$$p = exp(\beta o + \beta(Age)) / exp(\beta o + \beta(Age)) + 1$$
$$= e\text{^}(\beta o + \beta(Age)) / e\text{^}(\beta o + \beta(Age))$$
$$+ 1 \quad (3.4)$$

Using (a), (b), and (c), we can redefine the probability as:

$$p = e\text{^}y / 1 + e\text{^}y \quad (3.5)$$

where p is the probability of success. *This (d) is the Logit Function*

If p is the probability of success, 1-p will be the probability of failure, which can be written as:

$$q = 1 - p = 1 - (e\text{^}y / 1 + e\text{^}y) \quad (3.6)$$

where q is the probability of failure.
On dividing, (d) / (e), we get,

$$\frac{P}{1-P} = e^y \quad (3.7)$$

After taking log on both sides, we get,

$$log\left[\frac{p}{1} - p\right] = y \quad (3.8)$$

Log (p/1-p) is the link function. Logarithmic transformation on the outcome variable allows us to model a non-linear association linearly. After substituting the value of y, we get:

$$log\left[\frac{p}{1} - p\right] = \beta_0 + \beta_{age} \tag{3.9}$$

In Logistic Regression, this is the formula. There is an odd ratio in this situation (p/1-p). When the log of probability value is positive, there is always a greater than 50% chance of success [58]. Below is an example of a logistic model visualization. The likelihood will never fall below () or rise above one, as shown in Figure 3.3.

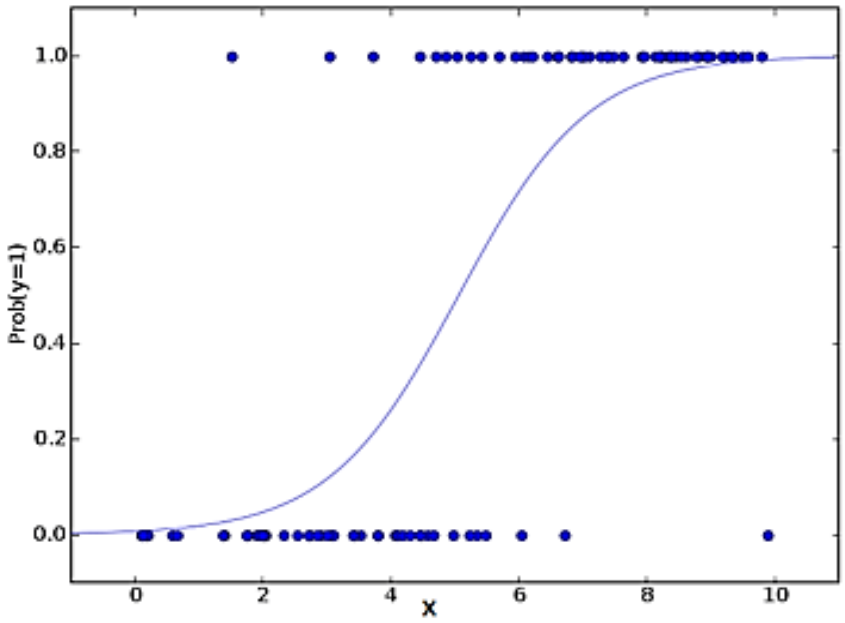

Figure 3. 3 Performance of Logistic Regression Model

3.2. TEXT FEATURES

Textual data formats are often incompatible with machine learning models. Machine learning models can only be compacted using numerical values. There are methods to translate this text data into numeric characteristics without sacrificing the data's significance. The following are examples of feature selection methods [71]:

3.2.1. The Word2Vec Model

Word2vec is a technique for efficiently generating word embeddings. Google developed this predictive deep learning-based model in 2013 to calculate and construct high-quality, distributed, and continuous dense vector representations of words that capture environmental and semantic similarities [59] [73]. It is a form of unsupervised model that uses a huge corpus of words to build a vocabulary of words and dense word vectors for each word.

The words are converted into vectors so that machine learning algorithms may conduct algebra operations on numbers as opposed to words. This change is known as word embedding [54][74]. With diffused Hypothesis in Word2Vec, a word's lexicon is in its nearby words. A word may be anticipated based on its proximity to these terms.

3.2.2. Continuous Bag of Words

In the CBOW approach, the framework attempts to predict the current target word (often the middle word) based on the source context words, which are the neighboring words [49] [53]. The corpus is constructed such that each unique word in the dictionary may be extracted and mapped to a unique number identity. Kera preprocessing is the most often used Python package. The CBOW generator is then constructed with two variables: context and target. The CBOW model's deep learning architecture is constructed using Keras and Tensorflow. This model tends to do better with smaller data sets. It is quite quick to train the model, providing greater precision [75] [76].

3.2.3. Continuous Skip-Gram Model

This is the opposite of CBOW in that it uses the current target words to forecast the words immediately around them. A bigger data set improves the performance of this model. Predicting the meaning of a word's context is the goal. A corpus dictionary is constructed in such a way that each unique word in the dictionary may be retrieved and given a unique identity for use in this model. We also keep track of the mappings

That convert words into and out of their unique IDs. Next, a skip-gram generator is created, which outputs the relevant pair of words. On top of TensorFlow, Keras is used to create the skip-gram model. In order to recover the encoded words, a model must first be trained [53] [77].

It is possible to get Twitter data in several different methods. Python modules that can be used to extract tweets are the topic of this study, some of which include:

A. Tweepy

A Strem-Listener may be used to get tweets from this module. Tweepy's OAuthHandler is used to get access to the API. The OAuthHandler receives the consumer key and consumer secret key from Twitter to provide access to the user's data.

Users may choose what information they want to see, and only tweets that include this information are returned. Adding terms as example "coronavirus," "corona," "COVID," "social distancing," and the like to a list can allow a user to only see tweets that include that information, for example. Here is the tweepy documentation (www.tweepy.org) [78]. Installing tweepy in Python is as simple as typing the following code:

'pip install tweepy'.

Using tweepy has the disadvantage of being limited to seven days of data extraction. Twitter-scaper and other modules may be used to extract data from Twitter. Assembling this study's data was a snap using the twarc module in Python.

Table 3. 1 The Root-level and Child Attributes of Twitter Data

Attribute	Type	Description
created_at	String	The time stamp on this Tweet is UTC. For instance, "created at" maybe "Wed Oct 10 20:19:24 +0000 2021."
Id	Int64	This is the unique identification for the Tweet in integer form. An unsigned 64-bit integer is acceptable since certain computer languages may have difficulty deciphering this identification.
text	String	This is the text or conversation by the twitter user.
User	User Object	The person who tweeted it. See the link for further information on each characteristic in the user data dictionary.
Re-tweet count	Int	Retweets this tweet has accumulated.
lang	String	The language of the tweet is detected by this.

B. Twarc

Twitter data may be retrieved and archived using this command line program and a Python package. The code for its installation is as follows:

'pip install twarc'

If already have python 2.7 or a higher version installed, you may use it as a command tool in PowerShell. It has to be set up once it's been installed. There are two ways to setup twarc: PowerShell or the command - line interface.

'twarc configure'

A set of secret keys is needed to grant permission. If you choose 'Authorize app,' Twitter will be linked to twarc so that the APIs may be used to retrieve tweets. The tweet extraction may begin when all the procedures are followed.

The Zenodo database was used to retrieve the tweet ids of Twitter users who discussed the epidemic and its connection to mental health as part of this study. Each month, the tweet ids are classified to make the data easier to work with. Tweet ids are entered into a command line, which extracts the tweets (conversations) from these ids.

$$'twarc\ hydrate\ tweet\_ids.txt\ >\ tweets\_hydrated.jsonl'.$$

If the tweet ID remains live on Twitter, the data extracts all of the tweet's root and child properties. Open the git bash program and run the following line of code to extract the relevant columns from the file: id, text, date, and location.

$$awk\ -F\ "," "\{print\ \$1\ \$2\ \$4\ \$16\}"\ >\ output.txt$$

Positions 1, 2, 4, and 16 correspond to the data's text, id, and date/time. It is necessary to transform the JSON file into a text file. The text file is parsed into a data frame that includes the id, text, date, and location using python. Here, you may find the python code titled 'extract combine and convert to data frame' here. Unstructured data is stored as a text file with three columns (id, text, and date) in order to be processed in the following step.

3.3. TWEETS PREPROCESSING AND CLEANING

The most critical phase in data mining is preprocessing, which transforms and prepares datasets for knowledge extraction [78]-[80]. Preprocessing is a broad term that encompasses a variety of methods. The dataset is being cleaned, integrated, transformed, and reduced using some of these techniques.

Modeling may be done using the structured/clean data that is produced by this method. Analysis can't take place without first cleaning raw data from many sources; hence the raw data must be filtered before analysis can begin. Data cleansing accounts for around 70% of all project effort in all analytical projects. It's a pain, but there's no way around it. From January 2021 to April 2022, data was gathered for this research.

Only the tweet id, the date of the tweets, and the content of the tweets were used in the data extraction. The text itself is being examined, not the audio or video.

A preprocessing step on dirty datasets ensures model acceptability. It comprises three columns: id, text and month of tweeting. The pre-processed data is placed in a new column called tidy text. Column text is used for preprocessing. Listed below are some of the steps we took to prepare our raw data.

As an example, the pictures below depict two scenarios of office space – one is untidy, and the other is clean and organized (See Figure 3.4).

Figure 3. 4 Example scenarios of office space

There is work that looking for in this office. Which of the following is the most probable situation in which have no trouble locating the document? The less cluttered one, of course, since everything has a designated place. There is a lot of overlap in the data-cleaning process. Finding the correct information is much simpler when data is organized systematically.

Preparing the text data is a necessary step to make it simpler to extract data from the text and use machine learning algorithms. If this step is skipped, the risk of dealing with unreliable and inconsistent data. Noise, such as punctuation, special characters, numerals, and phrases that do not have much weight about the content, should be removed from the tweets in this phase [81][82].

We want to use our Twitter text data to extract quantitative characteristics later. To generate this feature space, we use all terms that are unique in the whole dataset. The quality of our feature space will improve if we preprocess our data adequately. Here is a sample of the dataset that was utilized in this study. Id, label, and tweet make up the data's three columns. The binary target variable is labeled, and the tweet includes the tweets to be cleaned and preprocessed to prepare them for use.

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
1	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in urð□□±!!! ð□□□ð□□□ð□□□ð□□□ð□□¦ð□□¦ð□□¦
4	5	0	factsguide: society now #motivation

Clean is a function that takes in text and outputs text that is free of any punctuation and numeric symbols, as the name implies. We used it on the review’s column and added the cleaned text to a new column called 'Cleaned Reviews.

	review	Cleaned Reviews
0	I called because my food was cold and not done...	I called because my food was cold and not done...
1	OMG, hands down the best pizza I've had from D...	OMG hands down the best pizza I ve had from Do...
2	This Domino's has the best pizza delivery and ...	This Domino s has the best pizza delivery and ...
3	My Sweetheart & I are very pleased with the qu...	My Sweetheart I are very pleased with the qual...
4	I called to place an order, The lady answered ...	I called to place an order The lady answered a...

Great, look at the above image, all the special characters and the numbers are removed.

3.3.1. Removing Twitter Handles (@user)

There are various Twitter handles (@users) in the tweets, which is how Twitter users are recognized. We delete all these Twitter handles from the database. The first is a combined train and test set for simplicity. This saves time and effort by not having to repeat the same actions.

3.3.2. Stop-Words Removal

A stop-word is a term in English that conveys little or no meaningful information. Text preparation necessitates their removal. Every language's stop-words are listed on nltk. Look at the English stop-words.

```
print(stopwords.words('english'))
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours',
'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'its
elf', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these',
'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doin
g', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about',
'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'o
ut', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'al
l', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so',
'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o',
're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'has
n', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'sha
n', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

3.3.3. Remove the URLs from the Text

A python code like given below is used to delete the URLs from a text, as displayed.

$$def\ remove\_url(row):$$

$$txt\ =\ str(row['tidy\_text']).split('https')[0]$$

$$return\ txt$$

$$data['tidy\_text']\ =\ data.apply(remove\_url, axis\ =\ 1)$$

3.3.4. Remove the Special Characters

Punctuation, numerals, and other special characters are ineffective. Instead of including them in the work, it might be wiser to delete them as we did with our Twitter accounts. We will use spaces instead of characters and hashtags in this section.

$$data['tidy\_text']\ =\ data.tidy\_text.str.replace("?!,\ \&:;\%()","\ ",regex\ =\ True)$$

3.3.5. Remove All Usernames With @

To delete and change usernames, use the code below.

$$data\ ['tidy\_text']\ =\ data['text'].str.replace('@[\w:]*',")$$

3.3.6. Convert To Lowercase

There are no upper or lowercase letters. To prevent the repetition of words with varying case values, the next line of code converts the letters to lowercase.

$$data['tidy\_text']\ =\ data['tidy\_text'].apply(lambda\ x: x.lower())$$

3.3.7. Remove Numbers from Characters

The regular expression (regex) pattern d+ is used to delete all numeric values. The addition sign guarantees that a number with more than one digit, such as 10, is treated as a single number and not as two distinct ones. The following line of code removes all numerical data from the body of the document.

$$data['tidy\_text'] = data['tidy\_text'].str.replace('\d+', "").$$

3.3.8. Drop Null Values

There may be missing values in certain unstructured/raw data. The term "null value" is occasionally used to describe these types of values. The most popular terms are often used to fill in the blanks, or they are altogether omitted. Our models will suffer if we don't deal with null values since machine representations don't accept nulls. Null values were removed in our preprocessing using the following line of code:

$$data['tidy\_text] = data['tidy\_text].dropna(inplace = True)$$

3.3.9. Removal of Stop-Words

A stop-word is a term that does not contribute meaning to a statement and may thus be ignored or omitted without altering the sentence's meaning. [Refer to the list of pause words]. Although stop-words may be found in many languages, English stop-words are being used for this project. If you want to analyze a person's emotions, you must eliminate all of their stop-words from their data. Stop-words like [I, me, mine, myself, we, our, ours, you, your] are often used in written communication. Here is a collection of often used English stop words. Python library for natural language processing nltk is used to import stop-words from of the corpus of nltk stop-words.

$$import\ nltk$$
$$from\ nltk.corpus\ import\ stopwords$$
$$stop\_words = set(stopwords.words('english'))$$

The λ function may be used to remove stop - words from the text. Also, remove terms with at least four characters in order to focus on words that are relevant to the study.

$$data['tidy\_text'] = data['tidy\_text'].apply(lambda\ x: ''.join([w\ for\ w\ in\ x.split()\ if\ w\ not\ in\ stop\_words]))$$

$$data['tidy\_text'] = data['tidy\_text'].apply(lambda\ x: ''.join([w\ for\ w\ in\ x.split()\ if\ len(w) > 4]))$$

The uncleaned text and the tidy text may be found here, along with the whole python code for data preparation, which is shown in the top 10 data below in Figure 3.5.

Out[29]:

	id	text	month	tidy_text
0	1381308761533452291	Le #Portugal a d\u00e9cid\u00e9 de suspendre ...	Apr	portugal decide suspendre provenance bresiln c...
1	1381308765841002501	Tommorrow is an exciting day as the Island mo...	Apr	tommorrow exciting island moves stage reconnec...
2	1381308766482751493	Nos hacemos acopio de esta noticia publicada ...	Apr	hacemos acopio noticia publicada informacifn i...
3	1381308768764370945	401 people	Apr	people
4	1381308768898576387	@mybmc UK\u00a0Clinical Trial Confirms SaNOti...	Apr	ukaclinical trial confirms sanotizes breakthro...
5	1381308770404466688	Th\u00eam 1 ca\u00a0COVID-19	Apr	theam caacovid-
6	1381308773944414209	Matthew Hancock MP	Apr	matthew hancock
7	1381308773998936074	@xotep my co-worker got a cert saying he got ...	Apr	co-worker saying vaccine conformed covid-
8	1381308775118766087	Stay Safe Mumbai! Masks on and battle against...	Apr	mumbai masks battle covid support government m...

Figure 3. 5 Unstructured and Pre-processed data

3.3.10. Stemming

A word's suffixes (such as "ing," "ly," "es," and "s") are removed using a set of rules called stemming. There are several distinct ways to say "play," such as "player," "played," "plays," and "playing." The stem of a word is the component that conveys its meaning. Stems and tokenization are two typical methods for locating root/stem words. For example, stemming often results in useless root words since it merely removes certain letters at the end of the process. Here's an example of how stemming and tokenization vary.

```
Text: He glanced up from his computer when she came into his office
Stem: ['glanc', 'comput', 'came', 'offic']
Lemma: ['glance', 'computer', 'come', 'office']
```

The result of Stemming is Stem, and the output of Lemmatization is Lemma, as seen in the preceding example. There is no grammatical significance to the stem glance in the word glanced. The Lemma look, on the other hand, is flawless. Steps 2-5 were now clear to us after using basic examples. Allow me to bring the conversation back to where it started the root of the issue.

3.3.11. Evaluation Performance

There are a few indicators to examine when evaluating the efficiency of a logistic regression model.

Logistic regression's modified R2 has an analog in AIC (Akaike Information Criteria). The number of model coefficients is considered while calculating the AIC, which is a measure of model fit. As a result, we always choose a model with a low AIC score above anything else. Null Deviance is the response predicted by a model with just an intercept, whereas Residual Deviance shows the actual response.

The better the model, the lower the value. The response anticipated by a model when independent variables are included is known as residual deviation. The better the model, the lower the value.

As the name suggests, the Confusion Matrix depicts the difference between actual and predicted values. For example, we may use this method to determine the model's accuracy and prevent overfitting shown in Figure 3.6.

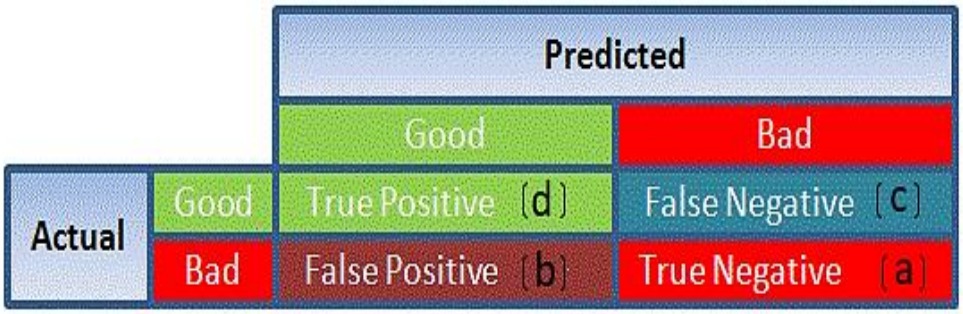

Figure 3. 6 Evaluation performance

To calculate the **accuracy** of the model is:

$$\frac{True\ Positive+True\ Negatives}{True\ Positive+True\ Negatives+False\ Positive+False\ Negatives} \tag{3.10}$$

From the confusion matrix, Specificity and Sensitivity can be derived as illustrated below:

$$sum\ to\ 1 \begin{cases} True\ Negatives\ Rate\ (TNR), Specificity = \frac{A}{A+B} \\ False\ Positive\ Rate\ (FPR), 1 - Specificity = \frac{B}{A+B} \end{cases} \tag{3.11}$$

$$sum\ to\ 1 \begin{cases} True\ Positive\ Rate\ (TPR), Specificity = \frac{D}{C+D} \\ False\ Negatives\ Rate\ (FNR), 1 - Specificity = \frac{C}{C+D} \end{cases} \tag{3.12}$$

PART 4

RESULTS AND DISCUSSIONS

Twitter data extraction and preprocessing are described in Part 3 of the thesis. Natural language processing models for statistical text characteristics presented in Part 2, are supplied into the preprocessed data. A variety of text feature selection approaches are used to extract the most valuable and accurate characteristics from the datasets. Filtering out unnecessary data is the goal of feature selection. The model's accuracy increases, and its training time is sped up if the proper subset is selected. Methods for selecting features include filtering, wrapper methods, and embedding.

4.1. DATA COLLECTION

Nowadays, people choose to express themselves through social media platforms such as Facebook, Twitter, TikTok, and others, rather than face-to-face interactions. For this study, Twitter data relating to the pandemic was gathered from the tweets of those who had been affected by it. The acquisition of Twitter data was the subject of a variety of studies. Unless you are a Twitter developer, you will not be able to access any of Twitter's data. This may be done by filling out the application and submitting the necessary information.

After approval, which may take anywhere from 24 to 48 hours, a user's profile is established, and API access is made available to them. The token, secret key and secret token may all be retrieved from the Twitter developer account profile, as can the consumer key and secret key. You can't get into any Twitter data until you've got these keys. A JSON file is often used to store the retrieved information. Root-level properties and child objects are often seen in the JSON file.

4.2. DATA EXPLORATION

Observations were made in the months of January and May of the following year. There are thousands of rows of data in all after cleaning. On a monthly basis, the number of tweets is plotted to observe how much debate there is on the categorization of tweets and its influence on social media.

A look at how the discussions progressed is shown in the following graph (Figure 4.2). It's a constant dialogue from January 2021 through April 2022. Much discussion occurred during this time period about the pandemic's severe impacts on employment, homeschooling, and social isolation. When vaccines were developed, the topic of vaccinations became a major topic of discussion. More nations opened their doors to the vaccination in February 2022, and a number of incentives were put in place, particularly in North American countries, to urge their citizens to get the vaccine in order to keep people safe and open their country to international communities. Covid-19, mental health, immunization, and the many kinds of authorized vaccinations were all discussion topics.

There was much discussion on how things are becoming better since more jobs are being generated, and people's living standards are rising as more individuals get completely immunized and no longer worry about catching the virus. The dread of being isolated decreases. Isolation in most nations has also decreased significantly.

4.3. UNDERSTANDING THE COMMON WORDS USED IN THE TWEETS: WORD-CLOUD

This phase examines the training dataset to determine how evenly distributed the supplied emotions are. Plotting word clouds might help to identify the most frequently used terms. The most frequently used terms are shown in the largest font in a word cloud, while the less frequently used words appear in the smallest font (See Figures 4.1, 4.7, and 4.3). The Word cloud of frequent words is classified as Positive and Negative Sentiments.

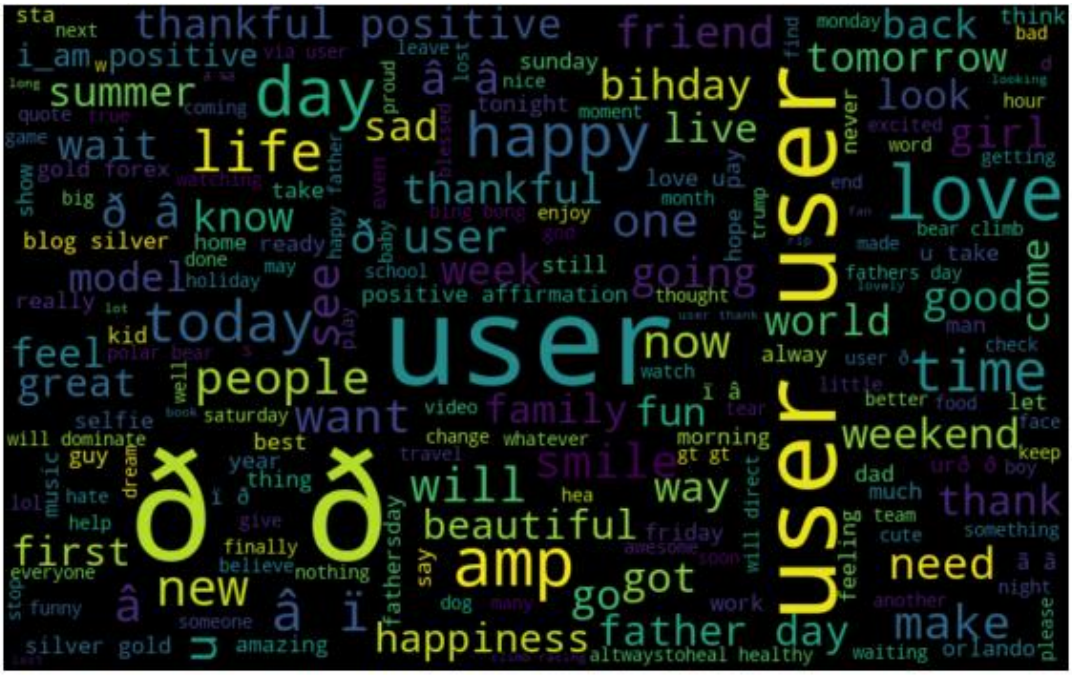

Figure 4. 1 Neutral sentiments

Most of the terms are either positive or neutral, as can be seen. With the most common ones being happiness and love. It does not offer us an insight into the language related to racist/sexist tweets. As a result, in Figure 4.2 we have distinct textual data for both the racist and sexist categories in our train data.

Figure 4. 2 Words in non-racist/sexist tweets (Positive sentiments)

Most of the terms are either positive or neutral, as can be seen. Love and happiness are the most often used expressions. As a result, most of the most often used terms are consistent with the mood of non-racist/sexist tweets. Figure 4.3 shows the other sentiment's word cloud. Negative, racist, and sexist phrases are to be expected.

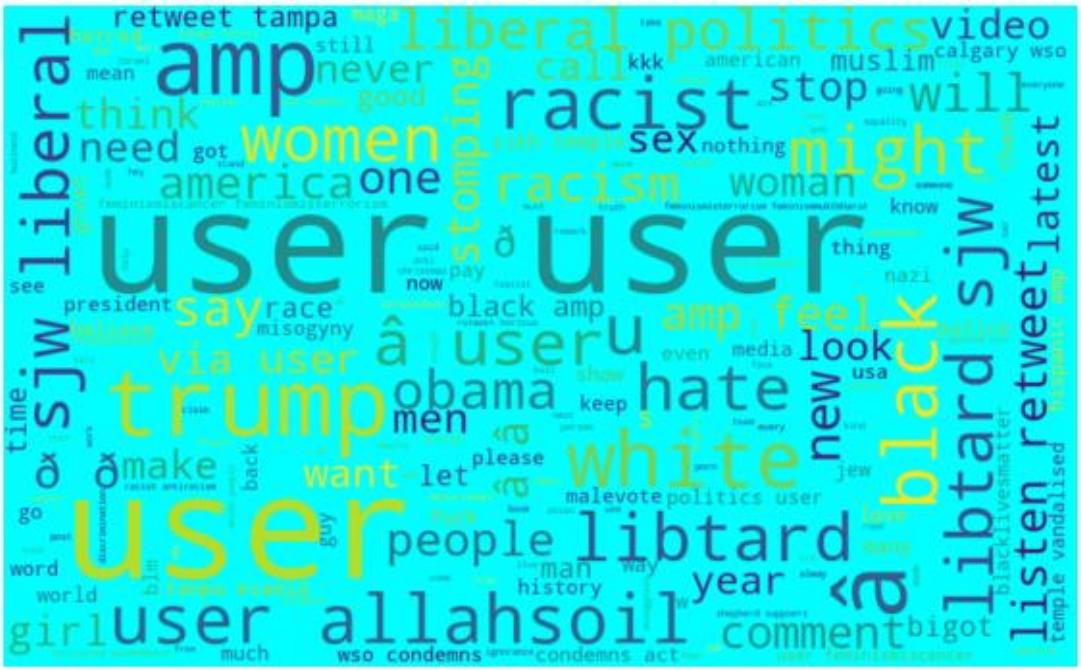

Figure 4. 3 Racist/Sexist Tweets

Most of the terms have negative meanings, as can be seen from the list above. In other words, it seems that the test data we have is of high quality. After that, we study the most popular Twitter hashtags and trends. Figure 4.4 displays the reviewers' most commonly used terms for the top 30 words.

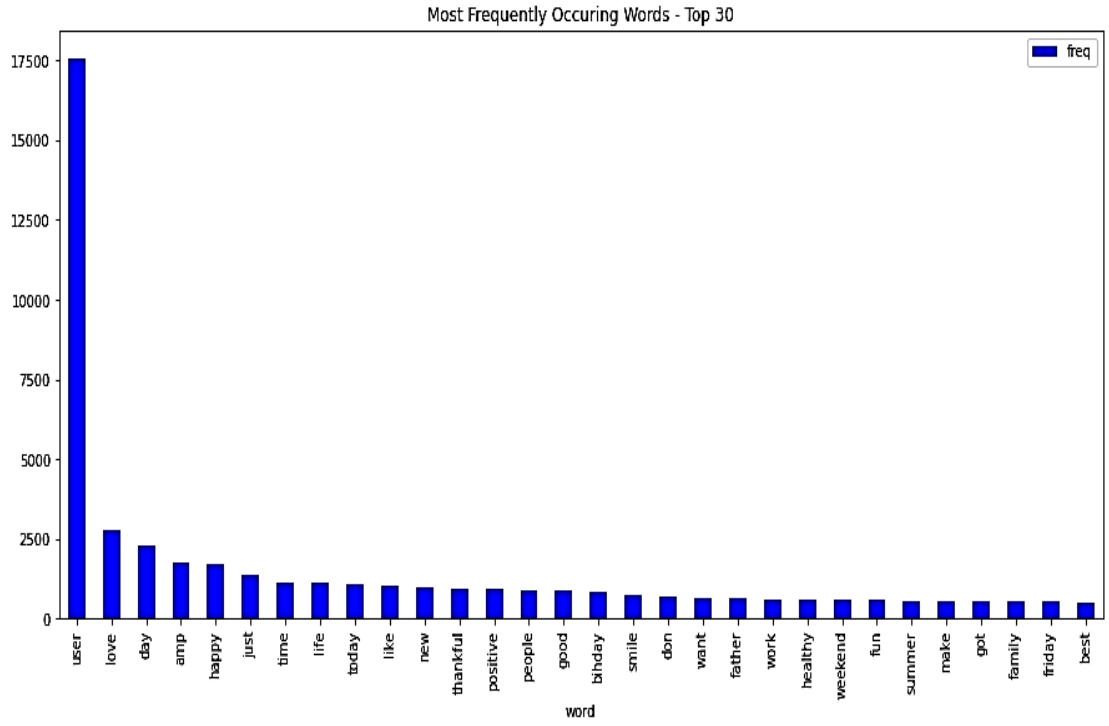

Figure 4. 4 Most frequently occurring words – Top 30

4.4. UNDERSTANDING THE IMPACT OF HASHTAGS ON TWEETS SENTIMENT

At any given period. Twitter's trending hashtags are identical to those hashtags. We need to see whether these hashtags can help us categorize tweets into distinct feelings, or if they are just a waste of time. A tweet from our dataset is shown below:

"what has today's attitude to women got in common with that of norman bates? #psycho #feminism #hollaback"

To our opinion, the tweet is sexist, and the hashtags used to describe it are sexist as well (see Figure 4.5).

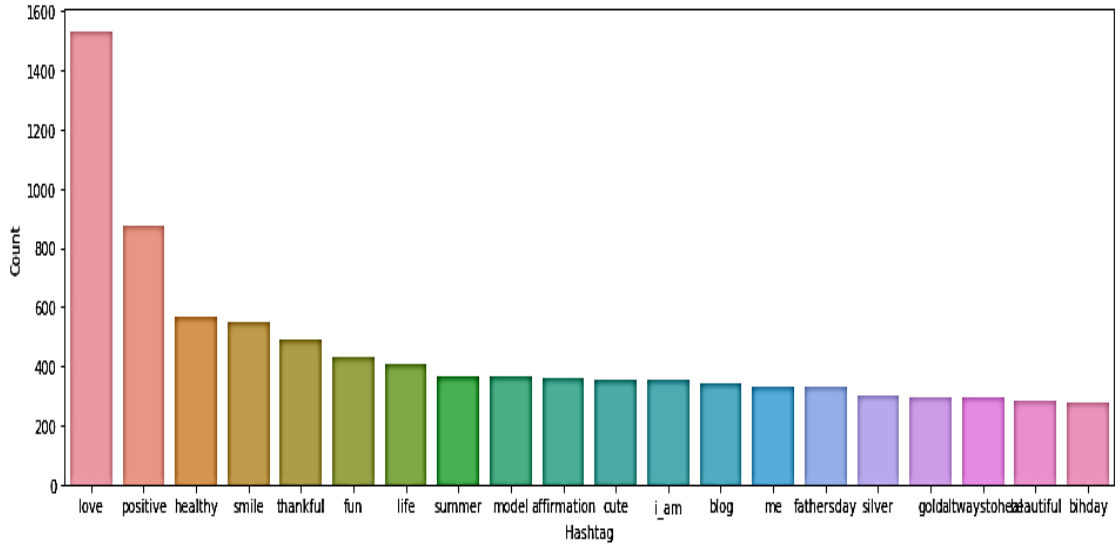

Figure 4. 5 Non-Racist/Sexist Tweets

It seems reasonable that all of these hashtags are good. The plot of the second list is expected to include negative words. Figure 4.6 shows the most frequently used hashtags in racist and sexist tweets.

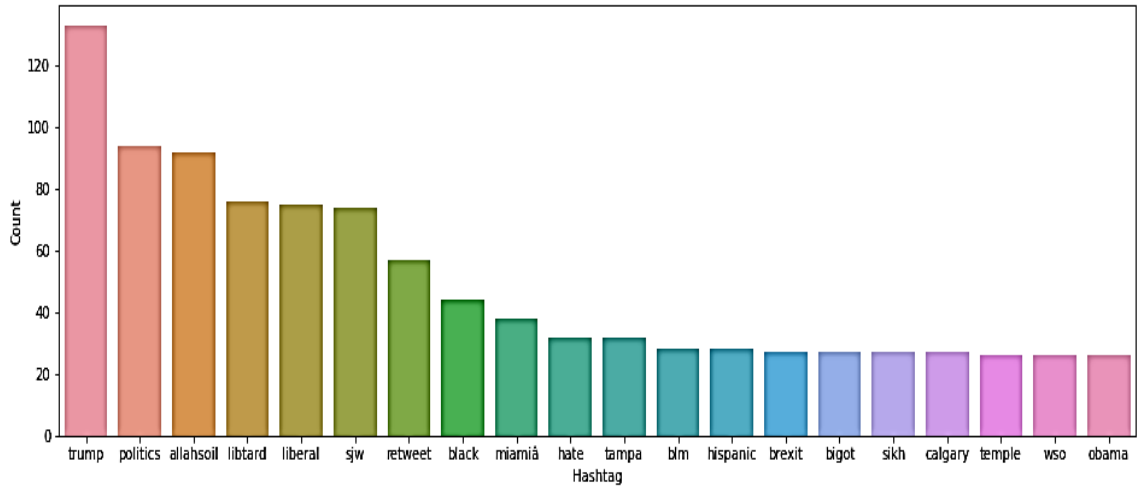

Figure 4. 6 Racist/Sexist Tweets

Most of the words are negative, although a handful is neutral. We should maintain these hashtags in our database since they provide essential information. Consequently, once the tweets have been tokenized, they can begin extracting characteristics from them.

4.5.EXTRACTING FEATURES FROM CLEANED TWEETS

Features must be generated in order to analyze pre-processed data. TF-IDF, Sentiment analysis, and Word vectors are a few of the methods that can be used to create text characteristics. There is no coverage of TF-IDF or Bag of Words in this work.

4.5.1. Term Frequency - (TF-IDF) Features

The bag-of-words method may be a useful place to start when studying a corpus, but it fails to account for the number of times a phrase occurs in a text or tweet [31]. Words that are frequent throughout the corpus yet often appear in a subset of texts gain from the method, which penalizes keywords that are common but seldom appear elsewhere.

The following definitions are crucial in the context of TF-IDF: As shown by the document's word count, It was necessary to extract features from both training and testing data for machine learning models to be trained, and the extracted features were utilized for categorization.

The TF-IDF score is a commonly used tool for information retrieval and summarization. According to its inventors, the TF-IDF measure is intended to illustrate how essential a statement is in a given text. TF-IDF is used to extract TF and IDF characteristics. IDF rewards those tokens the highest when a dataset has a limited number of tokens. If a unique phrase appears twice, it has a more significant impact on how each sentence should be interpreted.

For each word t, IDF is equal to the log (N/n), where, N is the number of documents, and n is how many times it has been in each document.

$$TF\_IDF = TF * IDF \tag{4.1}$$

4.6. MODEL BUILDING: SENTIMENT ANALYSIS

To collect the data in a usable manner, we have completed all pre-modeling steps. TF-IDF and W2V can be used to develop prediction models on the dataset. To create the models, we employed logistic regression. Using a logit function, it estimates the chance of an event occurring. In Logistic Regression, the following equation is used:
We used the W2V and TF-IDF features to train the machine learning algorithms (Logistic Regression, Random Forest Classifier, Decision Tree Classifier, SVM, and XGB Classifier) and it returned a training and validation accuracy, and F1-score on the validation set. Now that we have a model, we can use it to predict test results. Table 4.1 shows the performance evaluation of classification between different machine learning approaches.

Table 4. 1 Evaluation performance between different machine learning approaches

Methods	Range	Max Features	Training Accuracy	Validation Accuracy	F1 score
Logistic Regression	31962	2500	0.9951	0.9616	0.6133
Decision Tree Classifier	31962	2500	0.9991	0.9317	0.5321
Random Forest Classifier	31962	2500	0.999	0.9519	0.6089
SVC	31962	2500	0.9781	0.9521	0.4986
XGB Classifier	31962	2500	0.9445	0.9433	0.3537

As statistically shown in Table 1, the result of LR is more accurate than other machine learning classifiers, which is a validation accuracy is 0.9616 and an F1 score of 0.6133 with feature extraction (TF-IDF).

4.7. TEXT FEATURE SELECTION

4.7.1. Filter Methods

The filter technique uses univariate metrics to rank a dataset's characteristics. Finally, it chooses those attributes that have received the most votes. Following is a brief description of some of the filtering techniques used in this research project.

4.7.2. Information Gain

The scikits module in python is the features extraction package from which the information gain module is loaded. For example, it establishes the method by which one predicts the other. Using the train data from our experiment, the mutual information function was performed to determine which word had the most sentimental connotations. The feelings of the words in the dataset are shown in the bar graph by combining words.

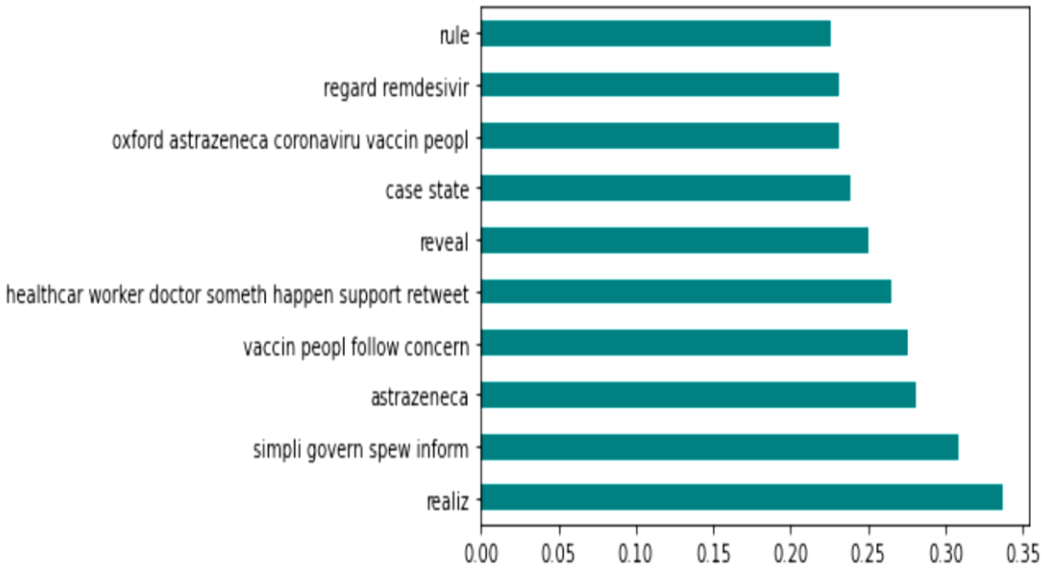

Figure 4. 7 Text Feature Selection Utilizing Information Gain

4.7.3. Fisher's Score

Text characteristics may be prioritized according to their relevance using this Text Feature Selection method. The terms in the following list are ranked the same way, depending on their relative significance. The fisher criteria select characteristics based on their scores. This results in a subset of characteristics that are not ideal.

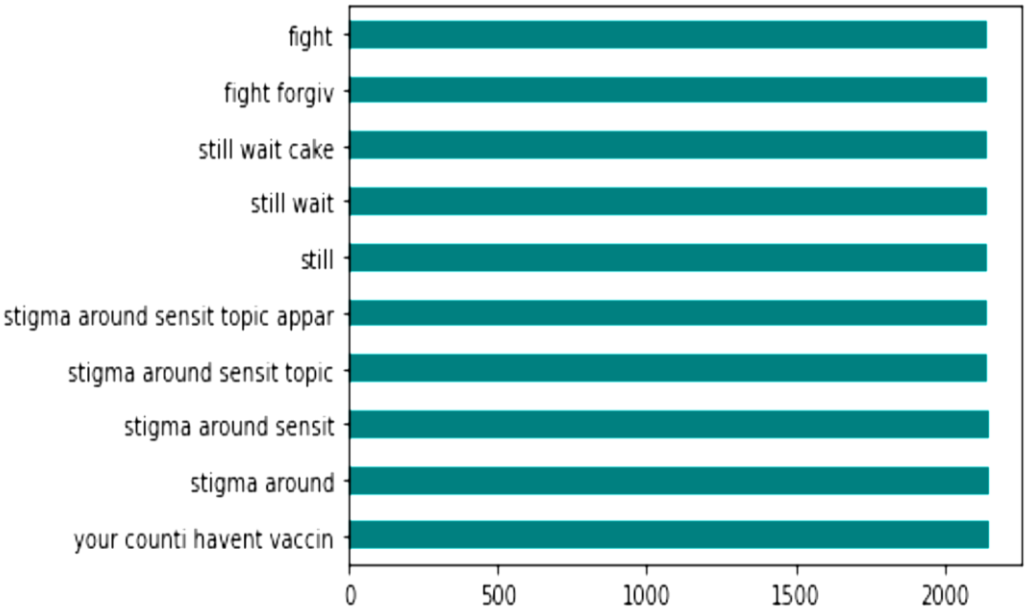

Figure 4. 8 Text Feature Selection Utilizing Fisher's Score

4.7.4. Chi-Square Test

The relationship between two things is explored in order to figure out how they vary from one another. This function eliminates characteristics that are most likely to be independent and so unimportant for classification. We used this test and our test results to construct the graph shown below.

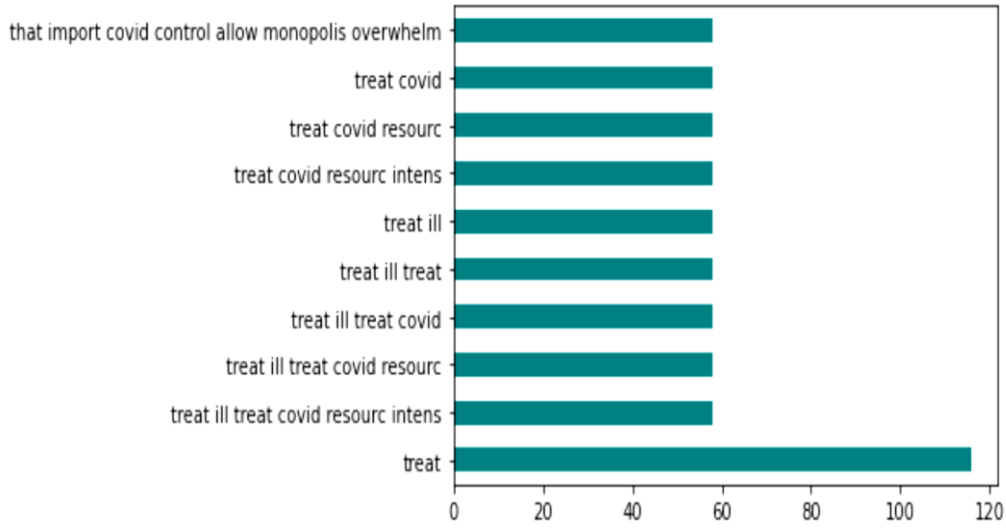

Figure 4. 9 Text Feature Selection Utilizing Chi-Square Test

4.7.5. Mean Absolute Difference (MAD)

While the variance applying dynamic includes and square, this method excludes the latter. Mean absolute difference from the mean value for a particular feature is calculated using this scaled variation. This is shown in Figure 4.10, where the data sample is used as an example of this strategy.

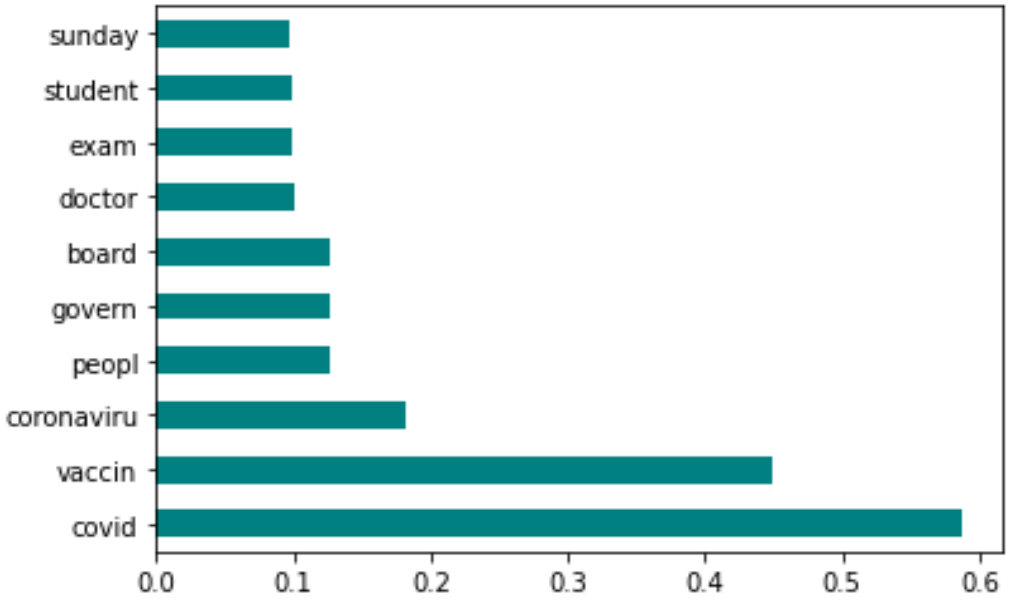

Figure 4. 10 Text Feature Selection Utilizing Mean Absolute Difference

4.7.6. Dispersion Ratio

The more dispersed a trait is, the more important it is. The dispersion ratio is derived by dividing the arithmetic average and the geometrical mean for the given feature. The term "tweets classification" has the greatest ratio in our sample data using this approach, as seen on the graph.

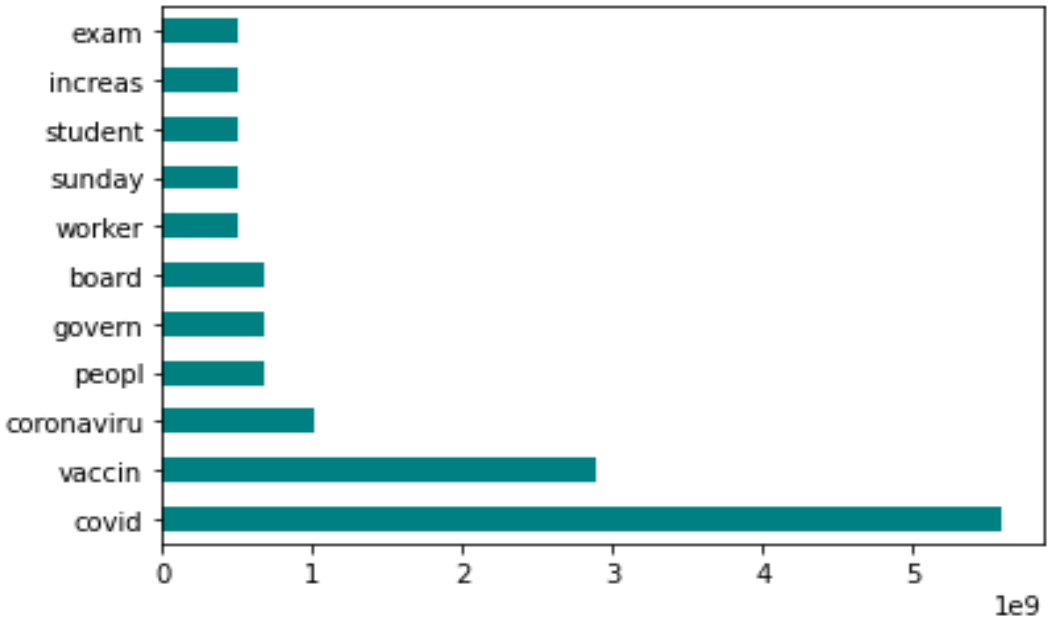

Figure 4. 11 Text Feature Selection Utilizing Dispersion Ratio

4.8. WRAPPER METHODS

The algorithm is trained using a subset of the characteristics in an iterative [27]. The algorithm considers all of the features that are available. The model is repeated till the accuracy of the model is good based on the user of the model. Finally, the best characteristics are chosen from the model. Recursive Feature Elimination (RFE) with random forest and Forward Feature Selection (FFS) are some wrapping approaches used.

4.8.1. Recursive Feature Elimination (RFE) with Random Forest

For its simplicity of use, flexibility in design, and ability to choose the essential characteristics for predicting the target variable, RFE is a popular feature selection method. The selection of relevant dataset columns and the technique for selecting these columns are two of the most critical aspects of RFE.

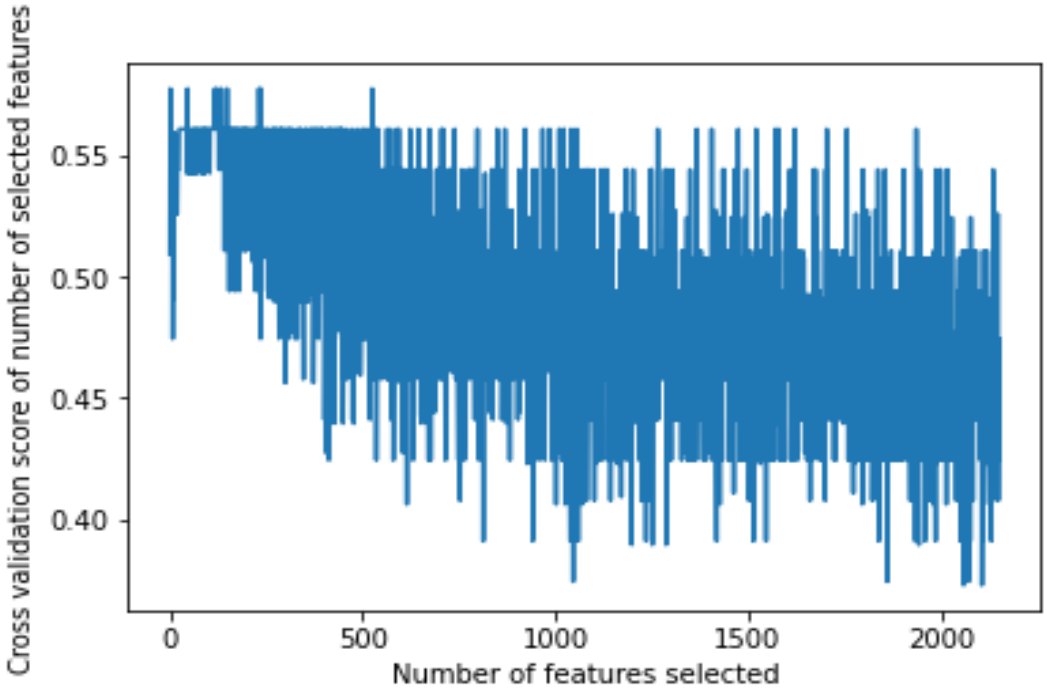

Figure 4. 12 Recursive Feature Elimination (RFE) with Random Forest

4.8.2. Forward Feature Selection

In order to begin the feature selection process, the algorithm begins with a set of features that are completely empty. Every time a new feature is added, the process repeats itself until the model's performance does not increase. Backward reduction, bi-directional elimination, comprehensive selection, and recursive elimination are some of the other methods that may be used here.

4.8.3. Embedded Methods

The filter and wrapper methods of feature selection are combined in this approach. Here, the algorithm contains built-in feature selection algorithms and evaluates a mixture of features. They are as quick as filter approaches, but they provide more precise results. Other methods used include the regularization methods (e.g., L1 and L2) and the tree-based approach, which employs feature significance to select features.

4.9. FEATURE ENGINEERING

It's, therefore, necessary to execute some feature extraction to convert the dataset features into matrices and create new dataset features. Below, we'll take a closer look at some of these features:

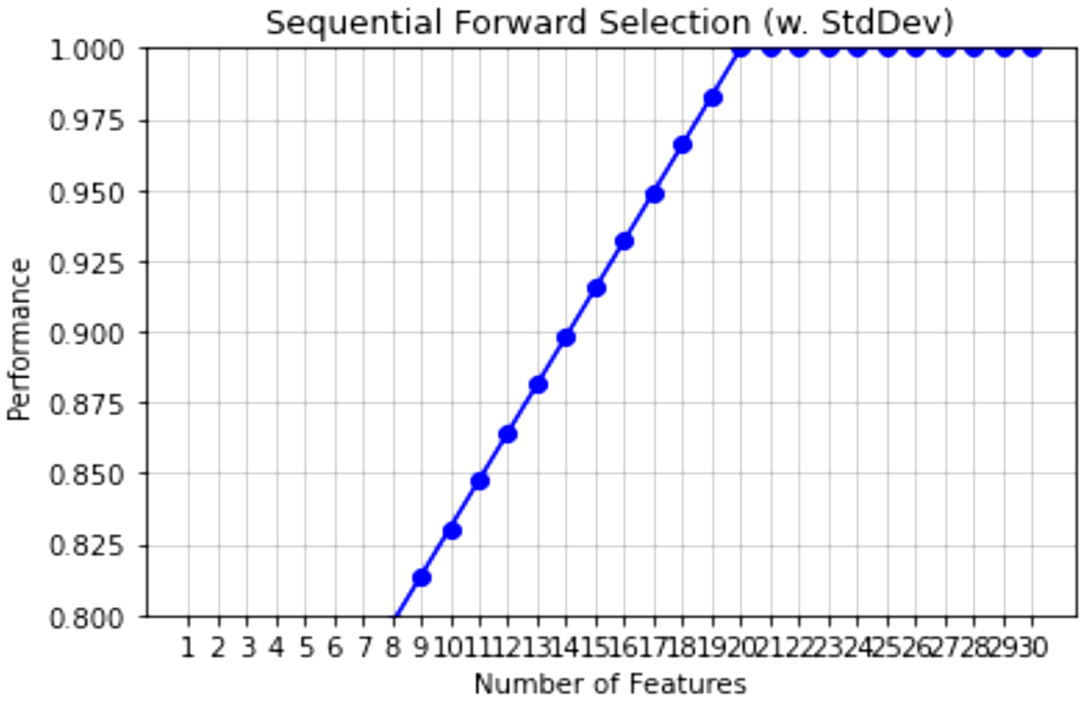

Figure 4. 13 Forward Feature Selection

4.9.1. Count Vectors as features

Numbers 1 and 0 represent text based on their location in the document. If a feature has text, the value is 1. Otherwise, it is 0. Every time the word appears, the counter goes up by one, leaving a 0 in all other places. Encoding in a single step is also known as one-Hot Encoding. A python tool called Count-Victimizer within the Scikit Learning activity may aid with this features extraction strategy since human efforts will be laborious.

4.9.2. Text / NLP-based features

In order to strengthen the text categorization model, more characteristics have been included here. Among the metrics derived from our dataset are the number of words, the number of characters, and the sentence length (or "word density").

4.10. Sentiment Classification

The tweet id, the date of the tweets, and the text that represents the dialogue in the tweets have been retrieved from the Twitter data. The data has already been analyzed. Natural language processing was used to eliminate all instances of stop words, special characters, and extraneous terms. Text characteristics have been extracted and prioritized for inclusion in the final product to ensure that the models execute as accurately as possible. A process known as Feature Engineering was used to turn the features into vectors. The accuracy, the F1 score, precision, recall, and kappa parameters of the scikit learn metrics were evaluated in the performance assessment. These settings were tested with 5 models. Figure 4.14 displays the metrics' outcomes for the various models.

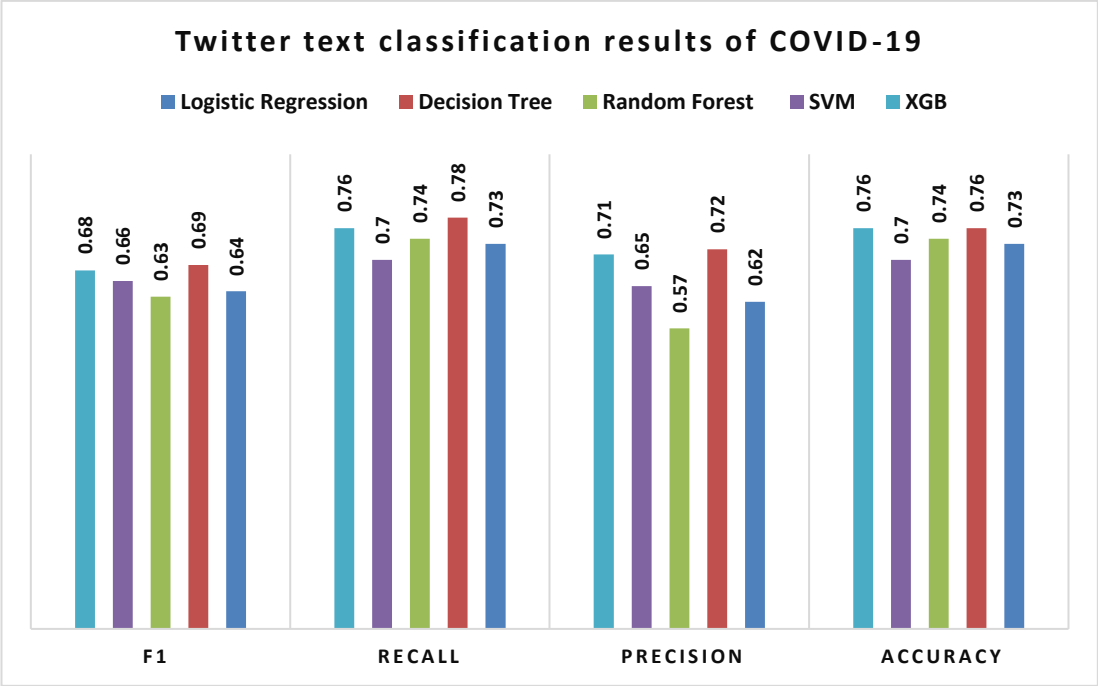

Figure 4. 14 Twitter Text Classification Results of COVID-19

4.10.1. Performance Evaluation with Supervised Learning Classification

Supervised learning algorithms are used to build models based on labeled data. The model was built using the dataset's labeled portion. Compared to actual results, the model classified and predicted both the train and the test score. Sustained learning classifiers are shown in the predicted scores in Figure 4.15 Logistic Regression Classifier is the best predictor for both train and test scores on the collected tweets dataset.

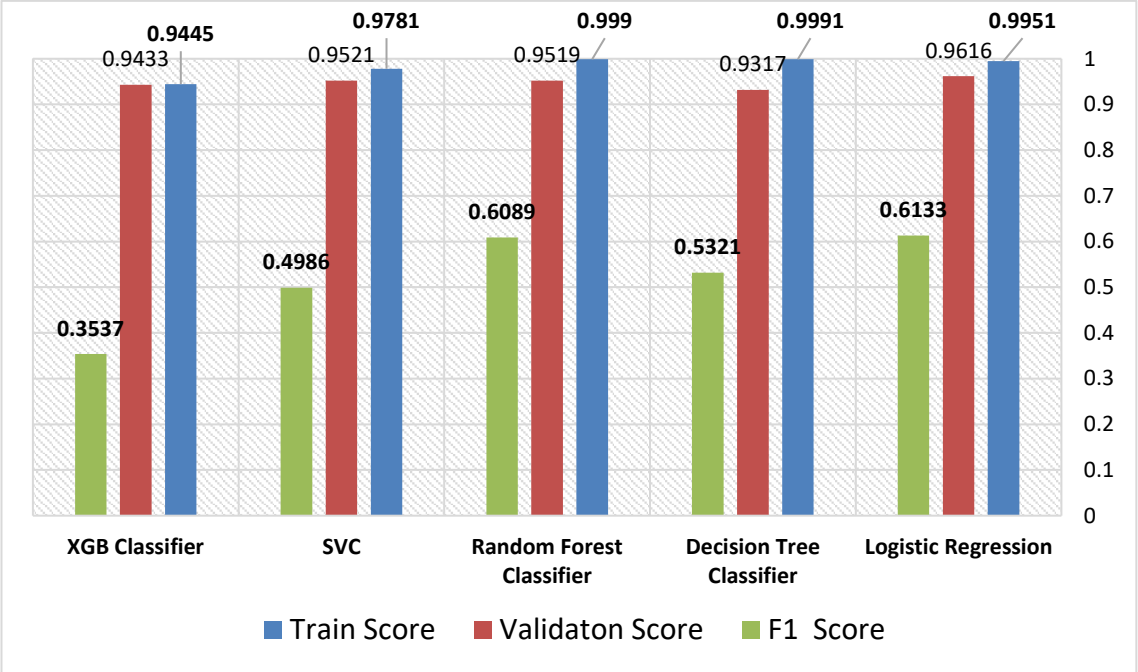

Figure 4. 15 Performance Evaluation for Supervised Learning

The Logistic Regression classifier had the most outstanding performance out of the supervised learning models, with 0.80. In other words, the model accurately predicted almost 80% of the adjusted test result.

PART 5

CONCLUSION AND FUTURE WORKS

Data has progressed to the point where significant insights may be gleaned from it to aid decision-making. Because of the vast amount of data gleaned from Twitter, this study focuses on that social media platform. According to a recent study, there are an estimated 200 million active Twitter users every day. Users' perspectives on the situation during the pandemic led to a large amount of data being acquired, evaluated using machine learning, and used to assist develop long-term remedies for the pandemic's influence on mental health. The sentiments of people's opinions were retrieved from tweets using machine learning. Positive and negative feelings were distinguished among the comments. Different models were created using machine learning in a semi-supervised and supervised learning environment.

A vote classifier based on logistic regression is proposed in this study. To integrate the likelihood of LR and TF-IDF, soft voting is used. Sentiment analysis may also be performed using different machine learning-based text categorization approaches. Users worldwide contributed to a Twitter dataset used for the research, examining the influence of feature extraction strategies such as TF-IDF and word2vec on model classification accuracy. Positive, negative, and neutral tweets were categorized using the specified classifiers. In addition to accuracy, validation accuracy and the F1 score were utilized as performance indicators.

This shows that TF-IDF feature extraction is better for tweet categorization based on the findings. An F1 score of 0.6133 and a validation accuracy of 0.9616 are achieved by the presented voting classifier using TF-IDF. When compared to non-ensemble classifiers, ensemble classifiers have better accuracy. TF-IDF feature extraction was also used in the implementation of other machine learning models. That doesn't fare well on the chosen dataset, as seen by these findings.

In addition, we found that supervised models outperformed semi-supervised models, with the best model performance predicted by the LR model at 0.80.

Natural Language Processing (NLP) and the Scikit machine learning algorithm were used to create AI models from Twitter data classification. This allowed us to determine that, between January 2021 and April 2022, mental health-related issues increased, but that, over time, people learned to deal with the situation better. COVID-19's influence, as an example, used to fluctuate over time but remained within a narrow range of variability. Vaccination changed the tone of the debate between January 2022 and April 2022. To incentivize vaccination, a slew of incentives was put in place.

As a result of the vaccines' success, those with job losses could return to work and generate new jobs. People are returning to work and returning to school. Using this research, doctors will be able to pinpoint the exact obstacles causing tweets classification in their accounts and aid in finding answers swiftly.

Suggestion for future work, detecting text polarities can be classified into other levels (strong, moderate, weak). We recommend the same system not only for texts but for speech recognition and cleaning noisy data in practical AI and Robotics. The proposed method can be used in AI industries and applied linguistics.

The study is still in process as we keep researching the influence of this virus on tweets classification utilizing social media data to bring about the final answer to positive and negative difficulties.

REFERENCES

1. Yousif, A., Niu, Z., Tarus, J. K., & Ahmad, A. (2019). A survey on sentiment analysis of scientific citations. ***Artificial Intelligence Review***, 52(3), 1805-1838.

2. Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. ***Review in Science (AAAS)***, 349(6245), 261-266.

3. Chen, A., Lu, Y., & Wang, B. (2017). Customers' purchase decision-making process in social commerce: A social learning perspective. ***International Journal of Information Management***, 37(6), 627-638.

4. De Haan, E., Kannan, P. K., Verhoef, P. C., & Wiesel, T. (2018). Device switching in online purchasing: Examining the strategic contingencies. ***Journal of Marketing***, 82(5), 1-19.

5. Rezaeinia, S. M., Rahmani, R., Ghodsi, A., & Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. ***Expert Systems with Applications***, 117, 139-147.

6. Sevinç, H. K., Karas, I. R., & Demiral, E. (2020). MOBILE-WEB-BASE VOLUNTEERED GEOGRAPHIC INFORMATION APPLICATION AND GEOMETRIC ACCURACY ANALYSIS FOR TRAFFIC ACCIDENTS. ***The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences***, 44, 375-378.

7. Nakayama, M., & Wan, Y. (2019). The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews. Information & Management, ***Association for Computing Machinery (ACM)***, 56(2), 271-279.

8. Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. ***International journal of information management***, 39, 156-168.

9. Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. ***Journal of Big Data***, 6(1), 1-16.

10. Poornima, A., & Priya, K. S. (2020, March). A comparative sentiment analysis of sentence embedding using machine learning techniques. In 2020 ***6th International Conference on Advanced Computing and Communication Systems (ICACCS)*** (pp. 493-496). IEEE.

11. Abd El-Jawad, M. H., Hodhod, R., & Omar, Y. M. (2018, December). Sentiment analysis of social media networks using machine learning. In 2018 ***14th international computer engineering conference (ICENCO)*** (pp. 174-176). IEEE.

12. Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques. In ***Cognitive Informatics and Soft Computing*** (pp. 639-647). Springer, Singapore.

13. Yadav, N., Kudale, O., Rao, A., Gupta, S., & Shitole, A. (2021). Twitter Sentiment Analysis Using Supervised Machine Learning. In ***Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI*** 2020 (pp. 631-642). Springer Singapore.

14. Boiy, E., & Moens, M. F. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. ***Information Retrieval Journal***, 12(5), 526-558.

15. Yi, S., & Liu, X. (2020). Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. ***Complex & Intelligent Systems***, 6(3), 621-634.

16. Kumar, P. K., & Nandagopalan, S. (2017). Insights to problems, research trend and progress in techniques of sentiment analysis. ***International Journal of Electrical and Computer Engineering***, 7(5), 2818.

17. Jain, A. P., & Dandannavar, P. (2016, July). Application of machine learning techniques to sentiment analysis. In 2016 ***2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*** (pp. 628-632). IEEE.

18. Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. ***Mathematical and Computational Applications***, 23(1), 11.

19. Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In ***A practical guide to sentiment analysis*** (pp. 1-10). Springer, Cham.

20. Wang, Y., Sun, A., Han, J., Liu, Y., & Zhu, X. (2018, April). Sentiment analysis by capsules. In ***Proceedings of the 2018 world wide web conference*** (pp. 1165-1174).

21. Liu, N., Shen, B., Zhang, Z., Zhang, Z., & Mi, K. (2019). Attention-based Sentiment Reasoner for aspect-based sentiment analysis. ***Human-centric Computing and Information Sciences***, 9(1), 1-17.

22. Trupthi, M., Pabboju, S., & Narasimha, G. (2017, January). Sentiment analysis on twitter using streaming API. In 2017 ***IEEE 7th International Advance Computing Conference (IACC)*** (pp. 915-919). IEEE.

23. Halibas, A. S., Shaffi, A. S., & Mohamed, M. A. K. V. (2018, March). Application of text classification and clustering of Twitter data for business analytics. In 2018 ***Majan international conference (MIC)*** (pp. 1-7). IEEE.

24. Soumya, S., & Pramod, K. V. (2019, November). Sentiment analysis of Malayalam tweets using different deep neural network models-case study. In 2019 ***9th International Conference on Advances in Computing and Communication (ICACC)*** (pp. 163-168). IEEE.

25. Kumar, S. S., Kumar, M. A., & Soman, K. P. (2017, December). Sentiment analysis of tweets in malayalam using long short-term memory units and convolutional neural nets. In ***International Conference on Mining Intelligence and Knowledge Exploration*** (pp. 320-334). Springer, Cham.

26. Rahul, M., Rajeev, R. R., & Shine, S. (2018). Social Media Sentiment Analysis for Malayalam., ***International Journal of Computer Sciences and Engineering***, Vol.06, Issue.06, pp.48-53, 2018.

27. Swathi, R., & Seshadri, R. (2017, June). Systematic survey on evolution of machine learning for big data. In 2017 **International Conference on Intelligent Computing and Control Systems (ICICCS)** (pp. 204-209). IEEE.

28. Kawade, D. R., & Oza, K. S. (2017). Sentiment analysis: machine learning approach. ***International journal of engineering and technology***, 9(3), 2183-2186.

29. Shamantha, R. B., Shetty, S. M., & Rai, P. (2019, February). Sentiment analysis using machine learning classifiers: evaluation of performance. In 2019 ***IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*** (pp. 21-25). IEEE.

30. Yadav, N., Kudale, O., Rao, A., Gupta, S., & Shitole, A. (2021). Twitter Sentiment Analysis Using Supervised Machine Learning***. In Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI*** 2020 (pp. 631-642). Springer Singapore.

31. Yogi, T. N., & Paudel, N. Comparative Analysis of Machine Learning Based Classification Algorithms for Sentiment Analysis. ***International Journal of Innovative Science, Engineering & Technology***, 7(6), 1-9.

32. Suryawanshi, R., Rajput, A., Kokale, P., & Karve, S. S. (2020). Sentiment Analyzer using Machine Learning. ***International Research Journal of Modernization in Engineering Technology and Science***, 2(6), 1-12.

33. Attri, V., Batra, I., & Malik, A. (2021, April). A Relative Study on Analytical Models. In 2021 ***2nd International Conference on Intelligent Engineering and Management (ICIEM)*** (pp. 52-57). IEEE.

34. Attri, V., Batra, I., & Malik, A. (2021, April). A Relative Study on Analytical Models. In 2021 ***2nd International Conference on Intelligent Engineering and Management (ICIEM)*** (pp. 52-57). IEEE.

35. Shirsat, V., Jagdale, R., Shende, K., Deshmukh, S. N., & Kawale, S. (2019). Sentence Level Sentiment Analysis from News Articles and Blogs using Machine Learning Techniques. ***International Journal of Computer Sciences and Engineering***, 1(1), 12-17.

36. Boiy, E., & Moens, M. F. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. ***Information retrieval***, 12(5), 526-558.

37. Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. ***Expert Systems with Applications***, 57, 117-126.

38. Yi, S., & Liu, X. (2020). Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. ***Complex & Intelligent Systems,*** 6(3), 621-634.

39. Mujahid, F., Anwar, S., Afzal, A., Riaz, L., & Saad, M. ENHANCED OBJECTIVE SENTIMENTAL ANALYSIS USING NLP TECHNIQUES. ***Journal of Natural and Applied Sciences Pakistan***, Vol 2 (1), 2020 pp 217-231

40. Machine Learning & its Applications Outsource to India. (2020). Retrieved on Oct18,2021,from,***https://www.outsource2india.com/software/articles/machine-learning-applications-how-it-works-who-uses-it.asp.***

41. Jain, A. P., & Dandannavar, P. (2016, July). Application of machine learning techniques to sentiment analysis. In 2016 ***2nd International Conference on Applied and Theoretical Computing and Communication Technology*** (iCATccT) (pp. 628-632). IEEE.

42. Deng, N., & Li, X. R. (2018). Feeling a destination through the "right" photos: A machine learning model for DMOs' photo selection. ***Tourism Management***, 65, 267-278.

43. El Rahman, S. A., AlOtaibi, F. A., & AlShehri, W. A. (2019, April). Sentiment analysis of twitter data. In 2019 ***International Conference on Computer and Information Sciences (ICCIS)*** (pp. 1-4). IEEE.

44. Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. ***Mathematical and Computational Applications***, 23(1), 11.

45. Wagh, R., & Punde, P. (2018, March). Survey on sentiment analysis using twitter dataset. In 2018 ***Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*** (pp. 208-211). IEEE.

46. Kumar, A., & Jaiswal, A. (2020). Systematic literature review of sentiment analysis on Twitter using soft computing techniques. ***Concurrency and Computation: Practice and Experience***, 32(1), e5107.

47. Berry, M. W., Mohamed, A., & Yap, B. W. (Eds.). (2019). Supervised and unsupervised learning for data science. ***Springer Nature***.

48. Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. ***Ai & Society,*** 30(1), 89-116.

49. Shoumy, N. J., Ang, L. M., Seng, K. P., Rahaman, D. M., & Zia, T. (2020). Multimodal big data effective analytics: A comprehensive survey using text, audio, visual and physiological signals. ***Journal of Network and Computer Applications***, 149, 102447.

50. Schnebly, J., & Sengupta, S. (2019, January). Random forest twitter bot classifier. In 2019 ***IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*** (pp. 0506-0512). IEEE.

51. Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. ***Journal of Applied Science and Technology Trends***, 2(01), 20-28.

52. Olaru, C., & Wehenkel, L. (2003). A complete fuzzy decision tree technique. ***Fuzzy sets and systems***, 138(2), 221-254.

53. Ding, B., Zheng, Y., & Zang, S. (2009, July). A new decision tree algorithm based on rough set theory. In 2009 ***Asia-Pacific Conference on Information Processing*** (Vol. 2, pp. 326-329). IEEE.

54. Zhu, L., & Yang, Y. (2016, November). Improvement of decision tree ID3 algorithm. In ***International Conference on Collaborative Computing: Networking, Applications and Work sharing*** (pp. 595-600). Springer, Cham.

55. Kaewrod, N., & Jearanaitanakij, K. (2018, November). Improving ID3 algorithm by ignoring minor instances. In 2018 ***22nd International Computer Science and Engineering Conference (ICSEC)*** (pp. 1-5). IEEE.

56. Rajeshkanna, A., & Arunesh, K. (2020, July). ID3 decision tree classification: An algorithmic perspective based on error rate. In 2020 ***International Conference on Electronics and Sustainable Communication Systems*** (ICESC) (pp. 787-790). IEEE.

57. Devi, B. L., Bai, V. V., Ramasubbareddy, S., & Govinda, K. (2020). Sentiment analysis on movie reviews. In ***Emerging Research in Data Engineering Systems and Computer Communications*** (pp. 321-328). Springer, Singapore.

58. Guerreiro, J., & Rita, P. (2020). How to predict explicit recommendations in online reviews using text mining and sentiment analysis. ***Journal of Hospitality and Tourism Management***, 43, 269-272.

59. Mehta, R. P., Sanghvi, M. A., Shah, D. K., & Singh, A. (2020). Sentiment analysis of tweets using supervised learning algorithms. In ***First International Conference on Sustainable Technologies for Computational Intelligence*** (pp. 323-338). Springer, Singapore.

60. Zhang, J. (2020). Sentiment analysis of movie reviews in Chinese. ***Uppsala University, Diva portal.***

61. López-Chau, A., Valle-Cruz, D., & Sandoval-Almazán, R. (2020). Sentiment analysis of Twitter data through machine learning techniques. In ***Software Engineering in the Era of Cloud Computing*** (pp. 185-209). Springer, Cham.

62. Addi, H. A., Ezzahir, R., & Mahmoudi, A. (2020, March). Three-level binary tree structure for sentiment classification in arabic text. In ***Proceedings of the 3rd International Conference on Networking, Information Systems & Security*** (pp. 1-8).

63. Patel, R., & Passi, K. (2020). Sentiment analysis on Twitter data of world cup soccer tournament using machine learning. ***IoT***, 1(2), 218-239.

64. Wang, Y., Chen, Q., Shen, J., Hou, B., Ahmed, M., & Li, Z. (2021). Aspect-level sentiment analysis based on gradual machine learning. ***Knowledge-Based Systems***, 212, 106509.

65. Baccouche, A., Garcia-Zapirain, B., & Elmaghraby, A. (2018, December). Annotation technique for health-related tweets sentiment analysis. In 2018 ***IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*** (pp. 382-387). IEEE.

66. Hameed, Z., & Garcia-Zapirain, B. (2020). Sentiment classification using a single-layered BiLSTM model. ***IEEE Access***, 8, 73992-74001.

67. Zhang, M. (2020, April). E-commerce comment sentiment classification based on deep learning. In 2020 ***IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*** (pp. 184-187). IEEE.

68. Solis, B., & Breakenridge, D. K. (2009). Putting the public back in public relations: *How social media is reinventing the aging business of PR*. ***Ft Press.***

69. Misopoulos, F., Mitic, M., Kapoulas, A., & Karapiperis, C. (2014). Uncovering customer service experiences with Twitter: the case of airline industry. ***Management decision: MD***. Vol. 52.2014, 4, p. 705-723

70. TUTOR, I., & TUTOR, C. Investigating Perceptual and Biological Feedbacks in ***Human Robot Interaction***. ***Anno Conseguimento Titolo***, 2018

71. Kirasich, K., Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogeneous datasets. ***SMU Data Science Review***, 1(3), 9.

72. Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. ***Journal of the Royal Statistical Society: Series B (Statistical Methodology)***, 70(1), 53-71.

73. Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. ***Journal of the Royal Statistical Society: Series A (General)***, 135(3), 370-384.

74. Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. ***CRC press***.

75. Cameron, A. C., & Windmeijer, F. A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. ***Journal of econometrics,*** 77(2), 329-342.

76. Ayer, T., Chhatwal, J., Alagoz, O., Kahn Jr, C. E., Woods, R. W., & Burnside, E. S. (2010). Comparison of logistic regression and artificial neural network models in breast cancer risk estimation. ***Radio Graphics***, 30(1), 13-22.

77. Cummins, N., Amiriparian, S., Ottl, S., Gerczuk, M., Schmitt, M., & Schuller, B. (2018, April). Multimodal bag-of-words for cross domains sentiment analysis. In 2018 ***IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*** (pp. 4954-4958). IEEE.

78. Kadhim, A. I. (2019, April). Term weighting for feature extraction on Twitter: A comparison between BM25 and TF-IDF. In 2019 ***international conference on advanced science and engineering (ICOASE)*** (pp. 124-128). IEEE.

79. Soares, E. R., & Barrére, E. (2019, October). An optimization model for temporal video lecture segmentation using word2vec and acoustic features. In ***Proceedings of the 25th Brazillian Symposium on Multimedia and the Web*** (pp. 513-520).

80. Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. ***PLOS ONE***, 16(2), e0245909.

81. Phan, H. T., Tran, V. C., Nguyen, N. T., & Hwang, D. (2020, September). A framework for detecting user's psychological tendencies on twitter based on tweets sentiment analysis. In ***International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*** (pp. 357-372). Springer, Cham.

82. Baccouche, A., Garcia-Zapirain, B., & Elmaghraby, A. (2018, December). Annotation technique for health-related tweets sentiment analysis. In 2018 ***IEEE international symposium on signal processing and information technology (ISSPIT)*** (pp. 382-387). IEEE.

RESUME

Firas Fadhil SHIHAB graduated first and elementary education in Baghdad-Iraq. He completed high school education at (Palestine High School) in Baghdad City, after that in 2010 he started a bachelor's program in the Department of Computer Science at Al-Mustansiriya University and completed it in the year 2014, He has worked in the computer field and in the field of airline ticket reservations and is still in this work. He moved to Turkey and began studying for his master's education at the department of Computer Engineering at Karabuk University in the year 2020, His aim is to complete his doctoral studies, Insha'Allah.