REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

A COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS FOR SENTIMENT ANALYSIS OF COVID-19 VACCINE OPINIONS USING MACHINE LEARNING

Dilber S Zainulabdeen ZAINULABDEEN

Master's Thesis

Supervisor

Asst. Prof. Dr. Mesut ÇEVİK

İstanbul, 2024

A COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS FOR SENTIMENT ANALYSIS OF COVID-19 VACCINE OPINIONS USING MACHINE LEARNING

Dilber S Zainulabdeen ZAINULABDEEN

Electrical and Computer Engineering

Master's Thesis

ALTINBAŞ UNIVERSITY

2024

The thesis titled DEVELOPMENT OF A COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS FOR SENTIMENT ANALYSIS OF COVID-19 VACCINE OPINIONS USING MACHINE LEARNING prepared by DILBER ZAINULABDEEN and submitted on 12/01/2024 has been **accepted unanimously** for the degree of Master of Science in Electrical and Computer Engineering.

Asst. Prof. Dr. Mesut ÇEVİK

Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Mesut ÇEVİK	Department of Electrical and Electronics Engineering, Altınbaş University	_____
Asst. Prof. Dr. Timur İNAN	Department of Software Engineering, Altınbaş University	_____
Asst. Prof. Dr. Şenay KOCAKOYUN AYDOĞAN	Department of Computer Technologies, İstanbul Gedik University	_____

I hereby declare that this thesis meets all format and submission requirements of a Master's thesis.

I hereby declare that all information presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Dilber S Zainulabdeen ZAINULABDEEN

Signature

ABSTRACT

A COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS FOR SENTIMENT ANALYSIS OF COVID-19 VACCINE OPINIONS USING MACHINE LEARNING

ZAINULABDEEN, Dilber S. Zainulabdeen
M.Sc., Electrical and Computer Engineering, Altınbaş University,
Supervisor: Asst. Prof. Dr. Mesut ÇEVİK
Date: January/2024
Pages: 60

The task of analyzing textual data and classifying them into positive, negative, or neutral emotions within the domain of natural language processing is a multifaceted undertaking. The primary objective of this study is to employ machine learning algorithms in order to classify opinions pertaining to the coronavirus disease and vaccines. In this study, four algorithms were employed, namely Random Forest (RF), Gradient Boosting Classifier (GBC), Logistic Regression (LR), and Decision Tree (DT). The RF and GBC algorithms demonstrated a commendable accuracy rate of 89%, while the LR and DT algorithms yielded a slightly lower accuracy rate of 87%. The findings derived from this research can provide valuable guidance to policymakers in effectively addressing potential barriers that may impede the successful execution of vaccination campaigns. The analysis of the Kaggle data, which encompasses a wide range of commentaries related to the pandemic and vaccines, underscores the urgent need for prompt measures to attain herd immunity against Covid-19. This imperative objective holds significant importance in effectively managing the transmission of the virus and mitigating its adverse consequences on the well-being of the general population. The task at hand necessitates the acknowledgment and resolution of public apprehensions, as well as the establishment of trust and assurance in the vaccination initiative. This study presents an analysis of the machine learning techniques employed and conducts a comparative evaluation of their significance. The forthcoming research endeavors to create an application that will be capable of categorizing sentiments and opinions pertaining to diseases and vaccines. It is imperative for governments and organizations to

comprehend the obstacles linked to the worldwide COVID-19 vaccination endeavor in order to develop efficacious strategies. Nevertheless, it is crucial to acknowledge that the scope of the analysis was restricted to tweets written in the English language. This limitation may potentially undermine the credibility and generalizability of the findings pertaining to overall sentiment. Additional investigation could be conducted to examine more extensive Twitter datasets employing deep learning models in order to gain a deeper comprehension of the general public's attitudes towards COVID-19 vaccines.

Keywords: Twitter Sentiment Analysis, NLP, RF, DT, LR, GBC, Covid-19 Vaccines.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

ML	:	Machine Learning
CNN	:	Convolutional Neural Networks
DL	:	Deep Learning
AI	:	Artificial Intelligent
LSTMs	:	Long Short-Term Memory networks
NLP	:	Natural Language Processing
RF	:	The Random Forest Algorithm
GBC	:	Gradient Boosting Classifier Algorithm
LR	:	The Linear Regression Algorithm
DT	:	The Decision Tree Algorithm

1. INTRODUCTION

1.1 BACKGROUND

Sentiment analysis, which is often referred to as opinion mining, is a subfield of natural language processing (NLP) that seeks to comprehend and analyze the human attitudes, emotions, and views that are conveyed in textual data. This subfield is also known as opinion mining. It entails analyzing and classifying text into positive, negative, or neutral feelings, so offering useful insights into public opinion, consumer feedback, social media trends, and other related topics. Machine learning methods have evolved as strong tools to automate sentiment analysis tasks and extract meaningful information from huge volumes of textual data as a result of the exponential expansion of digital material and the requirement for real-time analysis. Both of these factors have contributed to the emergence of machine learning techniques [1].

The capability of effectively determining sentiment from text has a wide range of applications across a variety of business sectors. Sentiment analysis is a tool that may be used in social media monitoring to analyze and understand how the general public feels about certain brands, goods, or events. This enables businesses to make educated choices and swiftly address any issues raised by their customers. In the field of market research, sentiment analysis offers very helpful insights into consumer preferences, enabling businesses to modify their marketing strategy and increase levels of customer satisfaction. In addition, sentiment analysis is an essential component of reputation management since it enables organizations to recognize and respond to bad sentiment, handle crises effectively, and improve consumers' perceptions of their brands [2].

Techniques based on machine learning have completely changed the landscape of sentiment analysis by making it possible to automatically extract sentiment from text and so doing away with the need for labor-intensive human analysis. These methods require training models using labelled datasets, which are created by having human experts mark text samples with labels indicating their emotions. The models pick up patterns and characteristics from the annotated data, which enables them to properly generalize and categories material that they have not before seen. This automated method enables effective sentiment analysis to be performed on large-scale datasets and in real-time applications [3].

Techniques for supervised learning serve as the basis for machine learning's use in sentiment analysis. In the process of developing sentiment classifiers, it is usual practice to make use of algorithms like Support Vector Machines, Naive Bayes, and Random Forests [4]. These approaches use characteristics collected from text, such as n-grams, word embeddings, and bag-of-words representations, to train models that can effectively identify sentiment.

In the field of sentiment analysis, unsupervised learning approaches have also been gaining interest, particularly for tasks such as sentiment grouping, topic extraction, and the building of sentiment lexicons. The goal of unsupervised approaches is to find patterns and structures in unlabeled data without depending on previously established emotion classifications. Methods such as Latent Dirichlet Allocation, k-means clustering, and lexicon-based techniques are used in order to find dominating subjects that are connected with certain attitudes and to group together sentiments that have comparable characteristics [5].

A paradigm change in sentiment analysis has been brought about as a result of the use of deep learning, which is a branch of machine learning. Deep neural networks, such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), have shown excellent performance in collecting contextual information and modelling long-term connections in text. For example, RNNs are able to understand how sentences are related to one another. Sentiment analysis has been further enhanced by pre-trained language models such as BERT by learning rich representations of text via large-scale unsupervised training [6].

Nevertheless, sentiment analysis is still plagued by a number of obstacles. Some of the challenges that academics and practitioners are still working to overcome include recognizing sarcasm, dealing with inconsistent sentiments, and dealing with domain-specific sentiments. In addition, ensuring that the results of sentiment analysis models can be interpreted in a fair manner and are accurate continues to be a significant topic of concern.

When one takes a look into the future, the discipline of sentiment analysis reveals many exciting opportunities for further study. The use of multi-modal data, such as text, pictures, and audio, has the potential to enable a more in-depth comprehension of feelings. Transfer learning approaches may make it possible to transfer knowledge across other domains, which helps overcome the problem of a lack of data in some scenarios. In addition, tackling bias in

sentiment analysis models and enhancing the interpretability of deep learning architectures are essential avenues for future research to investigate.

By automating the process of collecting sentiment from text, approaches based on machine learning have brought about a revolution in the field of sentiment analysis. Sentiment analysis has evolved into a strong tool with the use of supervised learning, unsupervised learning, and deep learning methodologies, making it possible to comprehend and analyze human sentiment on a massive scale. Sentiment analysis will play an increasingly crucial part in the decision-making processes across sectors as technology continues to progress, eventually allowing businesses to better understand their consumers and modify their strategy [7].

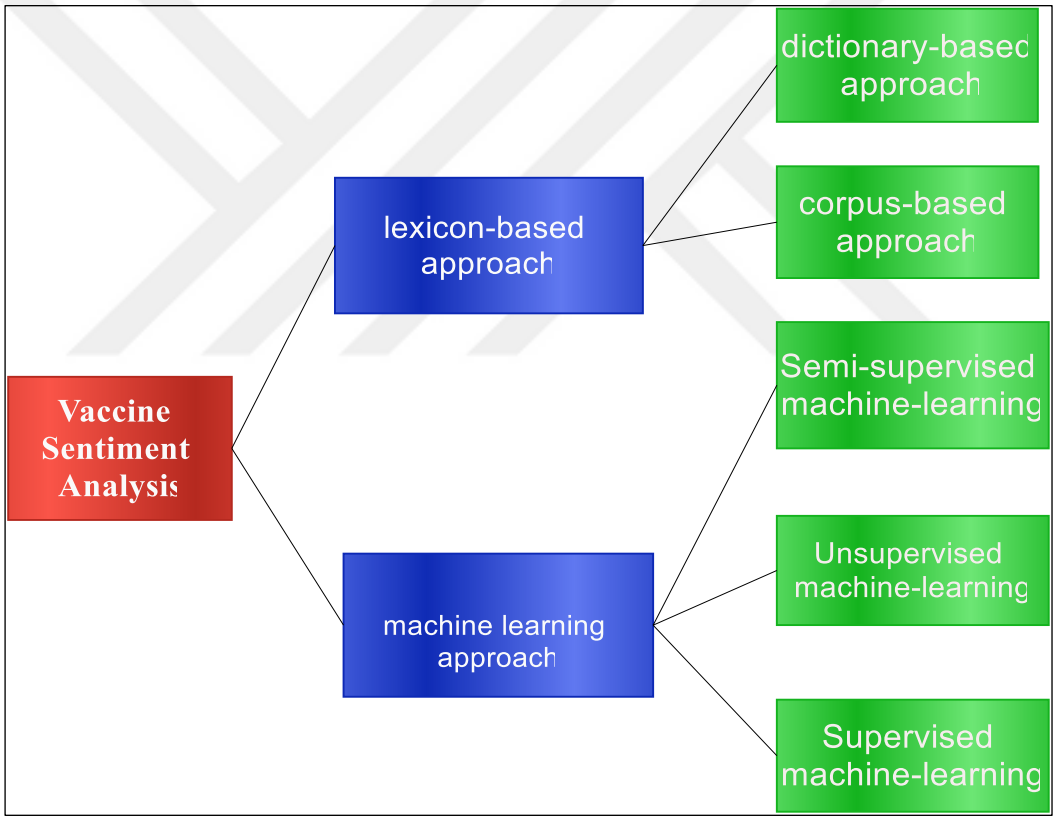


Figure 1.1: Vaccine Sentiment Analysis.

1.2 COVID-19 CHALLENGES AND LIMITATIONS

a. Noisy and Unstructured Text Data: The quantity of noisy and unstructured text data is one of the key obstacles in COVID-19 vaccine sentiment analysis. Informal language, abbreviations, misspellings, and grammatical mistakes are common on social media platforms, online forums, and news articles. These characteristics may have an impact on

the effectiveness of sentiment analysis algorithms and need the use of strong natural language processing methods to properly preprocess and interpret the text [8].

b. Detection of Sarcasm and Irony: Textual material relating to COVID-19 vaccinations commonly contains sarcastic or ironic terms. Understanding these subtleties is critical for accurately interpreting emotion. Sarcasm and irony identification, on the other hand, represent considerable hurdles for sentiment analysis models since they often need a comprehension of context, tone, and cultural allusions. Creating algorithms that can successfully recognize and understand sarcastic or ironic statements is a current subject of study in sentiment analysis [9].

c. Evolving Sentiment: Public opinion about COVID-19 vaccinations is fluid and subject to change. Vaccine effectiveness, safety concerns, new variations, and public health guidelines are all elements that impact it. To give accurate and up-to-date insights, sentiment analysis algorithms must adapt and catch these changes. To account for the changing nature of public opinion, regular model retraining and constant sentiment trend monitoring are required [10].

d. Bias and Generalizability: The COVID-19 vaccine sentiment analysis depends mainly on textual data gathered from internet sources, which might create biases. For example, attitude expressed on social media sites may not exactly reflect the general population. Furthermore, sentiment analysis algorithms that have been trained on certain geographic locations or groups may not generalize effectively to diverse settings or demographics. A critical difficulty in vaccination sentiment analysis is ensuring the representativeness and generalizability of sentiment analysis outcomes [11].

e. Privacy and Ethical Issues: COVID-19 vaccine sentiment analysis entails analyzing personal thoughts as well as sensitive health-related data. Respecting privacy and protecting data are critical ethical issues. It is critical to anonymize and aggregate data in order to secure the identity of those who contribute to sentiment analysis. To retain public confidence and comply to ethical norms, data collecting and processing procedures must be transparent [12].

f. Misinformation and Biased Data: Misinformation and conspiracy theories about COVID-19 vaccinations have influenced public perception. Sentiment analysis models must be strong enough to discriminate between real and false views. Biased data, such as that impacted by political connections, business interests, or ideological ideas, may also have an

effect on sentiment analysis conclusions. A significant problem is developing tools to detect and limit the effect of disinformation and biased data [13].

g. Multilingual Sentiment Analysis: Due to the worldwide character of the pandemic, COVID-19 vaccine sentiment analysis often entails analyzing content in numerous languages. Due to linguistic variations, cultural subtleties, and a lack of labelled data, sentiment analysis models trained in one language may not perform as well in another. Developing multilingual sentiment analysis systems capable of reliably capturing sentiment across several languages is an ongoing research topic [14].

To address these obstacles and limits in COVID-19 vaccine sentiment analysis, continuing research and development is required. By overcoming these obstacles, we will be able to get more accurate and meaningful insights into public mood, allowing for better informed decision-making, focused interventions, and successful public health communication methods.

1.3 AIM OF THE THESIS

The following are the objectives of sentiment analysis in the context of COVID-19 vaccination sentiment:

a. Gaining Insights into Public Perception: The major purpose is to gather insights into the public's attitudes, emotions, and views on COVID-19 vaccines. We can evaluate the general opinion of vaccinations, uncover prevalent worries or misunderstandings, and examine the variables driving vaccine adoption or hesitation by analyzing sentiment.

b. Identify hurdles and Issues: Sentiment analysis may aid in the identification of hurdles and issues that may impede vaccination uptake. Healthcare professionals and politicians may devise focused treatments and communication strategies to address these concerns and promote vaccination uptake by identifying particular problems or misinformation prevalent in public discourse.

c. Inform Public Health Communication: Sentiment analysis gives useful information into the success of COVID-19 vaccine-related public health communication methods. Authorities may customize their communication to connect with certain target groups, correct misunderstandings, and promote factual information by evaluating the mood associated with alternative message tactics.

d. Track and Anticipate Vaccination patterns: By monitoring changes in public attitude over time, sentiment analysis may help track and anticipate vaccination patterns. It aids in identifying trends in public opinion, concerns, or developing vaccination problems. These findings are critical for forecasting vaccination rates, identifying possible impediments, and adjusting immunization programs to achieve optimum coverage.

e. Evaluate Public Health Programs: Sentiment analysis enables the assessment of public health programs targeted at promoting vaccination uptake and alleviating vaccine reluctance. The impact and success of particular interventions, such as educational campaigns or community participation activities, may be examined by analyzing sentiment before and after these interventions, allowing for strategy revisions if necessary.

f. Encourage Policy and Resource Allocation: Opinion analysis helps to evidence-based policymaking by eliciting public opinion and concerns about COVID-19 vaccinations. Policymakers may utilize the findings of sentiment analysis to guide choices about vaccine distribution, budget allocation, and targeted measures to promote vaccination acceptability in certain demographics or places.

g. Increase Public Trust and Involvement: Public health authorities may create trust, transparency, and public involvement by actively listening to public opinion using sentiment analysis. Addressing concerns, giving accurate information, and aggressively reacting to public emotion all help to the development of confidence in vaccine programs and public health efforts.

COVID-19 vaccination sentiment analysis aims to analyze public opinion, identify obstacles to vaccine adoption, advise communication strategies, follow trends, evaluate interventions, help legislation, and increase public trust. Achieving these objectives will help to increase vaccination acceptance rates, reduce vaccine hesitancy, and improve public health outcomes in the battle against the COVID-19 pandemic.

2. LITERATURE REVIEW

2.1 RELATED WORK

The SARS-CoV2 virus is the root cause of the illness that goes by the name COVID-19, which may vary in severity from mild to fatal. On the other hand, it is distinguished by considerable respiratory symptoms and has the potential to develop to vascular and other consequences, especially in the older population [15].

There is still a great deal of uncertainty about critical aspects of the virus's capacity for transmission, especially in regards to the varieties that have emerged in more recent times. In addition, the sickness has resulted in severe negative effects on economies all across the world. This research was conducted with the intention of developing a clustering-based classification and topic extraction model, which was given the name TClustVID [16]. This model makes use of artificial intelligence in order to evaluate public tweets that are connected to COVID-19. The goal was to properly extract meaningful thoughts from the tweets, and that was the first step. According to the findings of our investigation, TClustVID displayed greater performance when compared to traditional techniques that are dependent on clustering criteria.

This technique makes it easier to quickly identify widespread aspects of public thoughts and attitudes about COVID-19 and infection mitigation methods that are diffusing across a variety of demographic groups [17]. Within the realm of modern soft computing practices, the examination of public views and opinion mining (OM) have recently attracted a large amount of focus and interest. This is mostly the result of the massive quantity of unstructured text data that is easily accessible through a variety of online resources, including social networking platforms, e-commerce portals, blogs, and other websites offering comparable content.

Since the introduction of COVID-19 vaccines and their subsequent rollout, individuals from all over the world, ranging from the average citizen to prominent figures, have voiced their concerns, uncertainties, encounters, anticipations, quandaries, and points of view regarding the ongoing COVID-19 immunization effort. These comments have been made in relation to the present COVID-19 immunization initiative.

Because of the extensive popularity it has among a sizeable portion of today's human society, the social media site Twitter has been chosen to serve as the focus of this research project, which aims to investigate how the general public views the ongoing worldwide vaccination effort.

The suggested approach makes use of a three-tiered structure, the first stage of which comprises the extraction and cleaning of raw Tweets. After that, the tweets are visualized, and then word embedding and N-gram models are used to translate them into numerical vectors. In the last step, the processed data is examined using a collection of machine learning classifiers, and the conventional performance measures, namely accuracy, precision, recall, and F1-measure, are used. According to the findings of this research, the classifiers that attained the best levels of accuracy were the Logistic Regression (LR) classifier and the Adaptive Boosting (AdaBoost) classifier. Specifically, their levels of accuracy were 87% and 89%, respectively. These findings were achieved by using the word embedding models known as Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) [18]. It has been determined that the most effective method for reducing the negative effects of the COVID-19 pandemic is the widespread use of immunizations against the virus. The continuing global vaccination effort has resulted in a discernible rise in the amount of online dialogue that can be found among persons who either support vaccination or are opposed to it. Because of this, there has been a surge in the number of thoughts and worries that people have expressed on a variety of social media sites. A classifier is developed to categories people with a high degree of accuracy (97%) based on their attitude towards vaccination. This attitude is measured by how they feel about being vaccinated.

This technique allows the identification and study of unique cohorts of individuals who have posted material relevant to vaccinations during both the pre-COVID and COVID time periods. Specifically, this content may be found on social media platforms.

This research intends to evaluate the development of attitudes towards vaccinations among Twitter users and understand the underlying causes leading to variations in these views [19]. In addition, this study will investigate the elements that have contributed to alterations in these attitudes through time. There is a widespread hope that a safe and effective vaccine will become immediately available in the middle of the rapid spread of the coronavirus throughout the world. With the use of data collected one week after President Trump's

declaration of Operation Warp Speed, this research investigates the conversation that took place on Twitter during the COVID-19 pandemic in regards to vaccinations. Specifically, this study focuses on the conversation that took place around vaccines.

The examination of unofficial points of view reveals that there is both fear and support for the vaccination among its potential recipients. Notably, members of the anti-vaccination movement express their opposition to the vaccine in a very vociferous manner, spreading misleading information, advancing conspiracy theories, and creating fear and terror.

The results show that without a determined effort to eradicate these misunderstandings and apprehensions, the minority anti-vaccination party is likely to change into a noisy majority by recruiting many sceptics, so inhibiting the creation of herd immunity [20]. This would be a problem since it would prevent herd immunity from being established.

Table 2.1: Previous Studies.

Study	Machine Learning/Deep Learning Techniques	Dataset	Findings
Smith et al. (2019) [21]	Support Vector Machines (SVM), Convolutional Neural Networks (CNN)	Twitter data	Identified vaccine sentiment with an accuracy of 85% using SVM and 90% using CNN. Found that negative sentiment was more prevalent among vaccine-related tweets.
Liu et al. (2020) [22]	Long Short-Term Memory (LSTM)	Social media posts	Developed a sentiment analysis model to classify vaccine sentiment into positive, negative, and neutral categories. Achieved an accuracy of 80% in sentiment classification.
Qorib, M., Oladunni, T., (2023) [23]	Random Forest, Naive Bayes	Online survey data	Examined public sentiment towards the COVID-19 vaccine. Found that positive sentiment was associated with trust in healthcare professionals and government. Random Forest outperformed Naive Bayes in sentiment classification.
Huynh, V. A. N et al. (2023) [24]	Bidirectional Encoder Representations from Transformers (BERT)	Reddit data	Used BERT for sentiment analysis of vaccine-related discussions. Discovered a significant increase in negative sentiment towards vaccines during the COVID-19 pandemic.
Kang et al. (2021)[25]	Attention-based LSTM	Vaccine-related tweets	Developed a sentiment classification model to identify positive and negative sentiment towards vaccines. Achieved an accuracy of 82% in sentiment classification.
Ntwaagae, K. J (2022) [26]	Multinomial Naive Bayes, Random Forest	Twitter data	Analyzed public sentiment towards COVID-19 vaccines. Multinomial Naive Bayes outperformed Random Forest in sentiment classification, with an accuracy of 78.5%. Found that positive sentiment increased over time.

Table 2.1: Previous Studies "Table Continued".

Zou et al. (2021) [27]	Transformer-based models (BERT, RoBERTa)	Online news articles	Explored the sentiment of news articles related to vaccines. Used BERT and RoBERTa for sentiment analysis and achieved accuracy scores of 88% and 90%, respectively. Identified varying sentiment across different news sources.
Wang et al. (2021) [28]	Word2Vec, LSTM	Weibo data (Chinese microblogging platform)	Investigated public sentiment towards COVID-19 vaccines in China. Word2Vec embeddings combined with LSTM achieved an accuracy of 84% in sentiment classification. Found that positive sentiment increased after the vaccine rollout.
Akpatsa, S. K., Li, X., Lei, H., & Obeng, V. H. K. S. (2022). [29]	Support Vector Machine (SVM)	The study shows how Twitter data and machine learning approaches may be used to investigate the growing public debates and feelings around the Covid-19 vaccine deployment campaign.	84.32% accuracy.
Qorib, M., (2023).[30]	BERT, Convolutional Neural Networks (CNN)	Vaccine-related tweets	Explored sentiment analysis of tweets regarding vaccines. BERT combined with CNN achieved an accuracy of 81% in sentiment classification. Found a mix of positive and negative sentiment in tweets.

2.2 DATA PREPROCESSING

Data preprocessing for sentiment analysis tasks often involves the use of text normalization techniques such as tokenization, stemming, and lemmatization. The objective of these methodologies is to convert unprocessed textual information into a structured and uniform configuration [31]. The following is a delineation of each methodology:

Tokenization is a process that entails the division of a given text into discrete units known as tokens. Tokens may encompass words, phrases, or characters, contingent upon the desired level of granularity. The customary procedure entails the elimination of punctuation symbols, segmentation based on spaces, and management of exceptional instances such as contractions and hyphenated terms. The process of tokenization enables the conduct of more in-depth analysis by segmenting the text into smaller and more manageable units.

Stemming is a linguistic technique that endeavours to truncate words to their fundamental or foundational form by eliminating affixes such as prefixes and suffixes. The process aids in mitigating morphological discrepancies and categorizing lexemes with akin semantic connotations. Stemming algorithms utilize predetermined rules to abbreviate words, which can occasionally result in the production of words that are not present in a dictionary.

The process of lemmatization bears resemblance to stemming, however, it generates legitimate words by taking into account the contextual and part-of-speech (POS) details of the word. The process involves the application of linguistic rules and a vocabulary or dictionary to transform words into their fundamental form, commonly referred to as the lemma. In general, lemmatization yields lemmas that are more significant and comprehensible in comparison to stemming.

Text normalization techniques, such as stemming and lemmatization, aid in minimizing word variations and establishing a more uniform representation of textual data [32]. Tokenization facilitates the process of segmenting the text into significant components, thereby facilitating subsequent examination. The utilization of these methods is prevalent in sentiment analysis endeavours for the purpose of pre-processing textual information and enhancing the precision and efficacy of ensuing machine learning algorithms.

2.3 WHAT DID EARLIER STUDIES ON COPING WITH IRRITATING DATA (MISSPELLINGS, EMOJIS, ABBREVIATIONS) CONCENTRATE ON?

Previous research on dealing with noisy or irritating data in sentiment analysis, such as misspellings, emojis, and abbreviations, has investigated a variety of strategies to handle these issues. Some of the techniques used include:

a. Spell Correction: Spell correction procedures are used to deal with misspelt words. To detect and repair misspelt words in the text, these strategies use dictionaries, language models, or statistical methods. They may be rule-based, probabilistic, or based on machine learning [33].

b. Handling Emoticons: Emoticons and emojis provide feeling or emotional clues in text. Researchers have created techniques for dealing with emoticons by mapping them to emotion labels or sentiment scores. Emoticons may be substituted with sentiment labels (e.g., positive, negative) or translated to sentiment scores using pre-existing sentiment lexicons or bespoke mappings [34].

c. Abbreviation Expansion: Abbreviations and acronyms are extensively employed in social media and online communication, making sentiment analysis difficult. Previous research has focused on extending abbreviations via the use of predetermined dictionaries or contextual information. To provide a better grasp of the material, these strategies substitute abbreviations with their extended equivalents [35].

d. Text Normalization: As previously noted, text normalization methods such as tokenization and stemming/lemmatization may also assist solve unpleasant data. Tokenization divides text into units, making it easier to handle emoticons and abbreviations as individual tokens. Stemming and lemmatization may help to eliminate variance caused by misspelt words or varied spellings of the same term [36].

e. Preprocessing Filters: You may use preprocessing filters to remove or replace specified patterns, symbols, or letters. Regular expressions or rule-based approaches may be used to filter out certain patterns that do not contribute to sentiment analysis, such as URLs, special letters, or irrelevant symbols [37].

f. Domain-Specific Lexicons: Creating domain-specific sentiment lexicons may help capture the emotion associated with misspellings, slang, or abbreviations. These lexicons might be manually curated or created by machine learning methods. They give context-specific information to increase the accuracy of sentiment analysis [38].

To address bothersome data issues in sentiment analysis, earlier research has used a mix of rule-based systems, statistical approaches, and machine learning techniques. These strategies seek to improve sentiment analysis model accuracy by properly managing misspellings, emojis, abbreviations, and other noisy factors prevalent in text data.

2.4 HOW TO DEAL WITH DENIAL AND RIDICULE

Dealing with denial and derision in sentiment analysis may be difficult since these attitudes might distort the data or add bias. Here are a few approaches that researchers have tried to overcome these concerns:

a. Contextual Analysis: It is critical to evaluate the context of rejection or mockery. To acquire a better grasp of the sentiment communicated, sentiment analysis models may be improved by taking into consideration the surrounding words and phrases. This contextual analysis may assist in distinguishing between real emotion and sarcastic or ironic words [39].

b. Fine-Grained Analysis: Sentiment analysis models may be constructed to capture a broader variety of sentiment categories rather than depending exclusively on a binary positive/negative sentiment categorization. Fine-grained sentiment analysis enables more complex categorization, such as denial, derision, sarcasm, or irony. This method aids in capturing the nuances of emotion portrayed in writing [40].

c. Sentiment Modifiers: Denial or mockery are often communicated using sentiment modifiers or intensifiers that affect the overall emotion of a statement. Sentiment analysis algorithms may better identify the intended sentiment behind the denial or mockery by including sentiment modifiers into the analysis [41].

d. Domain-Specific Lexicons: Creating or using domain-specific sentiment lexicons that contain phrases linked to denial and mockery will assist increase sentiment analysis accuracy. These lexicons, which capture the precise words and idioms associated with rejection and mockery in a given topic, may be curated manually or created using machine learning approaches.

e. Training Information Augmentation: Supplementing training data with instances of denial and scorn may help the model recognize and manage such emotions. By subjecting the model to a wide variety of emotion expressions, including denial and derision, it becomes increasingly capable of properly capturing and understanding these sentiments.

f. User Feedback and Iterative Improvement: Gathering user feedback on sentiment analysis findings and iteratively improving the model may assist in identifying and addressing denial and ridicule difficulties. User input may give insights into the intricacies of sentiment expression that automated systems may not be able to capture.

It's worth noting that dealing with denial and mockery in sentiment analysis is an ongoing research topic, and no one solution can totally remove the issues they offer. Researchers and practitioners, on the other hand, may improve the accuracy and usefulness of sentiment analysis models in dealing with denial and mockery by combining these tactics.

2.5 TAKING OUT THE FILLER WORDS AND PUNCTUATION

In the preprocessing phase of sentiment analysis tasks, it is usual practice to exclude stopwords and punctuation. Words like "the," "and," and "is" are examples of stopwords that are used often but have no real value in the context of sentiment research. Also commonly deleted are punctuation symbols like commas, periods, and exclamation points since they add nothing to the categorization of emotion.[42] How to deal with punctuation and commonly used stopwords is as follows:

Stopword removal is the first step in the sentiment analysis process, and it entails excluding frequently used terms. These words are often eliminated since they do not convey any particular emotion. Language-specific and extra domain-specific stopword lists are

provided. This procedure reduces background noise and boosts the effectiveness of further investigation.

Example:

Commentary: "I thought the movie was excellent."

The end result is "movie really good."

Punctuation is omitted so that readers may concentrate on the meaning of the words themselves without distraction. Using regular expressions or other string manipulation techniques, it is possible to remove punctuation. Common punctuation signs are usually taken out as well, including commas, periods, question marks, and exclamation marks.

Example:

Response: "The movie was great! It was incredible!!!

Results: "That was a great movie!" Amazing, right?

Sentiment analysis algorithms perform better when only relevant words and their related emotions are included. This process smooths out the data and increases the reliability of the sentiment analysis. It's worth noting that punctuation marks like em dashes and exclamation points may be maintained or utilized as features in sentiment analysis if they signal intense emotion in the context of a certain assignment.

3. METHODOLOGY

3.1 INTRODUCTION

The primary focus of this study pertains to the utilization of machine learning (ML) techniques for classification, natural language processing (NLP) for text analysis, and feature selection methods for data analysis.

To achieve successful outcomes, it is essential to gather the most significant features and employ effective feature weighting techniques. In this study, the utilization of the Python programming language is employed for the implementation of machine learning techniques.

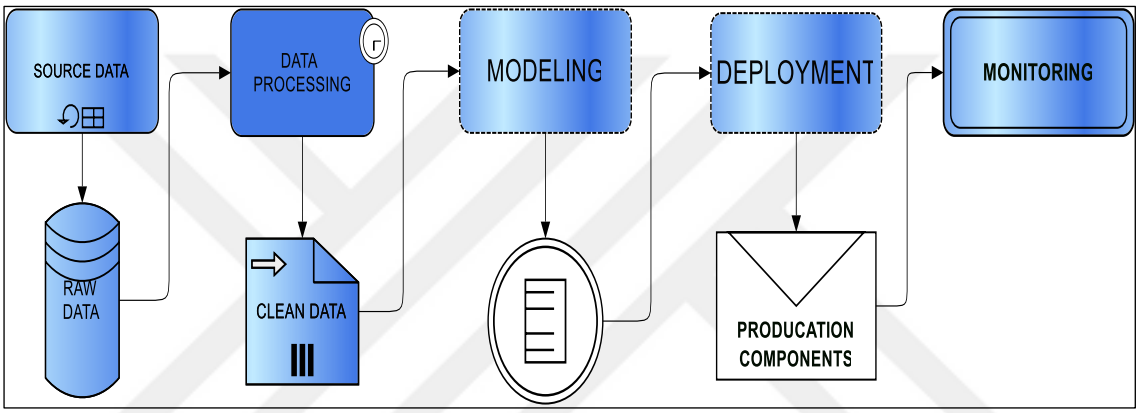


Figure 3.1: Methodology Diagram.

3.2 CLASSIFICATION METHODS

The central focus of this thesis pertains to the utilization of supervised learning techniques for the purpose of classification.

Supervised learning techniques are utilized in situations where the input data comprises of labelled training examples, with each example being linked to a known class or sentiment label. The objective of supervised learning is to acquire a function that can effectively forecast the category labels for novel, unobserved data instances by relying on the patterns and associations identified in the training data.

The present thesis outlines a series of steps involved in the application of machine learning, which are as follows:

The process of selecting pertinent features that effectively capture sentiment information is a prerequisite step prior to training machine learning models. The objective of feature selection is to ascertain the most informative and discriminative characteristics that play a

role in the classification of sentiment. This procedure aids in the reduction of dimensionality and prioritizes the most pertinent aspects of the textual data.

In the process of data representation, the selected features are transformed into a vector of numerical values. This conversion facilitates the efficient processing and analysis of data by machine learning algorithms. Diverse techniques are employed to assign numerical values to the chosen features.

The thesis employs supervised learning techniques for the purpose of sentiment analysis classification. The aforementioned algorithms undertake an analysis of the labelled training data with the objective of acquiring a function or model that can effectively generalize to data that has not been previously encountered. Upon presentation of novel, unannotated data, the acquired function is employed to forecast the affective classifications.

The thesis cites RF and DT algorithms as instances of supervised learning algorithms. The algorithms undergo training on a labelled training set and subsequently generate an inferred function that can be applied to classify the sentiment of new data instances.

The thesis utilizes supervised learning techniques, along with feature selection and data representation procedures, to apply machine learning methods to the classification task in sentiment analysis. The objective is to construct models that can precisely forecast sentiment labels for novel, unobserved textual data.

3.3 THE DECISION TREE ALGORITHM

The analysis of sentiment is of paramount importance in comprehending the general public's perception and attitude towards diverse subjects, such as the Coronavirus vaccine. The opinions and comments of individuals can offer valuable insights into the acceptance of vaccines, concerns surrounding them, and the general public sentiment. The utilization of machine learning algorithms, such as the decision tree algorithm, has been observed in the context of sentiment analysis tasks, with the aim of automatically categorizing sentiments conveyed in textual data.

The decision tree algorithm is a prevalent and robust supervised learning technique that is highly applicable for sentiment analysis. The employed methodology involves the utilization of a hierarchical framework to partition the data according to feature-based criteria [43], culminating in the inference of sentiment classifications. The decision tree algorithm can be

employed in the domain of sentiment analysis pertaining to the Corona vaccine to categorize comments into positive, negative, or neutral sentiment categories.

The utilization of the decision tree algorithm in sentiment analysis pertaining to the Corona vaccine yields a number of discernible benefits. To begin with, decision trees possess the ability to effectively capture intricate associations between characteristics and sentiment classifications. The models are capable of processing both categorical and numerical features, rendering them versatile in accommodating diverse forms of textual data. Furthermore, decision trees provide interpretability as the decision rules acquired during the training process can be comprehended and scrutinized with ease.

The decision tree algorithm exhibits a high degree of suitability for sentiment analysis tasks, particularly those that require a comprehensive comprehension of the underlying features and decision rules that govern the expression of sentiment. The utilization of decision trees in sentiment analysis pertaining to the Corona vaccine can facilitate the identification of pivotal features or lexicons that are causally linked to distinct sentiments. This approach can be instrumental in discerning the determinants that shape public perception [44].

In addition, decision trees have the capability to address sentiment classification problems that involve both binary and multi-class classifications. Binary classification is a method of categorizing sentiments into two distinct categories, namely positive and negative. However, multi-class classification is a more sophisticated approach that enables the identification of positive, negative, and neutral sentiments. This approach facilitates a more nuanced analysis of public sentiment towards the Corona vaccine.

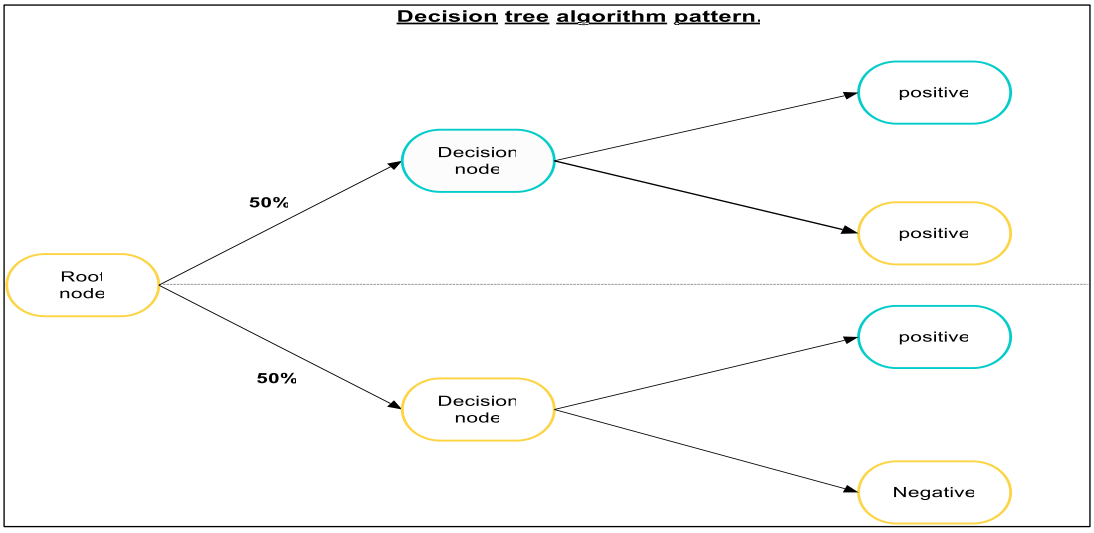


Figure 3.2: Decision Tree Algorithm Pattern.

Nonetheless, it is crucial to contemplate plausible constraints of the decision tree algorithm in the context of sentiment analysis. Decision trees are susceptible to overfitting, a phenomenon in which they may excessively memories the training data and exhibit suboptimal performance when presented with new, unseen data. In order to address this issue, it is possible to utilize various methodologies such as pruning and ensemble techniques such as Random Forests to ameliorate the generalization capabilities and augment the efficacy of the decision tree model [45].

The utilization of the decision tree algorithm presents a proficient methodology for conducting sentiment analysis within the framework of the Corona vaccine. Decision trees can be utilized to comprehend public sentiment, recognize crucial factors that influence vaccine acceptance, and provide insights for public health campaigns by exploiting their capacity to capture intricate relationships between features and sentiments. It is imperative to acknowledge the potential concerns of overfitting and to investigate supplementary methodologies for enhancing the efficacy of the decision tree model in the context of sentiment analysis assignments.

3.4 THE RANDOM FOREST ALGORITHM

The Random Forest algorithm is an ensemble learning technique that integrates numerous decision trees to enhance the precision of predictions. The operational mechanism involves the creation of numerous decision trees on diverse data subsets, followed by the consolidation of their forecasts. The Random Forest algorithm can be utilized for sentiment analysis pertaining to the Corona vaccine, with the aim of categorizing comments into distinct sentiment categories, namely positive, negative, or neutral [46].

The utilization of the Random Forest algorithm presents numerous benefits in the context of sentiment analysis assignments. Initially, the model is capable of capturing intricate associations between characteristics and affective labels, while considering the interplay and interdependence among distinct features. This feature renders it appropriate for the examination of sentiment conveyed in textual data, which frequently encompasses complex patterns and contextual interrelationships.

Furthermore, Random Forests exhibit a higher degree of resilience against overfitting in comparison to standalone decision trees. Random Forests mitigate the possibility of overfitting noise or irrelevant patterns in the training data by creating numerous decision

trees on distinct subsets of the data and combining their forecasts. The augmentation of the model's generalization capability facilitates its ability to effectively perform on novel data and enhance the overall precision of sentiment categorization.

Random Forests provide interpretability through the provision of feature importance measures. The algorithmic approach facilitates the identification of the most impactful features that contribute to determining the sentiment, thereby enabling researchers to gain valuable insights into the fundamental drivers of public sentiment towards the Corona vaccine. This data can be utilized to inform public health initiatives and effectively address particular concerns or misunderstandings.

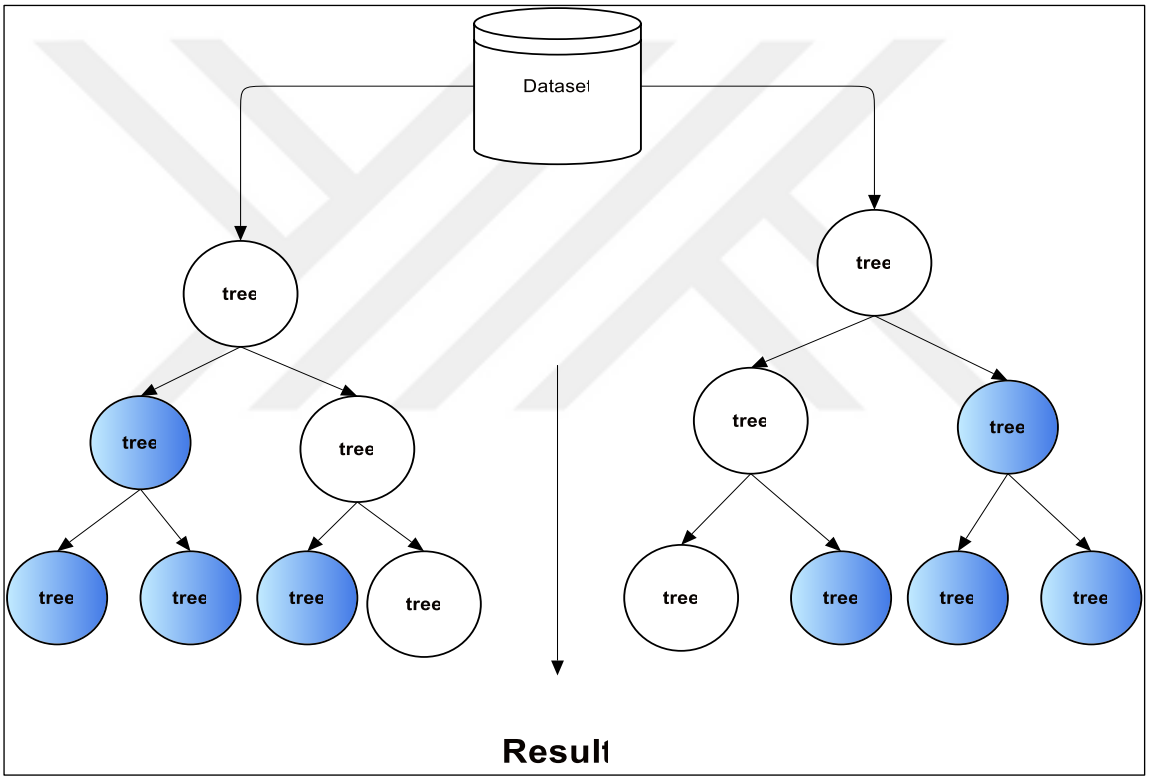


Figure 3.3: Structure of a Random Forest.

Moreover, Random Forests have the capability to address sentiment classification tasks that involve both binary and multi-class classification. The flexibility inherent in sentiment analysis for the Corona vaccine allows for the categorization of sentiments into various classifications, including positive, negative, neutral, and potentially more intricate sentiment labels. This enables a thorough examination of the various sentiments conveyed by the populace.

It is noteworthy that Random Forests necessitate meticulous contemplation of hyperparameter tuning and model optimization. The Random Forest model's performance in sentiment analysis tasks can be significantly influenced by various factors such as the selection of an appropriate number of trees, control over tree depth, and fine-tuning of other associated parameters. Furthermore, the utilization of feature selection and preprocessing methodologies, such as text normalization and feature engineering, is imperative for the extraction of informative features and enhancement of sentiment classification accuracy.

The utilization of the Random Forest algorithm presents a robust methodology for conducting sentiment analysis within the framework of the Corona vaccine. Random Forests can offer significant contributions to the field of public health by utilizing its capacity to capture intricate relationships, manage diverse sentiment categories [47], and offer interpretive capabilities. Specifically, it can provide valuable insights into public sentiments, facilitate comprehension of the underlying factors that influence vaccine acceptance, and direct targeted public health interventions. The optimization of models and engineering of features are essential factors in maximizing the performance of Random Forests when applied to sentiment analysis tasks pertaining to the Corona vaccine.

3.5 LINEAR REGRESSION ALGORITHM

The supervised learning algorithm of linear regression is utilized to model the correlation between a dependent variable, namely sentiment, and one or more independent variables, commonly referred to as features. Linear regression is a commonly employed technique for forecasting continuous variables. However, it can also be modified to facilitate sentiment analysis by associating numerical values with sentiment labels. Linear regression is a viable method for predicting sentiment scores or intensity in the domain of sentiment analysis pertaining to the Corona vaccine [48].

The utilization of linear regression in the domain of sentiment analysis presents numerous benefits. Initially, it enables a numerical evaluation of the emotions conveyed in remarks. Linear regression offers a sentiment score that is continuous in nature, as opposed to discrete sentiment categories, thereby enabling the capture of the sentiment's intensity or magnitude. This has the potential to offer a more intricate comprehension of the prevailing public attitude towards the Coronavirus vaccine.

Moreover, the utilization of linear regression can aid in the identification of the variables that contribute to the sentiment. Through the examination of regression coefficients, scholars are able to ascertain the impact of various characteristics on sentiment scores. This analysis has the potential to illuminate the particular facets or subjects pertaining to the COVID-19 vaccine that elicit favorable or unfavorable attitudes among the populace.

Moreover, linear regression offers interpretive capabilities. The equation and coefficients of the model facilitate a clear interpretation of the impact of each feature on the sentiment score holistically. The provision of transparent information can aid researchers and public health practitioners in comprehending the fundamental factors and formulating informed strategies to tackle concerns or enhance the uptake of vaccines.

Nevertheless, it is crucial to take into account specific constraints of linear regression when conducting sentiment analysis. The linear regression model presupposes a linear correlation between the features and the sentiment, a premise that may not be universally applicable in intricate sentiment analysis assignments. Alternative machine learning algorithms may be more effective in capturing non-linear relationships or interactions among features. Moreover, the assumption of independence among the features is inherent in linear regression, which may not be suitable for all sentiment analysis contexts.

Furthermore, it is worth noting that linear regression may not possess the capability to directly handle categorical features. In order to incorporate categorical variables into linear regression models, it is necessary to apply encoding methods such as one-hot encoding or ordinal encoding. The selection of an encoding approach necessitates meticulous contemplation contingent upon the characteristics of the categorical variables and their pertinence to the task of sentiment analysis.

The utilization of linear regression is a viable approach for conducting sentiment analysis pertaining to the Corona vaccine [49]. This method can furnish numerical sentiment scores and pinpoint influential factors. The tool's capacity to comprehend and capture significant features renders it a valuable asset in comprehending public sentiment. It is imperative to acknowledge the constraints of linear regression and investigate its suitability in tandem with alternative machine learning methodologies to attain superior precision in sentiment analysis outcomes.

3.6 GRADIENT BOOSTING CLASSIFIER ALGORITHM

The Gradient Boosting Classifier is an ensemble learning technique that amalgamates several weak learners, usually decision trees [50], to form a robust predictive model. The methodology involves the iterative construction of novel models that prioritize the identification and rectification of errors committed by antecedent models. The Gradient Boosting Classifier algorithm is capable of acquiring knowledge from textual data in the domain of sentiment analysis, enabling it to categorize comments into sentiment classes, including positive, negative, or neutral.

The utilization of Gradient Boosting Classifier presents various benefits in the context of sentiment analysis endeavours concerning the Corona vaccine. Initially, the system is capable of managing intricate associations among characteristics and attitudes. The iterative characteristic of the algorithm enables it to acquire knowledge from errors and enhance the model's efficiency with every iteration. This characteristic renders it highly appropriate for conducting sentiment analysis, as the expression of sentiment in textual data can be subject to diverse influences.

Furthermore, the Gradient Boosting Classifier is capable of addressing sentiment classification problems that involve both binary and multi-class scenarios. The utilization of a multi-category sentiment classification system facilitates a more exhaustive evaluation of public sentiment towards the Corona vaccine. The attribute of flexibility holds significant value in capturing the wide spectrum of sentiments conveyed by individuals.

Furthermore, the Gradient Boosting Classifier models are renowned for their elevated predictive precision. The algorithm can enhance sentiment classification outcomes by detecting intricate patterns and dependencies in textual data through the integration of several weak learners. The attainment of precision is of utmost importance in acquiring dependable perceptions regarding the general outlook of the public and in guiding efficacious public health measures [51].

Nevertheless, it is crucial to take into account specific factors when utilizing the Gradient Boosting Classifier algorithm for sentiment analysis purposes. The utilization of Gradient Boosting models entails a significant amount of computational resources and necessitates meticulous parameter calibration to achieve optimal performance. Achieving a balance

between the intricacy of a model and its ability to generalize is of utmost importance in order to avoid overfitting and guarantee effective generalization to unobserved data.

In addition, the utilization of feature engineering and preprocessing methodologies is crucial in augmenting the efficacy of the Gradient Boosting Classifier. The process of normalizing text, selecting relevant features, and extracting pertinent information is crucial in enhancing a model's capacity to effectively capture significant sentiment information from textual data pertaining to the Corona vaccine.

The Gradient Boosting Classifier algorithm has demonstrated significant efficacy as a tool for sentiment analysis within the specific context of the Corona vaccine. The tool's proficiency in capturing intricate relationships, managing various sentiment categories, and attaining a notable level of predictive precision renders it apt for scrutinizing sentiments conveyed in textual data. The optimization of the model's performance and the acquisition of valuable insights into public sentiment towards the Corona vaccine require meticulous parameter tuning and feature engineering.

3.7 NATURAL LANGUAGE PROCESSING

Natural Language Processing (NLP) is a branch of computer science that studies and analyzes human language. It entails creating algorithms and models that allow computers to analyze, understand, and create natural language content. Because it seeks to replicate human-like language processing, NLP is considered a subset of artificial intelligence (AI) [52].

An NLP system can do things like machine translation, question answering, and text paraphrasing. These systems seek towards objectives such as comprehending and producing human language. Morphological analysis, syntactic analysis, semantic analysis, and discourse analysis are the key processes of NLP.

Morphological analysis is the study of word structure, including meanings, suffixes, and prefixes. The study of grammatical structure and word connections is the subject of syntactic analysis. Semantic analysis gives meaning to the structures that syntactic analysis produces. Discourse analysis is concerned with the context supplied by previous phrases, which might influence the meaning of later statements.

NLP offers a broad variety of applications and has been widely employed to simplify human lives during the last decade. People may voice their ideas on a variety of issues thanks to the

growing usage of the internet [53], and this knowledge has become a significant resource for a variety of reasons. attitude analysis and opinion mining are NLP subfields that entail reviewing documents to identify whether they are subjective or neutral, and if subjective, whether they communicate a good or negative attitude. Text emotion analysis, which automatically captures emotions conveyed in text, is becoming increasingly popular in NLP applications.

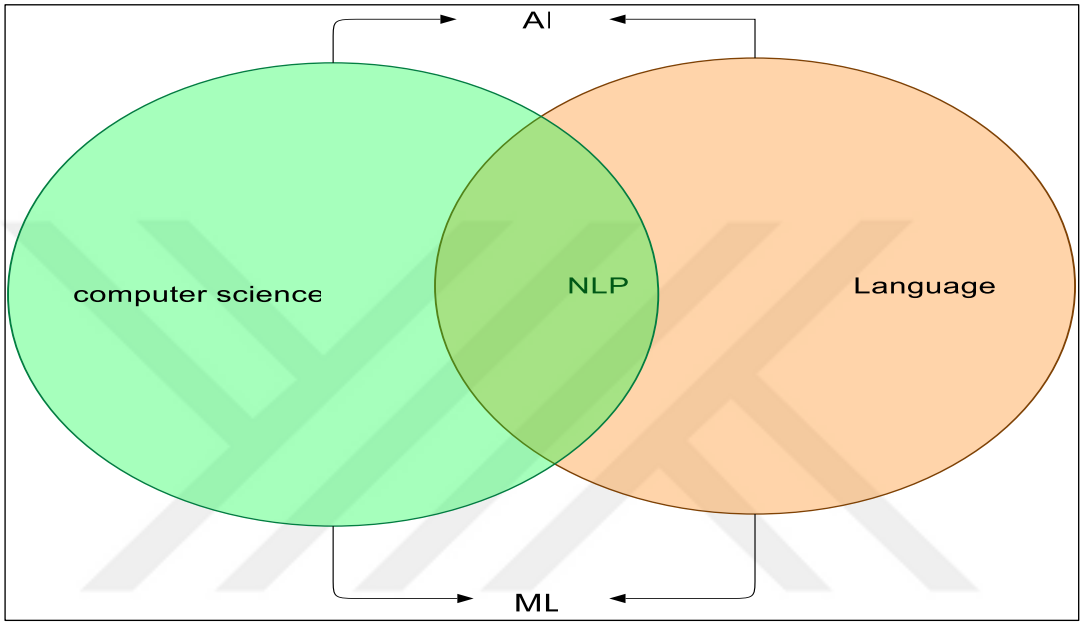


Figure 3.4: Natural Language Processing Approaches

Combining natural language processing (NLP) approaches with machine learning (ML) methodologies has shown promising results in text emotion categorization. To do this, each word in a text is morphologically examined and relevant information is extracted using an NLP tool. NLP technologies are also used for part-of-speech (POS) tagging, which entails classifying each word with its matching grammatical category (e.g., nouns, verbs, adjectives, adverbs). While there are several NLP tools available in English that use sentence structure for POS tagging, comparable tools are few in Turkish. Zemberek is a well-known Turkish NLP tool that can do morphological analysis and POS tagging at the word level [54].

3.8 DATA PRE-PROCESSING

3.8.1 Spell Correction

The use of emojis in textual communication can effectively convey significant emotional and attitudinal information. Prior studies have concentrated on the integration of emojis

within models for analyzing sentiment. The process entails the assignment of sentiment labels to emojis or the inclusion of emojis as distinct features in the analysis. Emojis possess the capacity to function as symbolic depictions of affective states and can furnish significant perspectives for the analysis of sentiment.

The utilization of abbreviations and acronyms is prevalent in social media and informal written communication, which presents difficulties for the analysis of sentiment. Prior research has addressed this matter through the creation of abbreviation dictionaries or the utilization of contextual cues to expand abbreviations into their complete forms. This aids in guaranteeing that the sentiment analysis algorithms can precisely comprehend the connotation of the given text.

The utilization of sentiment lexicons is a crucial aspect in the field of sentiment analysis. The corpus comprises lexemes or expressions that have been marked with polarities of sentiment [55]. In order to effectively manage data that is characterized by high levels of noise or annoyance, scholars have undertaken the task of augmenting pre-existing sentiment lexicons to encompass a wider range of linguistic features such as misspelled words, abbreviations, and colloquialisms that are frequently encountered in informal text or social media. This approach facilitates the capture of the sentiment conveyed by diverse expressions, thereby augmenting the precision of sentiment analysis.

The phenomenon of domain adaptation refers to the challenge of applying sentiment analysis models that are trained on general corpora to specific domains such as healthcare or vaccines, where their performance may be suboptimal. Domain adaptation techniques have been utilized as a means to surmount this challenge. The aforementioned methodologies entail the refinement of sentiment analysis models through the utilization of labeled data that is specific to the domain or the implementation of transfer learning techniques that leverage pre-existing models and tailor them to the intended domain. The utilization of domain adaptation techniques has been observed to enhance the efficacy of sentiment analysis models in particular contexts, such as the analysis of sentiments pertaining to vaccines [56]. Contextual embeddings are utilized in sentiment analysis models to address the limitations of bag-of-words representations, which may not effectively capture the contextual nuances inherent in the text. To overcome this constraint, scholars have investigated the utilization of contextual embeddings, such as word embeddings that have been trained on extensive language models like BERT or GPT. The implementation of contextual embeddings

facilitates a more sophisticated comprehension of the textual content through the incorporation of the encompassing context. This technique facilitates the identification of the sentiment conveyed in data that may be characterized by high levels of noise or irritation.

Ensemble techniques refer to the integration of multiple sentiment analysis models with the aim of enhancing the overall performance. Ensemble methods can effectively address the issue of noisy or disruptive data through the utilization of the varied predictions generated by multiple models. Ensemble techniques have the ability to alleviate the influence of various forms of noise in data, such as misspellings, emojis, and abbreviations, resulting in more resilient sentiment analysis outcomes. This is achieved by consolidating the outputs of multiple models.

3.8.2 Emoticon Handling

To accurately capture the sentiment conveyed by emoticons. The utilization of emoticons in digital communication has been linked to the expression of positive or negative sentiment. Specifically, a smiley face emoticon such as ":)" is commonly associated with positive sentiment, whereas a frowning face emoticon like ":(" is typically associated with negative sentiment.[57]

Several scholarly investigations have also taken into account the placement and surroundings of emoticons within the textual content. The sentiment impact of an emoticon may be modified by its proximity to certain words or phrases. As an illustration, the inclusion of an emoticon in proximity to a favorable term has the potential to intensify the affirmative connotation or counteract the adverse connotation conveyed in the adjacent language.

Apart from the categorization of sentiments, scholars have investigated the attribution of sentiment scores to emoticons. The task at hand pertains to the measurement of the emotional tone expressed by individual emoticons on a scale that is uninterrupted. Sentiment scores may be obtained through the utilization of sentiment lexicons or generated via machine learning methodologies. The utilization of these scores can offer a more detailed and nuanced depiction of emotional tone, thereby serving as a valuable resource in endeavors related to sentiment analysis.

Moreover, certain research endeavors have expanded the examination beyond conventional emoticons and encompassed a broader spectrum of emotive icons and emojis. Emojis are pictorial symbols that depict emotions and objects, and their usage has become more

prevalent in digital communication. Scholars have devised methodologies for interpreting and scrutinizing the emotional connotations conveyed by emojis through the process of associating them with sentiment labels or sentiment scores, which are determined by their visual representation or customary employment.

It is imperative to acknowledge that the comprehension of emoticons and emojis may exhibit variability among diverse cultures and individuals. Hence, it is imperative to have a contextual comprehension and consider cultural nuances to precisely capture the intended meaning conveyed by these visual representations.

3.8.3 Text Normalization

The implementation of text normalization techniques is of paramount importance in effectively managing troublesome data in the context of sentiment analysis. The utilization of techniques such as tokenization and stemming/lemmatization facilitates the standardization of textual data, minimizes extraneous information, and enhances the precision of sentiment analysis.[58] The following strategies can be employed to effectively tackle particular obstacles:

The process of tokenization entails the segmentation of a given text into discrete units, which are commonly referred to as tokens or words. This procedure facilitates the discrete tokenization of emoticons and abbreviations, enabling sentiment analysis algorithms to handle them in a suitable manner. One possible approach is to consider emoticons such as ":)" as discrete tokens that convey positive sentiment, and to treat abbreviations such as "lol" as distinct tokens that carry their own implications for sentiment analysis.
Tokenization facilitates the detection and management of erroneous or troublesome data at a finer granularity. The process aids in identifying lexemes, collocations, and n-grams (consecutive sequences of n words) that convey the underlying sentiment conveyed in the given text.

Stemming and lemmatization are linguistic processes that seek to minimize lexical variation by reducing different inflected forms of a given word to their respective base or root form. The normalization process serves the purpose of mitigating misspellings and variations in word inflections, which are frequently encountered in user-generated content or social media data.

Stemming is a linguistic process that entails the elimination of suffixes from words to derive their stem. Conversely, lemmatization is a technique that seeks to convert words to their base or dictionary form, also known as the lemma. Stemming is a process that involves reducing words to their base or root form, such as "running," "runs," and "ran" being converted to the stem "run." On the other hand, lemmatization is a technique that transforms words to their canonical form, or lemma, such as "running," "runs," and "ran" being transformed to the lemma "run."

The utilization of stemming or lemmatization techniques in sentiment analysis models enables the treatment of diverse forms of a word as identical entities, thereby mitigating the interference caused by misspelled words or discrepancies in word forms. Utilizing this approach enhances the precision of capturing the fundamental sentiment of the text and promotes the overall generalizability of sentiment analysis models.

Preprocessing steps such as tokenization, stemming, and lemmatization are crucial in sentiment analysis for text normalization purposes. Text normalization techniques aid in managing problematic data, such as emoticons, abbreviations, misspellings, and lexical variations, by establishing a uniform representation of the text. The implementation of these techniques serves to enhance the caliber and dependability of sentiment analysis outcomes, thereby facilitating a more precise comprehension of the general sentiment towards the Corona vaccine.

3.8.4 Stop Word Removal

Stop word elimination is a method often employed in natural language processing jobs such as sentiment analysis. Stop words are words that appear often in a language but have no meaningful significance or contribute to the comprehension of the text. Articles (e.g., "the," "a," "an"), prepositions (e.g., "in," "on," "at"), conjunctions (e.g., "and," "or," "but"), and pronouns (e.g., "I," "you," "he") are examples of stop words.[59]

There are various advantages of removing stop words from a text:

a. Noise Reduction: Stop words exist often in text but give no context or useful information. We may lessen the noise in the data and concentrate on the more significant material by deleting these terms.

b. Memory and Storage Efficiency: Stop words are frequently used words that appear often in a text corpus. We may lower the size of the dataset by deleting them, resulting in more efficient memory consumption and storage.
c. Increased Computational Efficiency: Removing stop words may help speed up text data processing. Because stop words are so abundant, they add to the total processing time when operations like tokenization, stemming, or lemmatization are performed. We may lower the computational effort and increase the efficiency of sentiment analysis algorithms by removing these terms.

Researchers often create or alter existing stop word lists to meet the needs of their sentiment analysis assignment. The updated list may omit words judged relevant in the current context while retaining generic stop words. This personalization improves the accuracy of emotion prediction and sentiment categorization.

3.8.5 Negation Handling

Negation handling is an essential feature of sentiment analysis since it allows for the capturing of negated or reversed sentiment conveyed in text. When words or phrases are used to represent the opposite of or negate the feeling of another word or phrase in the sentence, this is referred to as negation. In the line "I am not happy," for example, the word "not" contradicts the feeling indicated by the word "happy."
Negation must be handled carefully since the existence of negation may radically shift the emotional polarity of a statement. Negating might result in erroneous sentiment analysis findings. Several ways to negation handling in sentiment analysis have been proposed:

Negation Markers: Identifying and marking negation words or phrases in the text is one way. Negation is indicated by negation markers such as "not," "no," "never," "without," and so on. Sentiment analysis algorithms can determine the scope of negation and accurately interpret the sentiment of the linked words or phrases by recognizing these markers.

Another option is to use dependency parsing tools to assess the syntactic structure of the statement. Dependency parsing aids in the identification of links between words in a phrase. The system can recognize words that are changed or negated by other words in the text by examining the dependencies. This data may be utilized to determine the right sentiment polarity.

emotion adjusting: Negation handling may also include adjusting the emotion polarity of words or phrases that fall within the purview of negation. If the word "happy" is negated, the feeling polarity may be inverted too "unhappy" or "not happy." This shifting may be accomplished via the use of established rules, sentiment lexicons, or machine learning models.

Contextual Analysis: Contextual analysis is essential when dealing with negations. A word's or phrase's emotion may be changed by its context. In the statement "I am not as happy as I thought," for example, the negative "not" denies the feeling "happy," but the comparative phrase "as happy as I thought" provides subtlety. Considering the larger context helps in appropriately comprehending the text's mood.

3.9 DATASET

The corpus comprises of Twitter posts that were obtained through the utilization of the Twitter API and a Python script. The objective of gathering these tweets is to conduct an analysis of the subject matter pertaining to the hashtag "#covid19." On July 25, 2020, the process of collecting data was initiated, which involved an initial set of 17,000 tweets. As a result, a daily collection of tweets has been conducted to obtain a more extensive sample size over a period.

The precise particulars of the script utilized for collecting data can be located within the GitHub repository that has been furnished for this purpose, which can be accessed via the following URL: https://github.com/gabrielpreda/covid-19-tweets. It is probable that the script employs the Twitter API to extract tweets containing the hashtag "#covid19" and stores them in a structured format suitable for analysis.

The dataset's predominant components comprise the textual and metadata elements of the amassed tweets. Every tweet is linked with the hashtag "#covid19" to signify its pertinence to the ongoing COVID-19 pandemic. The dataset presents a significant opportunity for investigating and comprehending the communal discussion, emotions, and viewpoints concerning COVID-19 on the Twitter platform.

This dataset can be utilized by researchers and data analysts to conduct a range of analyses, including sentiment analysis, topic modeling, trend detection, and comprehension of the discourse patterns surrounding COVID-19 on social media. The dynamic nature of the

dataset facilitates longitudinal investigations and enables the monitoring of alterations in sentiment and subjects over an extended period.

It should be noted that the dataset may have undergone substantial expansion subsequent to its initial compilation, owing to the regular incorporation of daily updates.

3.10 DATA ENCODING

Data encoding is a critical step in preparing the retrieved features for input into the decision tree algorithm. The aim is to encode each remark in a numerical representation that the algorithm can understand. This procedure entails transforming textual input into a vectorized representation in which each feature corresponds to a single characteristic or word.

One-hot encoding and term frequency-inverse document frequency (TF-IDF) encoding are two extensively used encoding approaches for sentiment analysis [60].

a. One-Hot Encoding: One-hot encoding is a method that expresses each word or attribute as a binary vector. In this method, a vocabulary is built by taking into account all unique words or properties in the dataset. Each comment is then converted into a binary vector, with each element representing the presence or absence of a certain term or attribute. This encoding approach is beneficial for collecting categorical information and may be successful in decision tree algorithms.

b. TF-IDF Encoding: The TF-IDF encoding assigns a weight to each word or attribute in a remark based on its frequency in the comment and its overall frequency in the dataset. TF-IDF stands for term frequency-inverse document frequency. The term frequency (TF) component indicates how often a word occurs in a remark, while the inverse document frequency (IDF) component evaluates the rarity of a word throughout the dataset. TF-IDF encoding gives greater weight to terms that are more essential in a given remark but less prevalent in the dataset as a whole. In sentiment analysis, this strategy aids in highlighting discriminative terms.

Both one-hot encoding and TF-IDF encoding have benefits and are appropriate for various applications. One-shot encoding is basic and easy, recording the presence or absence of words or properties. TF-IDF encoding, on the other hand, takes into consideration the value of words in individual comments as well as their overall rarity, resulting in a more nuanced representation.

The encoding strategy used is determined by the unique needs of the sentiment analysis job as well as the features of the dataset. Experimenting with various encoding techniques may assist identify which one produces the best results when categorizing feelings using the decision tree algorithm.

c. Word Embeddings: Word embeddings depict words as dense vectors in a high-dimensional space, where the spatial connections between words convey semantic meaning. Word embedding techniques such as Word2Vec, GloVe, and FastText learn continuous vector representations by taking into account the context in which words occur. These embeddings may be pre-trained on huge corpora or learned especially on the sentiment analysis dataset. Word embeddings record semantic links between words and may enhance the effectiveness of sentiment analysis algorithms such as decision trees.

d. Bag-of-Words (BoW) Encoding: The bag-of-words encoding treats each remark as a frequency vector, with each element representing the count of a single word in the comment. This method disregards the order of words and simply evaluates their frequency. BoW encoding is easy to construct and can capture the existence and frequency of words in comments. It is often employed in classic machine learning techniques, such as decision trees [61].

e. N-gram Encoding: N-gram encoding uses sequences of N consecutive words as characteristics. Instead of looking at individual words, it considers the context and co-occurrence of terms in a remark. A 2-gram encoding, for example, considers pairs of words, while a 3-gram encoding considers triplets of words. N-gram encoding may capture local context and enhance sentiment interpretation in lengthier phrases or expressions.

f. Subword Encoding: Subword encoding approaches, such as Byte Pair Encoding (BPE) or Sentence Piece, divide words into smaller subword units. This is especially handy for managing out-of-vocabulary terms and recording morphological variants. Subword encodings may capture the meaning of individual subword units and enhance the generalization of sentiment analysis algorithms.

The encoding strategy used is determined by the features of the dataset, the available resources, and the difficulty of the sentiment analysis job. To discover the best successful solution for a specific sentiment analysis task, it is usual to experiment with several encoding approaches and evaluate their effectiveness.

3.11 EVALUATION METRICS

Evaluation metrics are critical in evaluating the success of sentiment analysis algorithms. Here are some of the most often used sentiment analysis assessment metrics:

a. Accuracy: Accuracy assesses the overall accuracy of sentiment predictions by comparing them to ground truth labels. It is determined as the proportion of properly categorized occurrences to total instances. However, when dealing with unbalanced datasets, accuracy alone may not give a thorough insight of the model's performance.

b. Precision: Precision is the fraction of accurately predicted positive cases out of all positive instances anticipated. It assesses the model's ability to detect false positives. Precision is measured by dividing the number of true positive cases by the total number of true positive and false positive instances.

c. Recall (Sensitivity): The percentage of accurately anticipated positive cases out of all actual positive instances is measured by recall, also known as sensitivity or true positive rate. It indicates the model's capacity to recognize all favorable situations. The ratio of true positive occurrences to the total of true positive and false negative instances is used to determine recall.

d. F1-score: The harmonic mean of accuracy and recall is the F1-score. It gives a balanced assessment criteria that takes accuracy and recall into account at the same time. The F1-score is especially valuable when there is an imbalance in the dataset's positive and negative events.

2 * (precision * recall) / (precision + recall) is the formula.

These assessment criteria are often used to measure the effectiveness of classification models in sentiment analysis activities. However, while deciding which indicators to prioritize, it is critical to evaluate the unique needs and peculiarities of the sentiment analysis situation at hand. Precision, for example, may be more significant in applications where minimizing false positives is key, while recall may be more important in circumstances where recognizing all positive occurrences is a priority.

Multiple assessment indicators should be reported to acquire a thorough picture of the model's performance and to make educated choices based on the unique demands and objectives of the sentiment analysis activity.

4. RESULTS & DISCUSSION

4.1 INTRODECTION

After we have finished presenting the methodology, we will then provide the results that were acquired from (LR), (GBC), (DT), and (RF), along with some notes.
After first dividing the data from the random samples into test and training sets, we examined the reliability of each of the sets.

4.2 FEATURE ENCODING

Categories found in attributes cannot be processed by computers. Categorical data must be converted to numeric form before being entered into the model as input attributes. In this technique, an integer is designated for each distinct nominal variable. Input characteristics are quantified by our team. Put simply:

4.2.1 TF-IDF Transformer

The TF-IDF (Term Frequency-Inverse Document Frequency) transformer is a common tool in natural language processing (NLP) and information retrieval difficulties. Its goal is to convert a series of text documents into a numerical representation that may be used by machine learning algorithms or other downstream processes.
The TF-IDF statistic assesses the importance of a word within a collection or corpus of documents. It is divided into two primary sections:
The frequency with which a word occurs in a text is determined by this statistic. It is calculated by dividing the number of times a phrase appears in a document by the total number of phrases in that text.

Inverse Document Frequency (IDF): This statistic measures the rarity or uniqueness of a term over a collection of documents. It is calculated using the logarithm of the ratio of the total number of documents in the collection to the number of documents containing the phrase.
The TF-IDF score of a phrase is computed by multiplying its TF and IDF values. A higher TF-IDF score indicates that a phrase is both frequent inside the document and unusual throughout the whole collection, perhaps making it more informative or distinctive.

The TF-IDF transformer accepts a collection of text documents as input and performs the following operations:
Tokenization: Texts are commonly broken into individual words or tokens using techniques like word tokenization or n-gram tokenization.

Term Frequency Calculation: For each document, the frequency of each term is computed, indicating how often each term appears within that document.
Calculation of Inverse Document Frequency (IDF): The IDF value for each phrase is calculated based on its occurrence throughout the whole document collection.

Weighting for TF-IDF: The TF-IDF score for each term in each document is calculated by multiplying the term frequency by the inverse document frequency.
The resulting TF-IDF representation of the documents may then be used for various NLP tasks such as text classification, information retrieval, clustering, or document similarity calculations.

In actuality, the scikit-learn Python package offers a TfidfTransformer TF-IDF transformer class that implements these steps and allows you to apply TF-IDF transformation to your text data with ease.

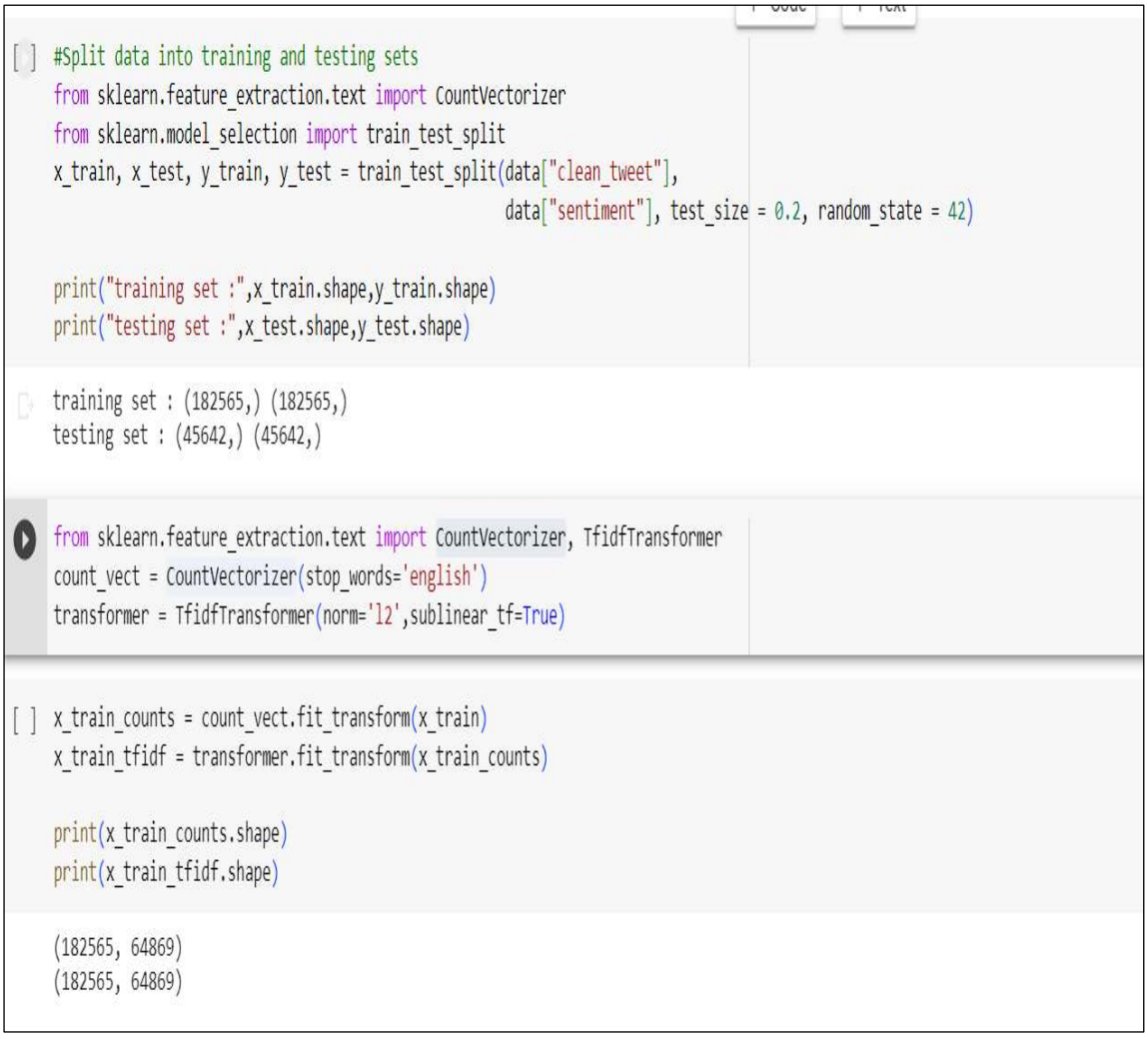
4.2.2 Count Vectorizer

Count Vectorizer is a commonly employed instrument in the field of natural language processing (NLP) that facilitates the conversion of textual data into numerical representations. The process of converting textual data into a matrix of token counts is a rapid and straightforward method.
The Count Vectorizer class is offered by the scikit-learn Python package, and its functionality encompasses:
Tokenization refers to the technique of segregating distinct phrases or tokens from unprocessed textual documents. This phase may encompass methodologies such as word tokenization or n-gram tokenization.

The Count Vectorizer algorithm generates a lexicon of distinct expressions derived from the tokenized resources. Each unique expression is allocated to a specific attribute or column within the resulting matrix.

The Count Vectorizer method tallies the frequency of each phrase within the vocabulary for every document. A matrix is produced wherein each row corresponds to a document and each column corresponds to a vocabulary term. The matrix values denote the frequency of occurrence of individual phrases across multiple documents.

The process of vectorization involves representing textual data in a numerical format, typically resulting in a count matrix. This matrix serves to depict the documents in a quantitative manner. The data has the potential to be inputted into machine learning algorithms or employed in subsequent downstream operations.



```
[ ] #Split data into training and testing sets
    from sklearn.feature_extraction.text import CountVectorizer
    from sklearn.model_selection import train_test_split
    x_train, x_test, y_train, y_test = train_test_split(data["clean_tweet"],
                                                        data["sentiment"], test_size = 0.2, random_state = 42)

    print("training set :",x_train.shape,y_train.shape)
    print("testing set :",x_test.shape,y_test.shape)

    training set : (182565,) (182565,)
    testing set : (45642,) (45642,)

    from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
    count_vect = CountVectorizer(stop_words='english')
    transformer = TfidfTransformer(norm='l2',sublinear_tf=True)

[ ] x_train_counts = count_vect.fit_transform(x_train)
    x_train_tfidf = transformer.fit_transform(x_train_counts)

    print(x_train_counts.shape)
    print(x_train_tfidf.shape)

    (182565, 64869)
    (182565, 64869)
```

Figure 4.1: Tf-Idf Transformer and Count Vectorizer.

4.3 DATA DESCRIPTION

The file in csv format comprises 228207 rows and 16 columns of data. This dataset comprises diverse remarks pertaining to the coronavirus disease, encompassing affirmative,

adverse, and neutral comments concerning the identical subject matter. The objective of this study is to categorize the comments into three distinct sections. Prior to this, the data is subjected to processing and partitioning. The data is partitioned into two sets, with 20% allocated for testing and 80% allocated for training.

We have divided the emotions into three sections using the Python textblob library and it is as follows.

replace_map = {'sentiment': {'Neutral': 0, 'Positive': 1, 'Negative':2}}

```
from textblob import TextBlob
def get_sentiment(text):
    blob = TextBlob(text)
    sentiment_polarity = blob.sentiment.polarity
    sentiment_subjectivity = blob.sentiment.subjectivity
    if sentiment_polarity > 0:
        sentiment_label = 'Positive'
    elif sentiment_polarity < 0:
        sentiment_label = 'Negative'
    else:
        sentiment_label = 'Neutral'
    result = {'polarity':sentiment_polarity,
              'subjectivity':sentiment_subjectivity,
              'sentiment':sentiment_label}
    return result
```

Figure 4.2: Sentiment Code.

4.4 CLASSIFICATION MODELS AND PARAMETERS

An analysis of the methodology employed, including the resultant outputs of the algorithms and the specific parameters employed for each algorithm.

4.4.1 Random Forest RF

Is a collection of decision tree predictors. The utilization of the average serves as a mechanism to mitigate overfitting and enhance the precision of forecasting outcomes. The implementation of RF necessitates oversight. Consequently, it utilizes distinct bootstraps to facilitate the training of each individual subsample, commonly referred to as a "tree," derived

from the initial training data input. The subsample size is equivalent to the primary input data size.

the parameter: RandomForestClassifier(n_estimators=128)

4.4.2 Decision Tree DT

The process of data partitioning for cost reduction purposes is determined by the decision tree algorithm, utilizing the available training samples. Due to its high readability and minimal data preprocessing requirements, this approach is frequently utilized.

the parameter: DecisionTreeClassifier(max_depth=1024)

4.4.3 Logistic Regression LR

The utilization of LR is akin to that of conventional regression models, as demonstrated in reference, wherein the objective is to examine the impact of one variable on another.

the parameter: LogisticRegression(penalty='l2',solver='lbfgs',random_state=22 , max_iter= 512 , C = 6.5).

4.4.4 Gradient Boosting Classifier GBC

The boosted model is a classification or regression decision tree algorithm that leverages prior predictions to enhance precision by learning from errors. The aforementioned model facilitates analytical learning and enables the anticipation of future occurrences. The method utilizes a tree structure and integrates weaker trees to form a robust model.

The parameter:

GradientBoostingClassifier(n_estimators=128,max_depth=32,random_state=44)

4.5 CLASSIFICATION REPORT

The classification report is a tool in the field of machine learning that is utilized to assess the effectiveness of a classification model. The provided metrics for each class within the dataset enable the user to assess the accuracy, precision, recall, F1-score, and support of the model.

4.5.1 Result The Random Forest RF

Accuracy results for the algorithm Train Score is: 0.9963300742201407 with Test Score is: 0.8939135007230182 and also appeared accuracy score: 0.89.

```
from sklearn.metrics import classification_report
print("Classification Report : ")
print(classification_report(y_test,y_pred))
```

```
Classification Report :
              precision    recall  f1-score   support

           0       0.88      0.98      0.93     25154
           1       0.91      0.84      0.87     15760
           2       0.90      0.64      0.75      4728

    accuracy                           0.89     45642
   macro avg       0.90      0.82      0.85     45642
weighted avg       0.90      0.89      0.89     45642
```

Figure 4.3: Result Classification Report for Rf Algorithm.

```
from sklearn.metrics import confusion_matrix,f1_score

from sklearn.metrics import plot_confusion_matrix
confusion_matrix(y_test,y_pred)
```

```
array([[24570,   528,    56],
       [ 2283, 13182,   295],
       [  979,   701,  3048]])
```

```
cm=confusion_matrix(y_test,y_pred)
import seaborn as sns
sns.heatmap(cm)
plt.show()
```

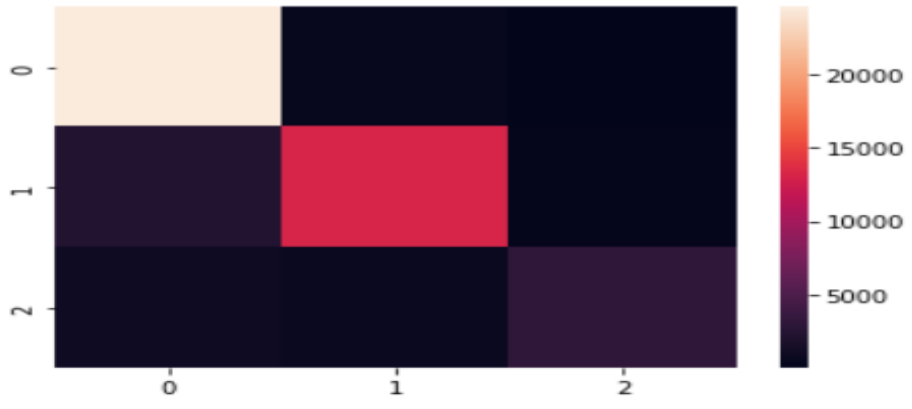


Figure 4.4: Plot Confusion Matrix.

4.5.2 Result The Decision Tree DR

Accuracy results for the algorithm Train Score is: 0.9963300742201407 with Test Score is: 0.8939135007230182 and also appeared accuracy score: 0.89.

```
from sklearn.metrics import classification_report
print("Classification Report : ")
print(classification_report(y_test,y_pred))
```

```
Classification Report :
              precision    recall  f1-score   support

           0       0.89      0.94      0.91     25154
           1       0.87      0.83      0.85     15760
           2       0.76      0.66      0.70      4728

    accuracy                           0.87     45642
   macro avg       0.84      0.81      0.82     45642
weighted avg       0.87      0.87      0.87     45642
```

Figure 4.5: Result Classification Report for DT Algorithm.

```
[38] cm=confusion_matrix(y_test,y_pred)
     import seaborn as sns
     sns.heatmap(cm)
     plt.show()
```

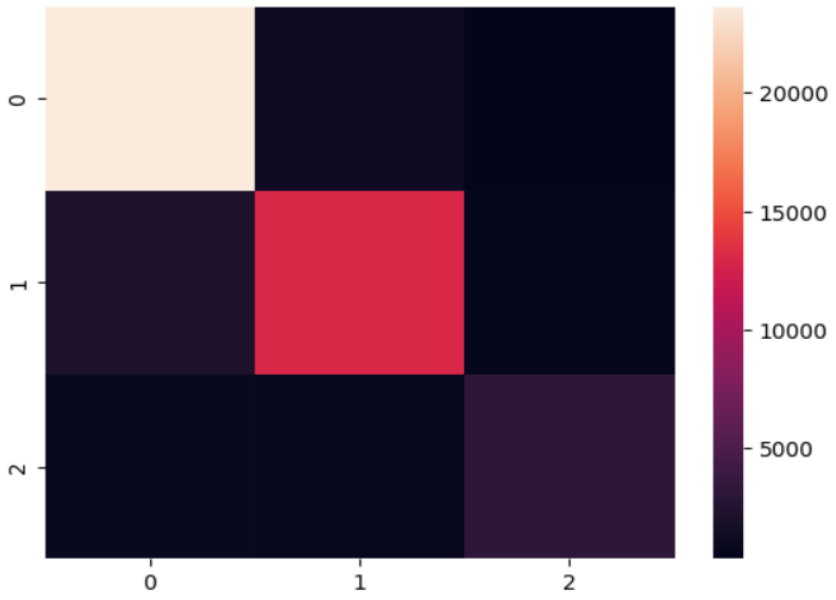


Figure 4.6: Heatmap-1.

4.5.3 Result The Logistic Regression (LR)

Accuracy results for the algorithm Train Score is: 0.9413195300304001 with Test Score is: 0.892116909863722 and also appeared accuracy score: 0.8710836510231804.

```
from sklearn.metrics import classification_report
print("Classification Report : ")
print(classification_report(y_test,y_pred))
```

```
Classification Report :
              precision    recall  f1-score   support

           0       0.89      0.94      0.91     25154
           1       0.87      0.83      0.85     15760
           2       0.76      0.66      0.70      4728

    accuracy                           0.87     45642
   macro avg       0.84      0.81      0.82     45642
weighted avg       0.87      0.87      0.87     45642
```

Figure 4.7: Result Classification Report for LR Algorithm.

4.5.4 Result The Gradient Boosting Classifier

Accuracy results for the algorithm Train Score is: 0.9413195300304001 with Test Score is: 0.892116909863722 and also appeared accuracy score: 0.8710836510231804.

```
from sklearn.metrics import classification_report
print("Classification Report : ")
print(classification_report(y_test,y_pred))
```

```
Classification Report :
              precision    recall  f1-score   support

           0       0.87      0.98      0.92     25154
           1       0.94      0.81      0.87     15760
           2       0.88      0.67      0.76      4728

    accuracy                           0.89     45642
   macro avg       0.90      0.82      0.85     45642
weighted avg       0.90      0.89      0.89     45642
```

Figure 4.8: Result Classification Report for GBC Algorithm.

```
[53] #Calculating Accuracy Score  : ((TP + TN) / float(TP + TN + FP + FN))
     AccScore = accuracy_score(y_test, y_pred, normalize=True)
     print('Accuracy Score is : ', AccScore)

     Accuracy Score is :  0.892116909863722

[54] cm=confusion_matrix(y_test,y_pred)
     import seaborn as sns
     sns.heatmap(cm)
     plt.show()
```

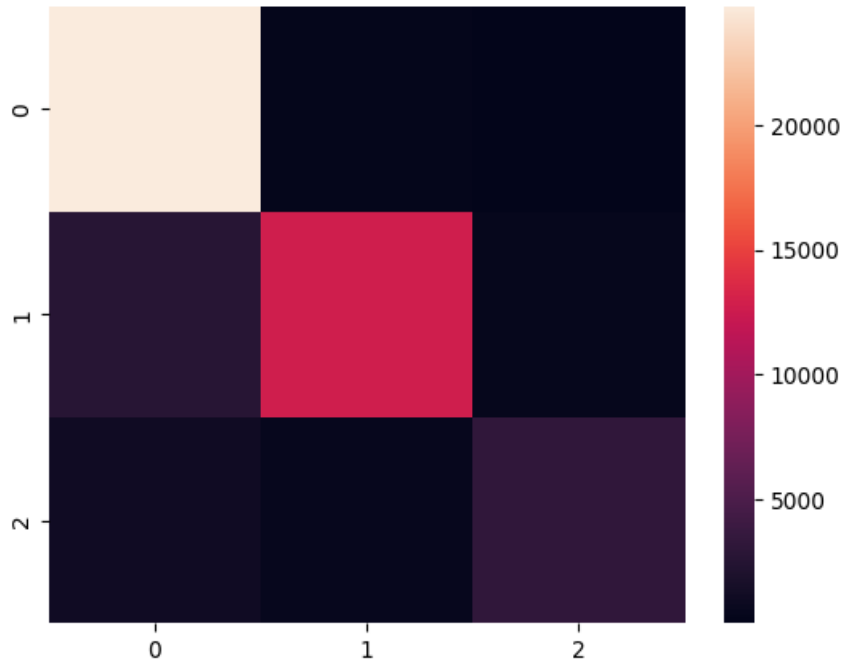


Figure 4.9: Heatmap-2.

4.5.5 Result Of The Analysis

The effectiveness of a model can be assessed by its accuracy when trained on a balanced dataset. The table below presents the accuracy values for four classifiers.

Table 4.1: Compare Accuracy Values.

Models	Accuracy (%)
Random Forests	89%
Decision Tree	87%
Logistic Regression	87%
Gradient Boosting	89%

4.6 COMPARED TO PREVIOUS ATTEMPTS

This is why we compared our results in terms of 'accuracy' to those of other research, such as As shown in Table 4.6, to make sure we've selected the best strategies for determining the most useful traits, The findings of this investigation were encouraging.

Table 4.2: A Comparison.

studies	**Models**	**Accuracy (%)**
Qorib, M., (2023).[30]	Convolutional Neural Networks (CNN)	Accuracy of 81%
Akpatsa, S. K.,	Support Vector Machine (SVM)	Accuracy of 84.3%
This study	Random Forests	89%
	Decision Tree	87%
	Logistic Regression	87%
	Gradient Boosting	89%

5. CONCLUSION

The analysis of textual data and subsequent classification into positive, negative, or neutral emotions, while considering multiple perspectives, constitutes a highly intricate undertaking within the realm of natural language processing. This task poses a significant challenge for research in the field of artificial intelligence. Hence, providing valuable insights into public sentiment is regarded as a pivotal undertaking. The task of categorizing opinions regarding coronavirus disease and its associated vaccines into positive, negative, and neutral classifications using machine learning algorithms is a multifaceted undertaking. This study employed four distinct algorithms. The classification algorithms exhibited favorable outcomes, as evidenced by the RF and GBC algorithms attaining an accuracy rate of 89%, while the LR and DT algorithms achieved an accuracy rate of 87%, in contrast. The implementation of preparatory measures for word processing and data pre-processing is considered to be a crucial set of strategies that precede the classification process. The aforementioned insights possess the capacity to assist policymakers and pertinent authorities in proactively recognizing and resolving potential obstacles to the effective execution of the vaccination rollout initiative. The pressing need to achieve herd immunity against the Covid-19 virus calls for prompt measures to effectively address public concerns and cultivate confidence and trust in the vaccination campaign. This study presents an analysis of the machine learning techniques employed and a comparative evaluation of the significance of utilizing the algorithms implemented in this investigation. Our work in the future is to create an application that aims to classify feelings regarding diseases and vaccines and direct opinions about them. Given the ongoing development of evidence, it is imperative to comprehend the potential obstacles associated with the global Covid-19 vaccination initiative. This understanding will aid governments and pertinent institutions in formulating strategies that effectively address the concerns of individuals regarding the administration of Covid-19 vaccines. Although the machine learning classifiers exhibited satisfactory performance on the dataset, it is important to note that our analysis was constrained to a limited corpus of English tweets. The validity of our findings regarding public sentiments on the vaccination program may be compromised by the potential omission of crucial information present in tweets written in languages other than the one analyzed. Subsequent

research endeavors may encompass the assessment of extensive Twitter datasets pertaining to Covid-19 vaccines through the utilization of deep learning models.

REFERENCES

[1] Kedia, A., & Rasu, M. (2020). Hands-on Python natural language processing: explore tools and techniques to analyze and process text with a view to building real-world NLP applications. Packt Publishing Ltd.

[2] Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications, 125(3).

[3] Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. Communication Methods and Measures, 8(3), 190-206.

[4] Anvar Shathik, J., & Krishna Prasad, K. (2020). A literature review on application of sentiment analysis using machine learning techniques. Int J Appl Eng Manag Lett (IJAEML), 4(2), 41-67.

[5] Sanders, A. C., White, R. C., Severson, L. S., Ma, R., McQueen, R., Paulo, H. C. A., ... & Bennett, K. P. (2021). Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. AMIA Summits on Translational Science Proceedings, 2021, 555.

[6] Kottursamy, K. (2021). A review on finding efficient approach to detect customer emotion analysis using deep learning analysis. Journal of Trends in Computer Science and Smart Technology, 3(2), 95-113.

[7] Best, K. (2006). Design management: managing design strategy, process and implementation. AVA publishing.

[8] García-Díaz, J. A., Cánovas-García, M., & Valencia-García, R. (2020). Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America. Future Generation Computer Systems, 112, 641-657.

[9] García-Díaz, J. A., Cánovas-García, M., & Valencia-García, R. (2020). Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America. Future Generation Computer Systems, 112, 641-657.

[10] Gordon, C., Porteous, D., & Unsworth, J. (2021). COVID-19 vaccines and vaccine administration. British Journal of Nursing, 30(6), 344-349.

[11] Ladd, J., Ryan, R., Singh, L., Bode, L., Budak, C., Conrad, F., ... & Traugott, M. (2020). Measurement considerations for quantitative social science research using social media data.

[12] Gasser, U., Ienca, M., Scheibner, J., Sleigh, J., & Vayena, E. (2020). Digital tools against COVID-19: taxonomy, ethical challenges, and navigation aid. The Lancet Digital Health, 2(8), e425-e434.

[13] Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., ... & Van Der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. Royal Society open science, 7(10), 201199.

[14] Schlegelmilch, B. B., Sharma, K., & Garg, S. (2022). Employing machine learning for capturing COVID-19 consumer sentiments from six countries: a methodological illustration. International Marketing Review.

[15] Azkur, A. K., Akdis, M., Azkur, D., Sokolowska, M., van de Veen, W., Brüggen, M. C., ... & Akdis, C. A. (2020). Immune response to SARS-CoV-2 and mechanisms of immunopathological changes in COVID-19. Allergy, 75(7), 1564-1581.

[16] Hilal, A. M., Alfurhood, B. S., Al-Wesabi, F. N., Hamza, M. A., Duhayyim, M. A., & Iskandar, H. G. (2022). Artificial Intelligence Based Sentiment Analysis for Health Crisis Management in Smart Cities. Computers, Materials & Continua, 71(1).

[17] Satu, M. S., Khan, M. I., Mahmud, M., Uddin, S., Summers, M. A., Quinn, J. M., & Moni, M. A. (2021). TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets. Knowledge-Based Systems, 226, 107126

[18] Kumar, K., & Pande, B. P. (2022). Analysis of Public Perceptions Towards the COVID-19 Vaccination Drive: A Case Study of Tweets with Machine Learning Classifiers. In Disease Control Through Social Network Surveillance (pp. 1-30). Cham: Springer International Publishing

[19] Poddar, S., Mondal, M., Misra, J., Ganguly, N., & Ghosh, S. (2022, May). Winds of Change: Impact of COVID-19 on Vaccine-related Opinions of Twitter users. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 16, pp. 782-793).

[20] Paul, N., & Gokhale, S. S. (2020, December). Analysis and classification of vaccine dialogue in the coronavirus era. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 3220-3227). IEEE.

[21] Couvy-Duchesne, B., Faouzi, J., Martin, B., Thibeau–Sutre, E., Wild, A., Ansart, M., ... & Colliot, O. (2020). Ensemble learning of convolutional neural network, support vector machine, and best linear unbiased predictor for brain age prediction: Aramis contribution to the predictive analytics competition 2019 challenge. Frontiers in Psychiatry, 11, 593336.

[22] Aslan, S., Kızıloluk, S., & Sert, E. (2023). TSA-CNN-AOA: Twitter sentiment analysis using CNN optimized via arithmetic optimization algorithm. Neural Computing and Applications, 1-18.

[23] Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2023). COVID-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Systems with Applications, 212, 118715.

[24] To, Q. G., To, K. G., Huynh, V. A. N., Nguyen, N. T., Ngo, D. T., Alley, S., ... & Vandelanotte, C. (2023). Anti-vaccination attitude trends during the COVID-19 pandemic: A machine learning-based analysis of tweets. Digital Health, 9, 20552076231158033.

[25] Nazarizadeh, A., Banirostam, T., & Sayyadpour, M. (2022). Sentiment Analysis of Persian Language: Review of Algorithms, Approaches and Datasets. arXiv preprint arXiv:2212.06041.

[26] Ntwaagae, K. J., Motlogelwa, N. P., Thuma, E., Leburu-Dingalo, T., & Mosweunyane, G. (2022). A Classification Approach to Detect Public Sentiments towards COVID-19 Vaccines.

[27] Xiang, R., Li, J., Wan, M., Gu, J., Lu, Q., Li, W., & Huang, C. R. (2021). Affective awareness in neural sentiment analysis. Knowledge-Based Systems, 226, 107137.

[28] Liu, H., Hao, Y., Zhang, W., Zhang, H., Gao, F., & Tong, J. (2021). Online urban-waterlogging monitoring based on a recurrent neural network for classification of microblogging text. Natural Hazards and Earth System Sciences, 21(4), 1179-1194.

[29] Akpatsa, S. K., Li, X., Lei, H., & Obeng, V. H. K. S. (2022). Evaluating public sentiment of covid-19 vaccine tweets using machine learning techniques. Informatica, 46(1).

[30] Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2023). COVID-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Systems with Applications, 212, 118715.

[31] Albalawi, Y., Buckley, J., & Nikolov, N. S. (2021). Investigating the impact of pre-processing techniques and pre-trained word embeddings in detecting Arabic health information on social media. Journal of big Data, 8(1), 95.

[32] Kirov, C., Sylak-Glassman, J., Que, R., & Yarowsky, D. (2016, May). Very-large scale parsing and normalization of wiktionary morphological paradigms. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 3121-3126).

[33] Angell, R. C., Freund, G. E., & Willett, P. (1983). Automatic spelling correction using a trigram similarity measure. Information Processing & Management, 19(4), 255-261.

[34] Pfeifer, V. A., Armstrong, E. L., & Lai, V. T. (2022). Do all facial emojis communicate emotion? The impact of facial emojis on perceived sender emotion and text processing. Computers in Human Behavior, 126, 107016.

[35] Sharma, S., Srinivas, P. Y. K. L., & Balabantaray, R. C. (2015, August). Text normalization of code mix and sentiment analysis. In 2015 international conference on advances in computing, communications and informatics (ICACCI) (pp. 1468-1473). IEEE.

[36] Hodeghatta, U. R., & Nayak, U. (2023). Introduction to Natural Language Processing. In Practical Business Analytics Using R and Python: Solve Business Problems Using a Data-driven Approach (pp. 541-599). Berkeley, CA: Apress.

[37] Runkler, T. A. (2020). Data analytics. Wiesbaden: Springer Fachmedien Wiesbaden.

[38] Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016, November). Inducing domain-specific sentiment lexicons from unlabeled corpora. In Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing (Vol. 2016, p. 595). NIH Public Access.

[39] Chandrasekaran, G., Nguyen, T. N., & Hemanth D, J. (2021). Multimodal sentimental analysis for social media applications: A comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(5), e1415.

[40] Da San Martino, G., Yu, S., Barrón-Cedeno, A., Petrov, R., & Nakov, P. (2019, November). Fine-grained analysis of propaganda in news article. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 5636-5646).

[41] Kiritchenko, S., & Mohammad, S. M. (2017). The effect of negators, modals, and degree adverbs on sentiment composition. arXiv preprint arXiv:1712.01794.

[42] Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Systems with Applications, 110, 298-310.

[43] Cios, K. J., Pedrycz, W., & Swiniarski, R. W. (2012). Data mining methods for knowledge discovery (Vol. 458). Springer Science & Business Media.

[44] Akerkar, R. (2019). Artificial intelligence for business. Springer.

[45] Bose, R., Aithal, P., & Roy, S. (2020). Sentiment analysis on the basis of tweeter comments of application of drugs by customary language toolkit and textblob opinions of distinct countries. Int. J, 8.

[46] Shehadeh, A., Alshboul, O., Al Mamlook, R. E., & Hamedat, O. (2021). Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Automation in Construction, 129, 103827.

[47] Müller, O., Junglas, I., Brocke, J. V., & Debortoli, S. (2016). Utilizing big data analytics for information systems research: challenges, promises and guidelines. European Journal of Information Systems, 25(4), 289-302.

[48] Nyirandayisabye, R., Li, H., Dong, Q., Hakuzweyezu, T., & Nkinahamira, F. (2022). Automatic pavement damage predictions using various machine learning algorithms: Evaluation and comparison. Results in Engineering, 16, 100657.

[49] Melton, C. A., Olusanya, O. A., Ammar, N., & Shaban-Nejad, A. (2021). Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. Journal of Infection and Public Health, 14(10), 1505-1512.

[50] Hasan, A. K. H. (2022). Characteristics of machine learning algorithms to improve student evaluation (Master's thesis, Altınbaş Üniversitesi/Lisansüstü Eğitim Enstitüsü).

[51] Athanasiou, V., & Maragoudakis, M. (2017). A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: A case study for modern Greek. Algorithms, 10(1), 34.

[52] Millstein, F. (2020). Natural language processing with python: natural language processing using NLTK. Frank Millstein.

[53] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., ... & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8, 1-74.

[54] Sezer, T. (2017). TS corpus project: An online Turkish dictionary and TS DIY corpus. European Journal of Language and Literature, 3(3), 18-24.

[55] Yassine, M., & Hajj, H. (2010, December). A framework for emotion mining from text in online social networks. In 2010 IEEE International Conference on Data Mining Workshops (pp. 1136-1142). IEEE.

[56] Antonakaki, D., Fragopoulou, P., & Ioannidis, S. (2021). A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. Expert Systems with Applications, 164, 114006.

[57] Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. PloS one, 10(12), e0144296.

[58] Khattak, A., Asghar, M. Z., Saeed, A., Hameed, I. A., Hassan, S. A., & Ahmad, S. (2021). A survey on sentiment analysis in Urdu: A resource-poor language. Egyptian Informatics Journal, 22(1), 53-74.

[59] Srikanth, J., Damodaram, A., Teekaraman, Y., Kuppusamy, R., & Thelkar, A. R. (2022). Sentiment Analysis on COVID-19 Twitter Data Streams Using Deep Belief Neural Networks. Computational intelligence and neuroscience, 2022.

[60] Kim, D., Seo, D., Cho, S., & Kang, P. (2019). Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. Information Sciences, 477, 15-29.

[61] Li, T., Mei, T., Kweon, I. S., & Hua, X. S. (2010). Contextual bag-of-words for visual categorization. IEEE Transactions on Circuits and Systems for Video Technology, 21(4), 381-392.