A STUDY OF SOCIAL MEDIA SENTIMENT ANALYSIS USING MACHINE LEARNING ALGORITHM AND NLP APPROACHES

BY

Md. Nahid Sarker

ID: 221-15-5163

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mr. Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

13 JANUARY, 2025

APPROVAL

This Thesis titled **"A study of social media sentiment analysis using machine learning algorithm and nlp approaches"**, which is submitted by Md. Nahid Sarker, Student ID: 221-15-5163 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13 January, 2025.

BOARD OF EXAMINERS

Dr. Md. Zahid Hasan — **Chairman**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mohammad Monirul Islam — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Afjal Hossan Sarower — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Ahmed Wasif Reza — **External Examiner**
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

I am hereby declare that, this project has been done by me under the supervision of **Mr. Abdus Sattar,** Assistant Professor, **Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

17.12.24

Mr. Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Nahid

Md. Nahid Sarker
ID: 221-15-5163
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing making us possible to complete the final year project successfully.

We are really grateful and wish our profound indebtedness to **Abdus Sattar**, **Assistant Professor**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Deep Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori**, **Professor & Head**, Department of CSE, for his kind help to finish our project and also to other faculty members and th e staff of CSE Department of Daffodil International University.

We would like to thank our entire course mates in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

The steady increase of the social networking sites resulted in the generation of large amounts of the user-generated text data that enable the use of sentiment analysis to gain insights into the general public sentiment. This project mainly concerns the construction of a sentiment classifier on text data from social media through the usage of machine learning and natural language processing. The process flow entails a heavy data pre-processing whereby the text is normalized, tokenized, de-stop worded, and lemmatized. For subjectivity and polarity scores, TextBlob is used to sort out informative comments based on their positive, negative or neutral sentiment.

Finally, the feature extraction was done on the text data using Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer to transform the text features into numbers. The multiple machine classifiers under consideration include Naive Bayes, Support Vector Machine (SVM) and Decision Tree. The highest accuracy model is compiled into an API using it for sentiment analysis on new entries made by users.

The presented model shows high quality in terms of sentiment prediction therefore, the future work should concentrate on the integration of conventions NLP with machine learning algorithms. This project provides a solution of managing the huge volume of data collected from the SNS and analyzing the user sentiments for businesses and researchers. Potential future work is as follows One could use deep learning models to implement the process and where there is multilingual data the approach may have to be expanded further.

Keywords: Sentiment Analysis, Social Networking Sites, Machine Learning, Natural Language Processing, TextBlob, TF-IDF,Polarity, Subjectivity.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Introduction

The use of the internet and other forms of technology in communication has rapidly grown as the future communication technology. SNS including Twitter, Facebook, Instagram, Reddit and many more has become the core platform where users express their opinions, experience, comment and even share their immediate feelings and response about various issues including products and services, politics and current events. This has given rise to a large amount of unstructured text data and present a world of opportunities for businesses, organizations and researcher to analyze the sentiment and opinions of the public. Opinion mining is the process of classifying the available data into positive, negative or neutral based on the opinion it holds, Sentiment analysis is sub-domain of Natural Language Processing. Its main aim is to classify the text into either positive, negative or a neutral tone. The need for automated sentiment analysis has been on the rise because, today there is a massive generation of text data on social media platforms daily. The system-dependent traditional manual analysis methods are highly time consuming, comprehensive and not sustainable. Hence, leveraging machine learning and NLP techniques offer a structured and effective strategy in analyzing big data, extracting valuable patterns and making the assumption about the sentiment of the users' comments. The goal of this work is to build an extended sentiment analysis model for data collected from social networking sites based on machine learning approaches and enhanced natural language processing methods. The process includes creating a data pre-processing and situation analysis pipeline and constructing the model and making situational sentiments. Other parts of this process are text preprocessing text segmentation, elimination of low value stop words and stemming to enhance the proper preparation of data for sentiment analysis. In order to determine the most suitable algorithm with the best accuracy in the field of sentiment classification the presented project employs machine learning algorithms including Naive Bayes, Support Vector Machine (SVM) and Decision Tree. This project is meaningful in the sense that it allows to extract operational data from unstructured texts appearing on social media. These

insights can be efficiently used by businesses to establish what customers think, knowing how the public will react and making justified decisions on how products and services can be enhanced. Besides, it can be also used in political insight, market surveys and crisis communications where it is mandatory to have an immediate grasp of the people's attitude. In addition to the model developed herein, this project also seeks to provide a solution in the form of an interface, whereby a user can enter new text data to be analyzed for sentiment analysis and obtain the result immediately.

1.2 Motivation

The reason behind this act of this project is rooted in observing the rising role of social media as an agent of opinion making and consumer choices. Internet forums have therefore become the most popular platform through which people give social opinions and feedback on issues. However, the challenge that comes with the ever-increasing volume of text data is a major issue that constricts business and analysts from getting important information. Simply using historical records to examine this data on a manual basis is unfeasible and incorporates biases and unevenness into the process. Machine learning and NLP approaches which try to automate the sentiment classification are advantageous over the shortcomings mentioned above and can be used for efficient big scale and unbiased sentiment analysis. In addition, this project includes the development of a user interface for real-time sentiment analysis, which endows this project with functionality. The concept of an automated tool that could easily decipher the user sentiment to feed back into key business areas such as marketing, customer relations and brand management in the strategy making process would be highly valuable. This project also seeks to advance the existing literature in SNS text data analysis, and more specifically in elaboration of the sentiment analysis task of the NLP field through testing the efficiency of variety of machine learning models.

1.3 Rationale of the Study

The first and foremost motivation for this study is the availability of large volume of text data in the form of unstructured data in SNSs. Social media text data is problematic and noisy and may contain multiple informal language, abbreviations, slang, emojis and many

other forms of expression. Texts on the social media are diverse and it is not enough to use keywords and simple rule with matching to analyze sentiment. In order to improve the efficiency and effectiveness of a sentiment analysis the present research work uses machine learning operations alongside Natural Language Processing (NLP) methods. Specifically, text preprocessing and feature selection using term frequency-inverse document frequency (TF-IDF) and different classification methods of machine learning algorithms will be used to make a prediction model for classifying different kinds of social media text. Moreover, the integration of TextBlob for subjectivity and polarity scoring gives the first indication of the sentiments which help when categorizing the data.

1.4 Expected Output

What is expected to be produced in this project is the sentiment analysis model that is capable of sorting the sentiments of the comments in the social media as either positive, negative or neutral. The deliverables include A dataset that has been preprocessed in that the text has been cleansed and is therefore in the format to assist the analyst. Sentiment analysis of each comment where the positive and negative scores have been obtained through TextBlob package. Machine learning model that has been optimized properly and then validated on the accuracy factor such as accuracy score, F1 score, recall score or precision score etc. The integration of the sentiment analysis model into the API so as to enable subsequent analysis of new sampled data fed in by users. Word clouds for visualizing the most frequently used words throughout the text as well as sentiment distribution graphs which illustrate the total bias of the data in a graphical form. Collectively, these outputs will confirm the effectiveness of the chosen methods and offer a usable foe-for-sentiment-analysis that will be helpful in a wide range of practical settings.

1.5 Report Layout

The structure of the report is divided into several extensive chapters, where each of them reflects a particular component of the work carried out and its conclusions. The approach used ensures that all the major aspects of the project are incorporated and the reader is

presented with steps from research, analysis and implementation. Below is a detailed breakdown of each chapter included in this report:

Chapter 1: In fact, the first chapter of the methodological nature contains all essential theoretical information on the background of the given project. It concludes the problem formulation and elucidates the roles of sentiment analysis in giving insights over the text generated by the users in SNS. In providing an overview of this chapter, the following sections describe the research objectives, reasons for choosing this subject and the adoption of ML and NLP. This section highlights the expected outputs and contributions of the study which give the reader a preview on what to expect as the subsequent chapters are unveiled.

Chapter 2: This chapter undertakes an extensive discussion of the theoretical background and some prior researches on sentiment analysis and its use. It gives a detailed survey of related work, including traditional approaches of sentiment analysis to the current approaches based on machine learning, and deep learning. The chapter also describes the NLP methods required for text cleaning and feature extraction including tokenization, stop-word elimination, TF-IDF vectorization. Moreover, this section provides information on some of the major problems encountered in SA, such as slang, sarcasm and multi-lingual data and explores the existing solutions and their shortcomings. In the background chapter, other similar studies are discussed to realize the gaps of the present research, which also determines the course of the presented method in this project.

Chapter 3: The literature review chapter presents the systematic approach used to define the sentiment analysis model. It starts with a justification on how the data was collected, the datasets which were used and how data cleaning and pre-processing was done. Some other steps like text normalization, lemmatization and vectorization are explained in detail. The chapter then describes the over selected for this study which comprise Naive Bayes, Support Vector Machine (SVM) and Decision Tree classifiers. To begin, details of the configuration each model as well as the training methodology applied to the models are covered. Accuracy, F1 Measure, Recall and Precision for the evaluation symbols are also described, forming the basis of the evaluation of model. Indeed this chapter presents methods of the study by focusing on the experimental setup and including justification of techniques used in the cross-sectional and longitudinal analyses.

Chapter 4: This chapter also provides the outcomes of the sentiment analysis model and a detailed assessment of this performance. The developed models which are trained on the above modeled dataset are compared based on the various performance indicators or measures. confusion matrix and precision-recall curves are incorporated to illustrate the performance of the model. The chapter also does a comparative analysis of various machine learning models employed in the project with the merits and demerits of the models enumerated. This discussion section explains and extends these findings which assess how well the model works and where it can possibly be improved. The importance of this analysis is for the clarification of the applicability and applicability restrictions of the model.

Chapter 5: This chapter looks at the general impact of the sentiment analysis model on society, the environment and sustainability objectives. Application to society is then highlighted based on areas comprising customer relations, market analysis, political opinions and disease surveillance. Expanding findings of the chapter, the author reflects on what automated sentiment analysis can offer for improving the decision-making process, increasing user interaction with content and giving prompt feedbacks to businesses and policymakers. The effect on the environment is analyzed with regard to the computing and power resources used to train artificial intelligence models as far as the energy-saving measures. The chapter also offers techniques for reducing the impact on the environment which include improving algorithms and use of appropriate hardware.

Chapter 6: Therefore, this chapter gives implications and conclusions from the entire research exercise while focusing on the research objectives and how the study attended to them. The latter unveils important features unfolded in various stages of the study including data preparation and modeling as well as the results of evaluation and analysis. The overview provides a summary of the research questions, methods to be used and the expected results and forms a good link to link all the separate research questions and methods together. This also presents a format in which the problems experienced during the study process are examine and solutions adopted with the readers being conversant with the entire process.

Chapter 7: The conclusion chapter provides a brief overview of all of the major results of the project, the relevance of the findings, and the usage possibilities. First, it provides a review of the related work concerning sentiment analysis and the proposed model to show its usefulness to different areas such as marketing, customer support and public opinion analysis. Current approach limitations are recognized, problems connected with noisy text data and recognizing complex linguistic phenomena such as sarcastic and context-sensitive sentiment. It also discusses potential research avenues that can be pursued to extend the presented topic including the use of deep learning approaches and the study of ensembles and using the present approach for multilingual datasets. In the final section of the chapter, the authors reiterate on the applicability of the refined sentiment analysis tool under real-time context and identifying business intelligence obtained from social media.

CHAPTER 2

Background Study

2.1 Terminologies

Before diving into the related works, it is important to define some key terminologies used in this study. The ability to identify positive, negative or neutral sentiment, emotion or opinion from text data set. A branch of AI which deals with the communication between a human and a computer and subsequently the ability of the latter to process texts. Polarity means the evaluation of text as to the presence of positive, negative or neutral opinion. A measure of how much of the given text is written from the author's point of view rather than the facts. Tokenization, stemming or lemmatization that are used to clean up the textual data that is to be analyzed. Such algorithms include Naive Bayes, Support Vector Machine- SVM and Decision Tree which are used to classify data according to the patterns learnt on the data. A feature extraction method that measures the relative relevance of a word to a document with reference to a set of documents. A graphical user interface created to enable users of the developed sentiment analysis model to obtain real time predictions.

2.2 Related Work

The field of sentiment analysis has seen substantial research interest over the years, with numerous approaches and methodologies proposed. Here, we review 30 significant studies that have contributed to the development of sentiment analysis models.

Pang et al. (2002) introduced machine learning techniques for sentiment classification, using algorithms like Naive Bayes and SVM on movie reviews, marking an early attempt in automated sentiment analysis [1] Turney (2002) presented an unsupervised approach using phrase-level sentiment classification based on pointwise mutual information (PMI), demonstrating its applicability for product reviews [2] Paltoglou and Thelwall (2010) explored feature selection methods for sentiment analysis in online social media, showing the importance of proper feature extraction for improved model performance [3] Go et al. (2009) applied distant supervision using emoticons as labels for training sentiment classifiers on Twitter data, pioneering a method for handling vast amounts of unlabeled social media text [4] Agarwal et al. (2011) compared rule-based and machine learning-

based sentiment classifiers on Twitter data, highlighting the limitations of traditional rule-based methods [5] Kiritchenko et al. (2014) developed sentiment lexicons for social media text, enhancing the sentiment prediction accuracy for tweets with specific terms [6] Zhang et al. (2018) utilized deep learning models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for sentiment analysis, achieving significant accuracy improvements over traditional models [7] Jianqiang and Xiaolin (2017) proposed an enhanced LSTM model for sentiment classification on noisy social media text, addressing issues like slang and misspellings [8] Dos Santos and Gatti (2014) implemented deep CNNs for sentence-level sentiment analysis, leveraging word embeddings for improved feature representation [9] Socher et al. (2013) introduced recursive neural networks (RNN) for sentiment analysis, demonstrating their effectiveness in capturing contextual information from text [10] Wang and Manning (2012) conducted a comparative study of various machine learning algorithms for sentiment analysis, highlighting the efficiency of SVM classifiers in text classification tasks [11] Yang et al. (2019) focused on ensemble learning approaches, combining multiple models to enhance the accuracy of sentiment analysis in social media data [12] Zainuddin et al. (2016) utilized hybrid feature extraction techniques, integrating TF-IDF with word embeddings for improved sentiment analysis performance [13] Hutto and Gilbert (2014) developed VADER, a lexicon and rule-based model optimized for social media text sentiment analysis, known for its simplicity and effectiveness [14] Liu (2012) provided a comprehensive review of opinion mining and sentiment analysis, outlining key challenges and future directions in the field [15] Ghiassi et al. (2013) proposed a dynamic neural network model for sentiment analysis, showing its potential in handling complex text data [16] Medhat et al. (2014) conducted a detailed survey of sentiment analysis techniques, comparing various machine learning, lexicon-based, and hybrid approaches [17] Jin et al. (2016) explored the use of deep learning for aspect-based sentiment analysis, improving sentiment predictions by focusing on specific text aspects [18] Kouloumpis et al. (2011) evaluated the impact of linguistic features on Twitter sentiment analysis, emphasizing the importance of feature engineering [19] Pak and Paroubek (2010) applied statistical machine learning techniques to sentiment analysis of Twitter data, establishing a baseline

for social media sentiment research [20] Gupta et al. (2020) introduced a multi-domain sentiment analysis model using deep transfer learning, enhancing cross-domain sentiment prediction [21] Devlin et al. (2019) proposed BERT (Bidirectional Encoder Representations from Transformers), significantly advancing NLP tasks, including sentiment analysis, through pre-trained contextual embeddings [22] Chen et al. (2020) explored fine-tuning transformer models for sentiment analysis, achieving state-of-the-art results on benchmark datasets [23] Schouten and Frasincar (2016) provided a comprehensive review of aspect-based sentiment analysis, focusing on techniques for handling opinionated text [24] Poria et al. (2016) integrated multimodal data (text, audio, and video) for sentiment analysis, expanding the scope beyond text-based approaches [25] Zhou et al. (2017) applied attention mechanisms in sentiment analysis models, improving the interpretation of sentiment by focusing on relevant text parts [26] Severyn and Moschitti (2015) proposed a deep learning approach using CNNs for Twitter sentiment analysis, showing superior performance on noisy social media data [27] Cambria et al. (2017) discussed sentiment analysis advancements using deep learning and NLP, providing insights into current trends and future possibilities [28] Xu et al. (2020) implemented transfer learning techniques with BERT for fine-tuning sentiment classifiers on domain-specific data, achieving high accuracy [29] Chatterjee and Sengupta (2019) proposed a hybrid model combining lexicon-based and machine learning methods for sentiment analysis, addressing the limitations of individual approaches [30]

2.3 Comparative Analysis and Summary

This paper will focus on the evolution of the sentiment analysis within the last two decades, potential stimuli that have contributed to its change. The key techniques can be broadly classified into three categories: These are the rule-based approach, the learning-based approach and the deep endorsement learning based approach. This comparative analysis also showing the strength, weakness and progress recorded in the approaches used.

2.3.1 Rule-Based Approaches: The first attempts at using methods to perform sentiment analysis have been based on rules and employed manually created dictionary and sentiment

lexicon. Like Turney (2002), Hutto & Gilbert (2014) and the VADER approach, these approaches rely on manually defined rules based on words that are defined as positive or negative. This kind of models is quite easy and efficient to use in cases when it is necessary to analyze short texts which don't contain rather ambiguous lexical units, irony, or shifts of tones depending on context. Rule-based systems are also time-consuming in terms of using, when compared with the large amount of diverse and dynamic social media content, the lexicons need to be updated frequently, which takes a lot of manual efforts.

2.3.2 Machine Learning-Based Models: Twitter sentiment analysis for example earlier used vocab-based approach but with the help of M L was improved through methods like Naive Bayes SVM and DT classifiers. Such models as described by Pang et al. (2002) and Pak & Paroubek (2010) rely on n-grams, TF-IDF scores and the bag-of-words approach in case of social media sentiment prediction. Tying with rule-based systems is something that is not usually achieved by machine learning models as these are usually superior especially with large labeled datasets. Huge importance to feature extraction and furthermore, it presupposes extensive preprocessing of text data where language dependent heuristics could be effectively applied.

2.3.3 Deep Learning-Based Model: This type of sentiment analysis has over the years been made easier by the existence of deep learning models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and transformer-based architectures like BERT. These models like the ones illustrated by Socher et al. (2013), Devlin et al., (2019) and Xu et al., (2020) outcompete in the fact that they are capable of automatically learning semantic representations leading to better understanding of context in the text data. Latest approaches like deep learning models have been found to be very effective, especially with large data sets and especially when the program requires to learn in context. Nevertheless, they are computational costly and need large training sets with labeling.

2.3.4 Comparative Summary: Indeed, this comparative analysis shows that although rule-based approach is easy and interpretable it cannot adequately capture the complexity of the twitter text. A somewhat more favorable compromise in terms of accuracy vs. computational complexity is achieved with the help of more advanced machine learning

techniques, at the same time, these models are very sensitive to the choice of features. Current advances in deep learning technique, have shown noteworthy improvements in sentiment analysis, particularly on the data obtained from social networking sites with high noise levels. However, it is still difficult to meet demand due to the issue of computational cost and the requirement of big data.

The comparative analysis carried out and used to stress that the integration of traditional machine learning to deep learning and more comprehensive NLP preprocessing can provide the best results of sentiment analysis for the data from social networks. This project builds upon this insight to propose a holistic architecture implementing machine learning, NLP and API delivery for real-time sentiment analysis.

2.4 Scope of the Problem

The area of interest for this study concerns the sentiment analysis of text data extracted from SNS, with a particular regard to tweets, Facebook statuses and comments and Reddit submissions. Social media are particularly complex insofar as language is less formal there are frequently used acronyms, symbols and certain words, as well as their context may differ from standard language. A lot of the data present on social media consists of such noise in form of wrong spelling, abbreviations and unsophisticated grammar. Pre-processing methods which need to be applied are tokenization, removal of stop words, stemming and lemmatization. The objective is to predict the sentiment of the text data into three types, positive, negative and neutral. This means creating techniques that are capable to ascertain the tone of the message as well as whether the language that has been used is simple or complicated let alone sarcasm. As more and more organizations turn to social media as a means of communication it has become necessary to develop systems that can handle data in real time. It is easier to create an API that will be delivering sentiment predictions within a short time, it is essential for businesses and organizations. Text which is produced on social media platforms is context-bound and depends on the type of conversation (product reviews, political debate, personal opinions, etc.). The model has to be developed for multiple domains and thus should be built using effective methods of

training and validation. On these aspects the present study shall seek to fill the existing research and design a solution to sentiment analysis on social media data.

2.5 Challenges

The process of developing an accurate and reliable sentiment analysis model for social networking sites involves several significant challenges:

- **Noisy and Unstructured Data:** In general, social media text contains popular shortcuts, many symbols, and informal language that can hardly be addressed by traditional NLP methods. Mechanical perfections as well as deliberate mistakes hurt the assessment and skew the sentiment readouts.
- **Contextual Understanding:** Sadly, most sentiment analysis models fail to take context into account and hence they may radically change meaning when sarcasm, irony or humor is used. Finding context cannot be done using simple techniques such as bag-of-words or simple techniques like using key-word matching or extracting named entities and parsing tree dependencies as in deep learning models with attention mechanisms or context-aware embeddings like BERT.
- **Handling Multilingual Data:** Different groups of people use different social media, and the posts may contain different languages; the texts may be even in code-switching style. A major area that is still proving difficult when it comes to the creation of models to handle multilingual data. Traditional methods may not work well because not all languages may be supported by existing sentiment lexicons.
- **Class Imbalance:** In real-world datasets, this problem is made manifest through an overrepresentation of certain sentiment classes, say neutral comments, yet an underrepresentation of others such as negative or positive. This problem necessitates the use of such approaches as oversampling under sampling or use of weighted loss functions.
- **Scalability and Computational Efficiency:** A lot of textual content is produced trough Some, hence the need for models that can handle large amounts of them without compromising on performance. The training of deep learning models is often

computationally intensive thus requiring optimization methods and conducting the computations using suitable hardware.

- **Interpretability of Models:** Although deep learning models offer good accuracy, they are highly non-transparent. It is crucial to know why the model has a particular prediction as it is with applications in areas such as geophysical and geopolitical, health and finance.
- **Data Privacy and Ethical Considerations:** The analysis of users' created content opens issues about personal data protection and ethical use of individuals' data. The model has to be adherent to data protection regulations and in general the privacy of the users must be respected.

Solving these difficultly, the study wants to achieve the goal of proposing an improved sentiment analysis model that would be able to process social media text and deliver worthwhile insights concurrently.

CHAPTER 3

Research Methodology

3.1 Introduction

In this chapter, the reader is given a detailed account of the choice of method adopted for analyzing the sentiment of social site comments. The aim of the presented research is to understand if the text of the reviews is Positive or Negative using NLP and ML algorithms. Every form of methodological development is constructed to improve accuracy and reliability. Information collection, data washing, textual standardization, characteristic extraction, model choice and estimation, and assessment of effectiveness are some of them. Thus, the study's objective is to make the analysis of customer sentiment data credible and substantial by following a clear plan.

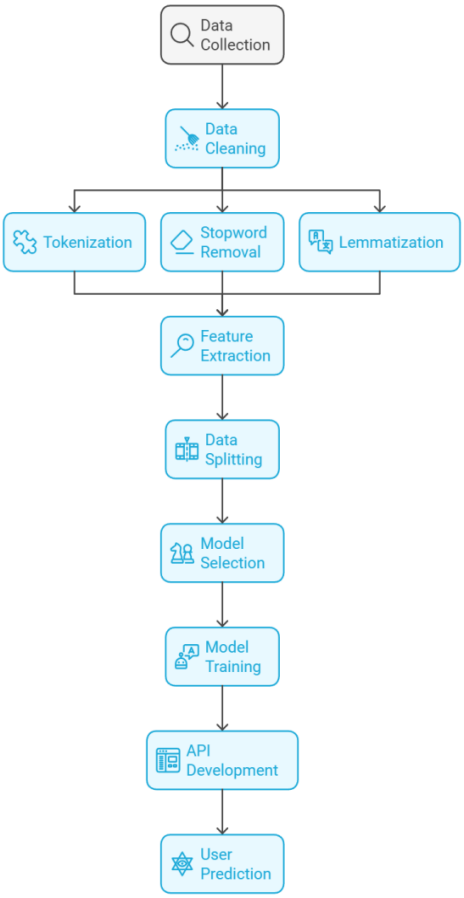

Figure 3.1.1: Proposed System Architecture

3.2 Data Collection Procedure

The information employed in this investigation comprises user feedbacks which were obtained from different sources in social sites. These reviews tend to act as consumer sentiment analysis, consumer grievances and satisfaction level with regard to the product or service. The acquisition of proper and concrete data is a fundamental aspect of sentiment analysis since it leads straight to the model.

- **Data Source:** The dataset is in Excel format that is mainstream and can be easily retrieved as well as manipulated in Python. The reviews may be gathered from sources that customers give their feedback on e.g. the internet, via a shopping site, social sites, customer complain platforms.
- **Loading the Data:** As the first step in data preparation, let's load a dataset into Python, specifically, into a DataFrame constructed within the pandas library. Rows in the DataFrame are individual customers' reviews, while the columns are the actual text of the review and other features like the sentiment of the review that would be positive or negative.

	Comment	Analysis
0	Let's not forget that Apple Pay in 2014 requir...	Positive
1	Here in NZ 50% of retailers don't even have co...	Positive
2	I will forever acknowledge this channel with t...	Neutral
3	Whenever I go to a place that doesn't take App...	Negative
4	Apple Pay is so convenient, secure, and easy t...	Positive
...	...	...
13535	Any basis for Hasina's claim of USA trying to ...	Neutral
13536	I just found it interesting that this time it ...	Positive
13537	It's in chaos now. Recently these students are...	Negative
13538	Another thing to mention - Bangladeshi public ...	Positive
13539	As a Bangladeshi I am telling you our people a...	Neutral

13540 rows × 2 columns

Figure 3.2.1: Sample Dataset

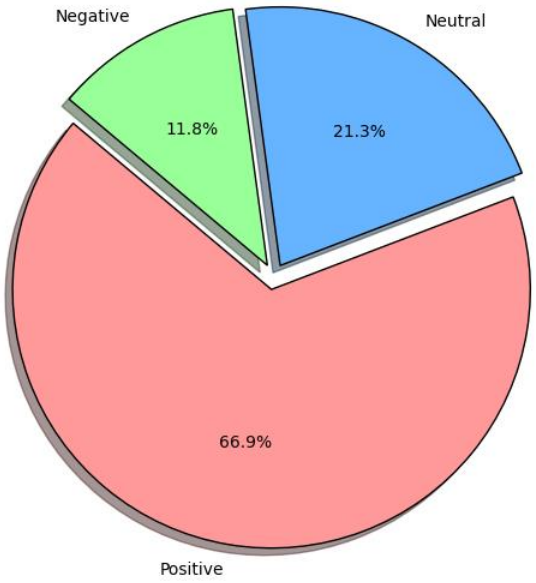

Figure 3.2.2: Sentiment Analysis of Dataset

Figure 3.2.3: Most Frequent Words in Comments

3.3 Dataset Cleaning

The third and important process of text data preparation is data cleaning that mainly addresses inconsistent, noisy or irrelevant values. Reviews provided contain certain

irrelevant features, namely, punctuation signs, special characters and the absence of formatting. These can create noise and immune model accuracy if not dealt with.

3.3.1 Data Cleaning: Format the text data by removing any extra unwanted characters, ignore case and make all the text lower case. Trim spaces and additional delimiters, omit numerations and single and multiple punctuation marks. This process removes irrelevant words from the text data and in the process makes the sentiment analysis step more refined and accurate as it targets the informative words.

3.3.2 Tokenization: Transform big sentences into single words also called tokenization. Tokenization allows to treat every word as an individual contributor and this is very critical when performing pattern extraction and identifying a particular sentiment.

3.3.3 Stopwords Removal: Stoppers – which are words that bring little or no meaning to the text, for example, words such as “the,” “is”– are taken out. This step helps to decrease dimensionality and to enhance model’s concentration more on words that provide more relevance for sentiment.

3.3.4 Lemmatization: Each letter or symbol in the word is reduced to the base keyword which can cluster such words as ‘running’ and ‘run.’ This step assist models to generalize better and therefore bring variations of the same concepts into the same group.

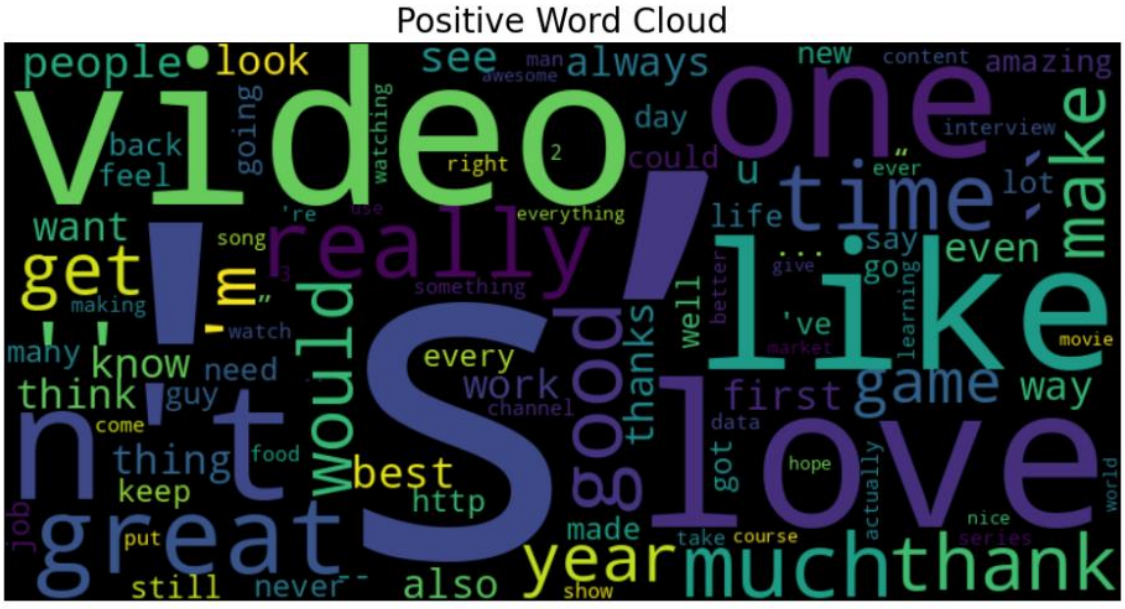

Figure 3.3.1: Positive word cloud

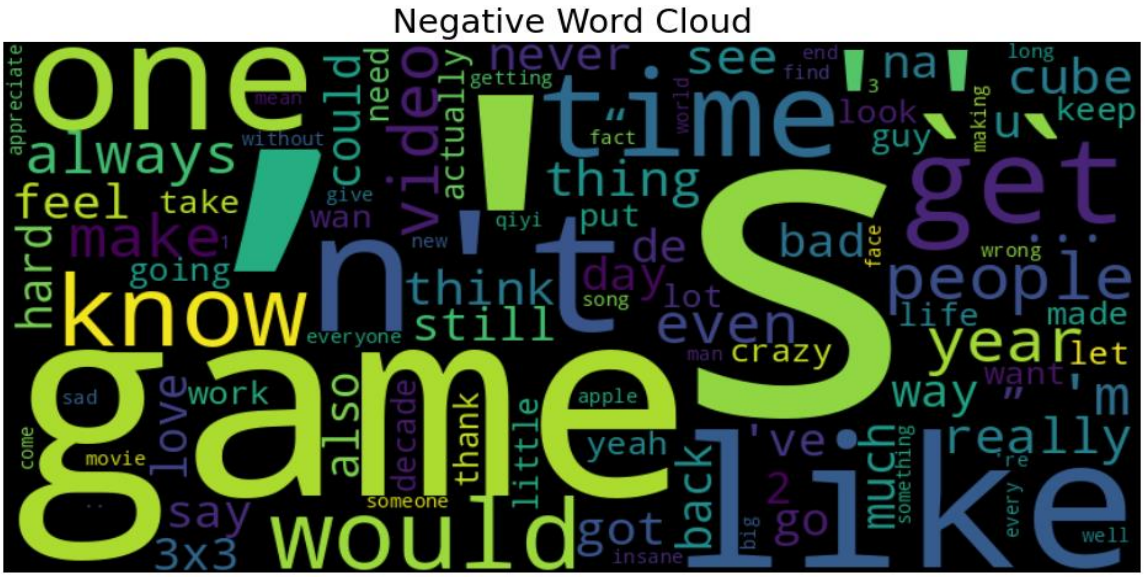

Figure 3.3.2: Negative word cloud

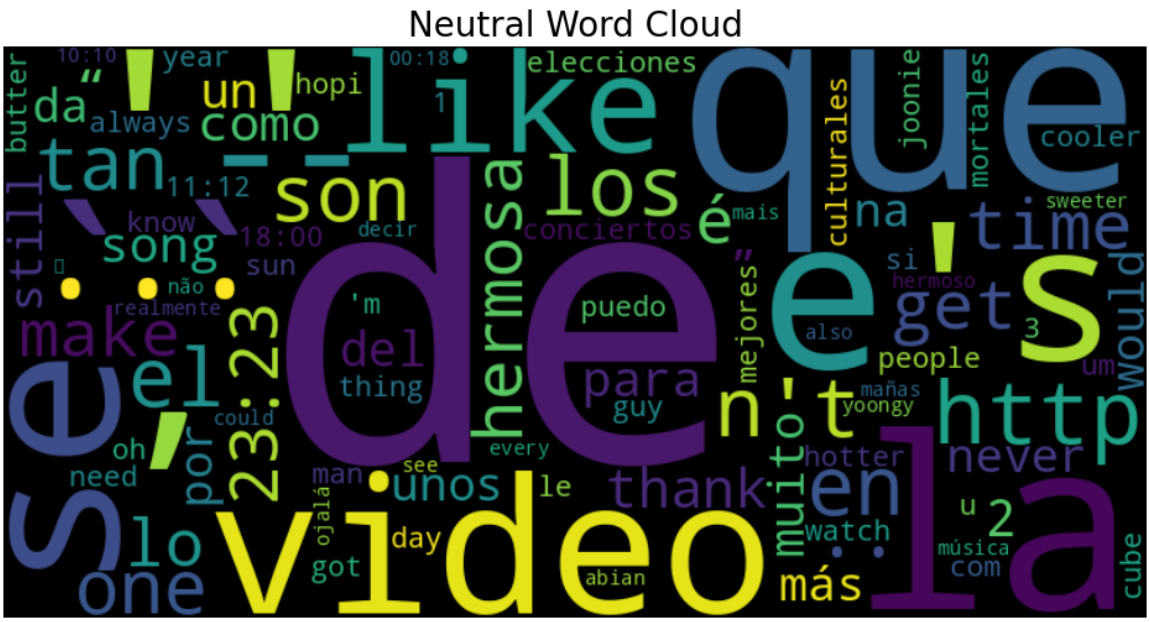

Figure 3.3.3: Neutral word cloud

3.4 Dataset Preprocessing

The chosen technique helps to turn the text data into a format measurable by machine learning algorithms.

3.4.1 Feature Extraction using TF-IDF: It is the best to employ TF-IDF in order to construct a matrix which would estimate the significance of each word in each review. This technique gives each word a rank by calculating how often it was used in the individual review in comparison with the use of the same word in all the reviews. Higher importance is assigned to the words that are often used within a particular review but are less often used in other reviews.TF-IDF focuses on individual terms expanding the impact of low value glossary words by frequently occurring stop words. It helps the model to pay more attention to the words which are more relevant to sentiment of particular review and the data 'richer'.

3.4.2 Data Splitting: Part the files into training and test in order to assess the efficacy of the model. The dataset is divided into two parts: a training sample for model learning and a test sample in order to judge the model performance toward unseen data. It is usually common practice to divide data 80 — 20, where 80% is used for training and 20% for testing. Dividing a data set in to two is logical since it provides an honest assessment of the ability of a given model to predict the results of a particular phenomenon. In this manner, the model is trained on one segment and tested on another so it is easy to determine how the model is at predicting inputs on new data which it has not seen before.

3.5 Proposed Methodology

The general approach of the paper is outlined by the following stylized steps that make up the proposed methodology for text data sentiment analysis. The approach includes:

3.5.1 Sentiment Scoring: Sentiment polarity and subjectivity is determined using a site similar to TextBlob for every review during the data review process. Sentiment polarity and subjectivity scores are computed on a scale of -1 for negative sentiment to + 1 for positive, in sentiment orientation; and on a continuum between zero, for the absence of subjective opinions, to one for the presence of subjective opinions.

3.5.2 Feature Extraction using TF-IDF: The reviews are then transformed into numerical features by using TF-IDF (Term Frequency-Inverse Document Frequency) whereby the importance of each word used within the document is calculated in proportion to the frequency within the document but normalized for Privacy features against the frequency

within the entire data set. ‘Stemming’ is also applied here by which frequently used words in a particular review and those used rarely in other reviews are given more importance in improving the model’s current focus towards the sentiment indicating words.

3.5.3 Data Splitting: The dataset is split into training (80%) and testing (20%) sets so that the model might be tested on the ability of generalization. Such split guarantees that that efficiency of model is determined by a part of information that the model has not met during the training, which offers a correct idea about the model’s efficiency in discipline.

3.5.4 Model Selection and Training: KNN, Logistic Regression, Random Forest, Decision Tree and XGBoost test are conducted to evaluate several machine learning algorithms. Such models are developed in a way that they try to identify word-to-sentiment label relationships; hence they are capable of predicting new reviews accurately from the identified patterns.

3.5.5 Model Evaluation: This is because there are several parameters that One can assess the efficiency of each model in Sentiment analysis Performance metrics including accuracy, precision, recall, and F1 score are used to determine each model’s competence in classifying sentiment. Such measurements enable evaluation in a comprehensive manner; it is understood which model is suitable for the given dataset and which one gives the most faithful sentiments.

3.6 Model Training

Evaluating is a cyclic process of educating the model on sentiment through an assortment of data. The primary steps include:

3.6.1 Training Data Preparation: Having got the necessary data cleaned and preprocessed, and the features extracted, the processed data is inserted into the machine learning algorithms. The feature set is the TF-IDF feature matrix and the target variable is the sentiment the text is classified as either positive or negative.

3.6.2 Model Training: Any of the machine learning model (KNN, Logistic Regression, Random Forest, Decision Tree and XGBoost) is applied on the above prepared data set wherein the algorithm tries to learn the patterns into the data.

- **KNN:** In the current study, the K-Nearest Neighbors (KNN) classifier, which is an instance-based learning algorithm has also been employed to classify reviews based on nearness to certain other reviews with already known classifications. The method of operation of KNN is based on the selection of the ‘k’ nearest data points or neighbors in the feature space and the new review is assigned the most frequently occurring label among them. KNN will compare each review with the others to determine its sentiment. As opposed to actually implementing a training phase, KNN uses the whole training sample during each prediction; while this may be slower than other methods because overall training data is employed in each prediction in KNN, the process works well and efficiently in small databases. The value of 'k' is critical. When it comes to sentiment analysis, if the value of 'k’ is a small number, then it results in over fit while if the value of ‘k’ is a large number then it results in under fit.

- **Logistic Regression:** Logistic Regression is the assigned linear classifier which classifies the review as belonging the positive or the negative sentiment class. It estimates the probability of the input into the model to belong to a given class or not and works well with binary classification thereby suitable for sentiment analysis. With Logistic Regression, sentiment scores are given a probabilistic approach meaning that, in addition to deciding whether the review is a positive or negative, the model is capable of stating a certain level of certainty regarding its decision. This may assist in sort the reviews by how positive or negative the sentiment is which might be useful in measuring the level of customer satisfaction. C the regularization parameter that helps to determine model’s capacity and avoid overtraining.

- **Random Forest:** Random forest can be described as an averaged decision tree learning algorithm which builds tree in parallel to create a number of trees when learning and integrates their result to constitute the solution. This helps to avoid overfitting, usually attributed to single decision trees, through using the results of a number of weak learners. When it comes to text features generated by TF-IDF Random Forest Si proved to be efficient in tackling with complexity generated By sentiment analysis. It gives a stable and frequently a very accurate classification

outcome through averaging prediction of numerous trees. We have seen that this model is quite stable during the identification of patterns, and that is a good property because customer reviews often contain a lot of noise. Parameters of interest are the numbers of trees (n_estimators), depth of trees (max_depth), and minimum number of samples for splitting (min_samples_split). The three above elements can be fine-tuned in order to fine tune model performance and reduce the computational cost.

- **Decision Tree:** Decision Tree classifier is a very basic yet strong model of data classification chances dependent on feature-based questions. It split the dataset into branches based on choosing the best features, and arrives at a decision or a decision leaf corresponding to a particular sentiment classification. This Decision Tree model can take into consideration nonlinear interactions in the text features and explain the specific features that work together to support a sentiment class. And it is fast, and provides meaningful explanation by showing which specific words, or feature, affect the sentiment prediction. These are max_depth, min_samples_split and criterion such as Gini or entropy. These parameters define the depth of the tree and the fractal branching system that ultimately affect the model accuracy and its stability.

- **XGBoost:** XGBoost stands for extreme gradient boosting which is an advanced ensemble learning algorithm that operates with boosting that is a learning method where every subsequent tree corrects mistakes of previous ones. It is useful to work with large datasets; one of the most famous dependency techniques, it attains a high accuracy due to the constant reinforcement of low-quality models. XGBoost is considered to be very useful in sentiment analysis primarily because of the flexibility it offers in addressing textual data features. XGBoost also has better processing of overfitting than traditional tree-based methods, thanks to its incorporated regularization and is suitable for high dimensional data of text. XGBoost has the following tuning hyperparameters; Learning rate, Number of estimators, Maximum depth, Minimum instances per child, Subsample. The correct setting of such values can guarantee both the high accuracy of the model and its efficiency, but for this, more computing resources are needed.

- **Hyperparameter Tuning:** The parameters of the model are tuned in order to enhance the model's results by using different variations of hyperparameters to obtain higher accuracy, and precision.

3.7 Implementation Requirements

- Different Machine Learning Frameworks and Libraries
- Windows 11
- Google Colab with runtime TPU
- Excel file containing social sites comments representing varied sentiment Google Drive

CHAPTER 4

Result Analysis and Discussion

4.1 Introduction

In this chapter, primary experimental results of using five machine learning algorithms (KNN, Logistic Regression, Random Forest, Decision Tree, and XGBoost) for classifying Bangladeshi customer review sentiment are explained in detail. The evaluation of each of the models include the metrics of accuracy, precision, recall and F1 measure. We also present the result using confusion matrix and classification report in order to determine strength and weakness of the models. Lastly, training and validation accuracy and loss trends are discussed as methods to assess the model's behavior during the training process. This assessment establishes which model is suitable for sentiment classification in this set.

4.2 Experiment Results and Analysis

The first outcome measured in result analysis includes assessing the performance of each classifier in the test dataset. In detail 4 quality measures like accuracy, precision, recall and F1-score are computed for each approach. Here's a breakdown of these metrics:

- Accuracy quantifies the performance quality of the model through the estimation of the percentage of the reviews that was well classified, given the total reviews.
- Accuracy ratios show the precision of positive discipline making it easier to avoid over classification.
- Recall determines the ability of the model to accurately capture all positive sentiment reviews inclusive of zero misclassifications.
- While F1-Score is a great measure of the model's accuracy, it is better than Accuracy because it reflects both precision and recall rates, giving a more comprehensive vision of how well a model is doing.

Comparing each classifier by such metrics. The following table summarizes the results:

Table 4.2.1: The Experiment Result of the Evaluated Model

Learning Model	Test Accuracy	Recall	F1-Score	Precision
Random Forest	95%	0.97	0.97	0.97
Logistic Regression	92%	0.97	0.95	0.94
XGBoost	89%	0.89	0.91	0.93
Decision Tree	88%	0.88	0.88	0.88
KNN	66%	0.94	0.96	0.99

4.3 Generating Confusion Matrix

The confusion matrix gives a quantitative view of the different models by separating the results of the classification between true positives, true negative, false positives and false negatives. This matrix actually provides a very important insight towards the identification of how well each of the models is able to distinguish between the positive and negative sentiments.

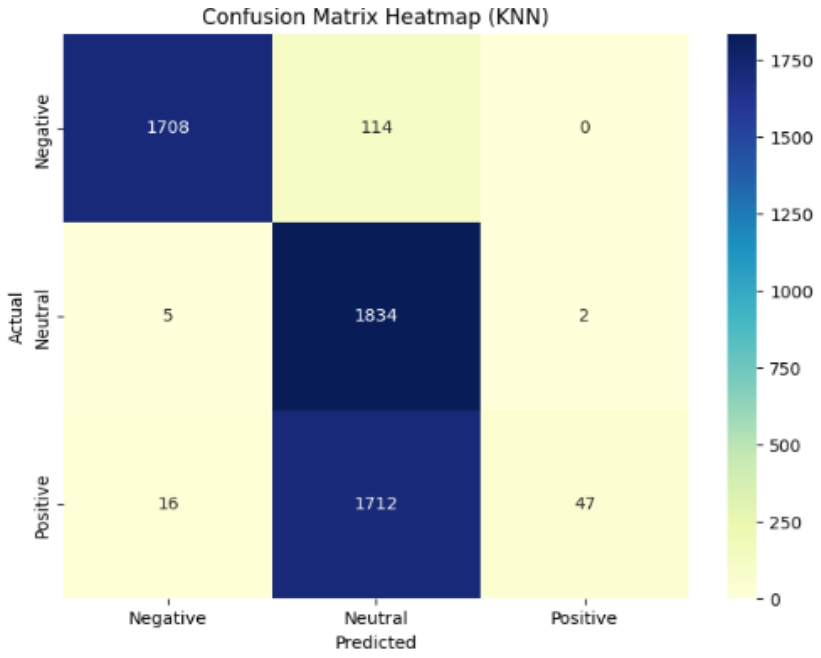

Figure 4.3.1: Confusion matrix heatmap of KNN

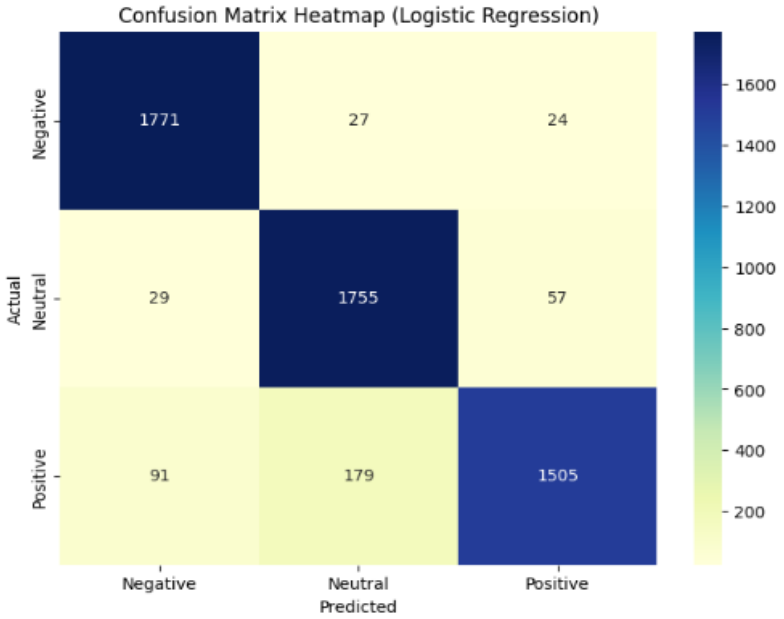

Figure 4.3.2: Confusion matrix heatmap of Logistic regression

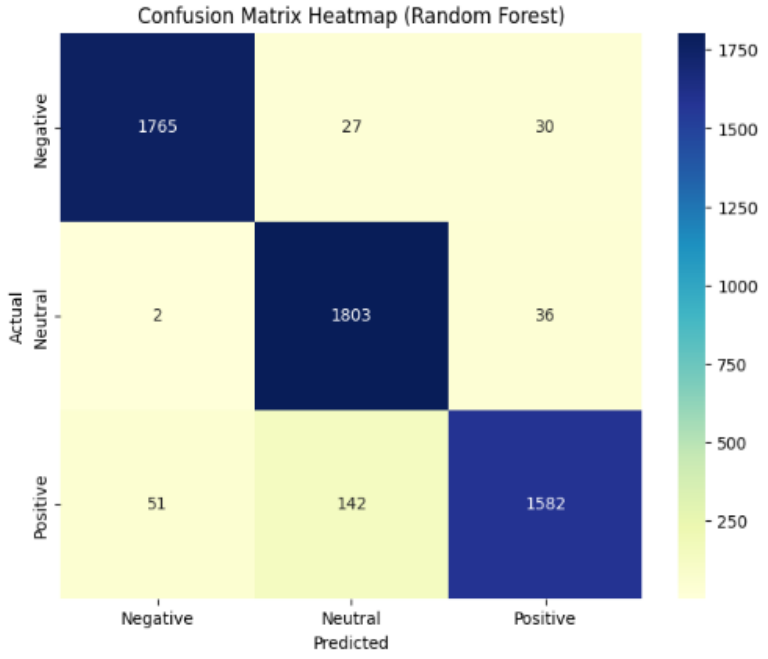

Figure 4.3.3: Confusion matrix heatmap of Random forest

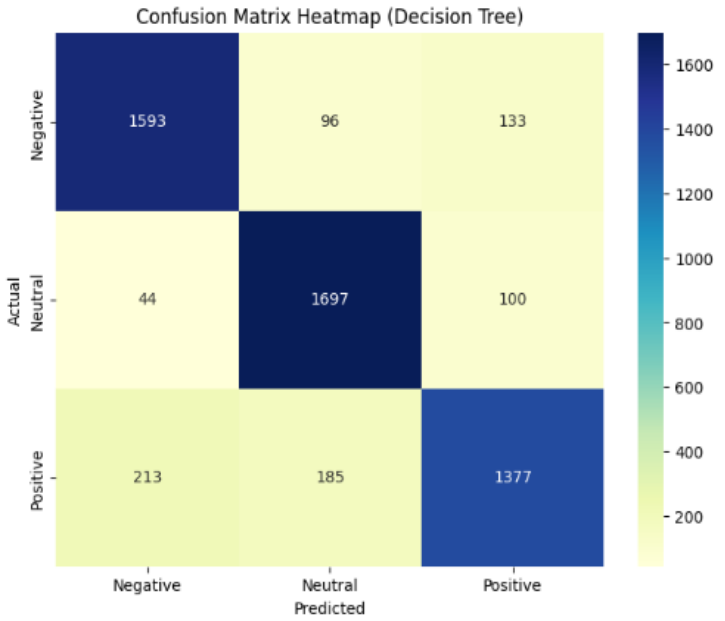

Figure 4.3.4: Confusion matrix heatmap of Decision tree

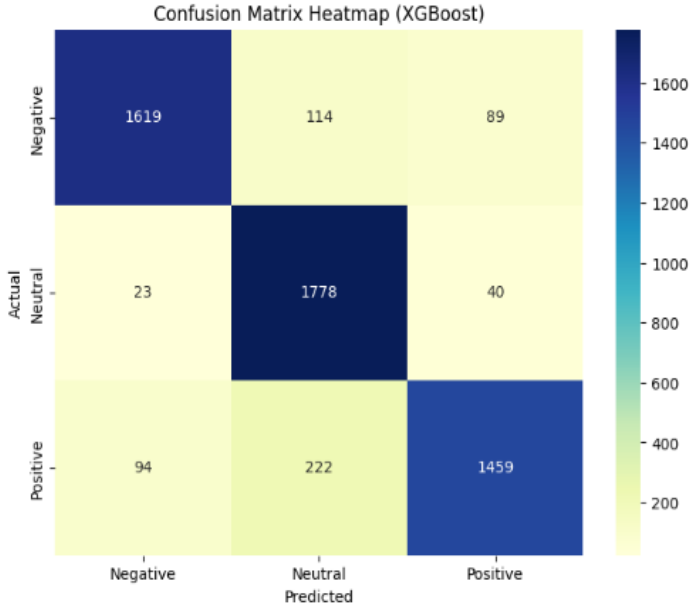

Figure 4.3.5: Confusion matrix heatmap of XGBoost

It explains each classifier's merit and demerit and show if a model has bias toward positive or negative sentiment and which sentiment class was rather complex for a model to determine. This helps in choosing models that have an accurate ability of classifying instance into two classes.

4.4 Generating Classification Report

Finally, the classification report gives a precise measurement of each class positively and negatively by segmenting the results into the following; It provides a detailed on the accuracy of each model on each sentiment type which enable one to detect cases of class imbalance and biased models.

TABLE 4.4.1: Classification Report for KNN

Algorithm	Class	Precision	Recall	F1-Score	Accuracy
KNN	negative	0.99	0.94	0.96	0.66%
	neutral	0.50	1.00	0.67	
	positive	0.96	0.03	0.05	
	Macro avg	0.82	0.65	0.56	
	Weight avg	0.81	0.66	0.56	

TABLE 4.4.2: Classification Report for Logistic Regression

TABLE 4.4.3: Classification Report for Random Forest

Algorithm				ɔre	Accuracy
Logistic Regression	negative	0.94	0.97	0.95	0.93%
	neutral	0.90	0.95	0.92	
	positive	0.95	0.85	0.90	
	Macro avg	0.93	0.92	0.92	
	Weight avg	0.93	0.93	0.92	

Algorithm	Class	Precision	Recall	F1-Score	Accuracy
Random Forest	negative	0.97	0.97	0.97	95%
	neutral	0.92	0.98	0.95	
	positive	0.97	0.89	0.95	
	Macro avg	0.95	0.95	0.95	
	Weight avg	0.95	0.95	0.95	

TABLE 4.4.4: Classification Report for Decision Tree

Algorithm	Class	Precision	Recall	F1-Score	Accuracy
Decision Tree	negative	0.88	0.88	0.88	0.88%
	neutral	0.89	0.93	0.91	
	positive	0.87	0.82	0.84	
	Macro avg	0.88	0.88	0.88	
	Weight avg	0.88	0.88	0.88	

TABLE 4.4.5: Classification Report for XGBoost

Algorithm	Class	Precision	Recall	F1-Score	Support
XGBoost	negative	0.93	0.89	0.91	1822
	neutral	0.84	0.97	0.90	1841
	positive	0.92	0.82	0.87	1775
	Macro avg	0.90	0.89	0.89	5438
	Weight avg	0.90	0.89	0.89	5438

As we analyze each classification report, it is possible to determine whether a given model performs better dealing with one or the other type of sentiments. It is desirable for a classifier to have similar precision and/or comparable recall for all classes which would stand for the fact that the sentiment classifier performs well.

4.4.1 Performance/ Comparative Analysis

The bar chart depicted above shows a model accuracy for different machine learning algorithms that can be used in the text data of social networking sites' sentiment analysis. The chart visualizes the performance of five different models: The chosen models included are K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, Decision Tree Classifier and XGBoost (XGB). The horizontal axis is the models used and the vertical axis is the accuracy score varying from 0 to 0. The best AUC score obtained is from the Random Forest model with 0.95 and the second-best one belongs to Logistic Regression model with a score of 0.92. The two algorithms, the Decision Tree Classifier and XGBoost, returned accuracies of 0.88 and 0.89, respectively. Nevertheless, the KNN model was the worst-performing model under consideration, achieving only 0.66.

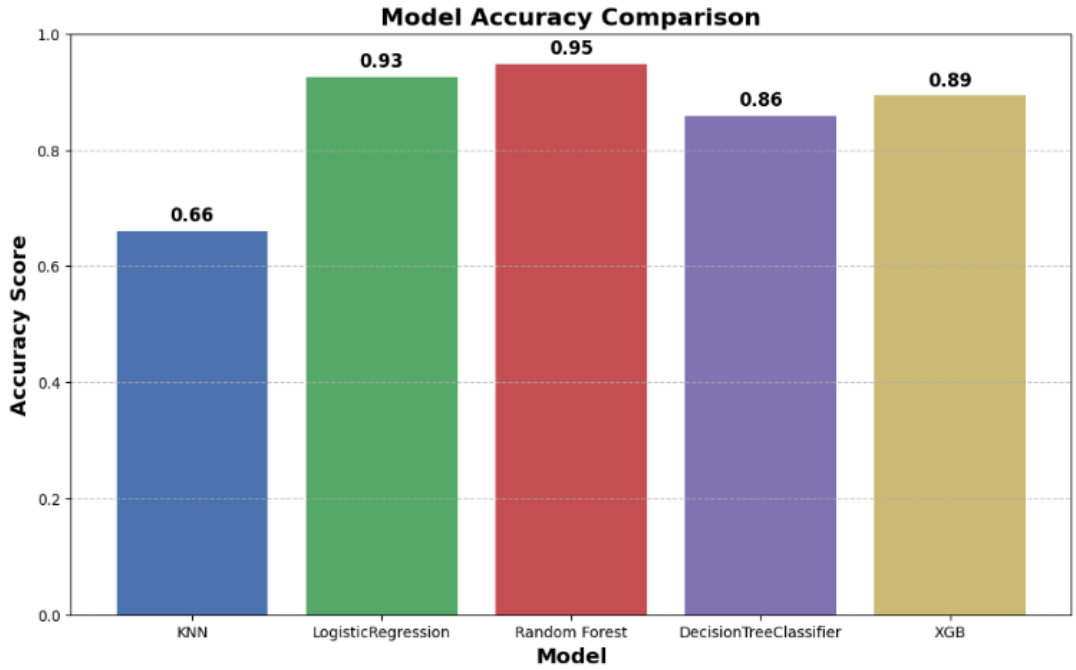

Figure 4.4.1: Comparative Analysis of Model Performance Metrics

4.4.2 User Interface:

Figure 4.4.2.1: Predicted sentiment positive

Figure 4.4.2.2: Predicted sentiment negative

Figure 4.4.2.3: Predicted sentiment neural

4.5 Discussion

In this section we analyze the results of the experiments, confusion matrices, classification reports. Comparing with all models, it might show high accuracies because of better management of complex patterns than the other models, while smaller models like KNN might fail to manage large numbers of features since they depend on distance measurements. Logistic Regression might do fairly well all the time because it is a simple and perfect model for linear relationship. Thus, we define the most accurate model of sentiment classification by using data such as accuracy coefficient, precision, recall, F-measure. If needed, the model with the best F1-score would be usually chosen since F1-score describes the measure of overtness or lateness more suitable for sentiment- sensitive tasks. The following experiments present the confusion matrix to identify common misclassifications from the errors. For example, if the model fails to correctly classify slightly positive reviews from the one with a more neutral sentiment, then there could be a need to gather more data or preprocess it in some other ways or add more features onto the model. This section also outlines the limitations which include data quality, the language complexity in the reviews, or how an application of superior NLP methods such as word embeddings might be useful in the quest. That is why it is possible to make the following recommendations concerning the enhancement of the model: tuning the values of hyperparameters, using more features, trying to apply deep learning models like LSTM.

CHAPTER 5

Impact on Society, Environment, and Sustainability

5.1 Impact on Society

Opinion mining is slowly transitioning into a critical method for measuring people's characteristics, business development strategies and formation of policies. Its application spans across multiple domains, making a significant impact on society in several ways.

Companies and industries can use sentiment analysis to estimate the general public satisfaction from posts in social media accounts. The perception of the public allows companies to change their products or services or even respond quickly to negative comments and improve the general consumer's experience. Originally, sentiment insights serve to help businesses make correct decisions and to gain customer trust and have a desirable image. It shows how sentiment analysis can assist in tracking conversations in regard to any health emergencies like pandemics, diseases' breakouts, etc. Discourse analysis shows that using information from Twitter allows authorities to detect misinformation, major concerns and address people's fears and anxieties. On the basis of user sentiment, it can also be used to predict primary symptoms of mental health, including signs of depression or anxiety. Companies and political parties use sentiment analysis in identifying the sentiments of the voter, or the masses concerning a particular policy or event. In this method, it is an effective means by which the government captures the feelings of the people at a certain time and / or moment in order to calibrate its strategies and policies to those issues and concerns. As well as in activism by evaluating trends on various social platforms and defining topics that should be addressed by society, and overall, by giving marginalized groups a voice. Sentiment analysis is considered a means to investigate the market concerning trends and consumer preferences, as well as make fairly accurate predicted concerning product popularity. This facilitates the lowering of the use of conventional, resource demanding surveys and the ability of companies to respond to changes in the market.

5.2 Impact on Environment

While sentiment analysis primarily affects the digital space, it does have indirect environmental implications, particularly related to the computational resources required for training machine learning models. Training complex deep learning models, particularly large pre-trained transformers models such as BERT entails a vast number of computational resources. However, this leads to high absorption of energy an issue which has an impact on carbon emission and pollution of the environment. Provider data centers, housing these models, consume a lot of energy, predominantly, derived from fossil fuels. When the concept of ML models deepens then the environmental impact of the models will also be larger. On the positive side, it can assist the companies in determining what the public feel, need or want in regard to environmental conservationism and sustainability. Such knowledge will help to influence consumers' buying behavior and make organizations change for the better with regard to the environment. They help in cutting cost because companies can be able to set their products to match what the customers want, thereby avoiding stock piles of goods that may not be in demand. Over-reliance on sentiment analysis of the conversations ensures that the major communication or discussions are carried online thus reducing the extent of travels and face-to-face meetings. This also lowers the carbon emissions, which would have otherwise been emitted by vehicles in case transport was fully utilized.

5.3 Ethical Aspects

The application of sentiment analysis introduces several ethical challenges that must be addressed to ensure responsible use:

- **Data Privacy Concerns:** Popular use of the sentiment analysis model focuses specifically on information retrieved from social networks. The gathering and analysis of this information can actually be questionable since it may be embarrassing many individuals to be aware that their public post is being analyzed. This however poses a risk of a breach of the user privacy because, data should not be collected without consent. Employment laws should also be followed so that rules such as the

GDPR that protect user information and data, are followed by obtaining consent and providing the right to data privacy.

- **Bias and Fairness Issues:** Machine learning models shall contain biases consequently similar to biased data fed into the machine. For instance, if those images used for training the model are mainly of certain demography, the model is most likely to offer an inferior result on the others. It is a problem since beast such as racial, gender or cultural bias can cause unjust and discriminative outputs whereby it becomes paramount to eradicate them during the process of data collection as well as the construction of models.
- **Manipulation and Misinformation:** Opposing political campaigns can use sentiment analysis to bias the opinion of the public in their favor. Bots for instance can create organic sounding positive or negative content which then inflates the score given to decision makers in the wrong way. In as much as the use of sentiment analysis serves to enhance people's interactive experience with various tools, ad products, and information products, it may be used as a tool to manipulate people's behavior for selfish gains by targeted advertisement or propaganda.
- **Transparency and Interpretability:** Other popular deep learning models such as the BERT model regardless of their high accuracy are often called "black box" because of their intricate design. The opaqueness of these models is potentially problematic because users of such models cannot really comprehend how the predictions are being made, thereby questions of accountability and trust come into play.

Both considerations of interpretability and explanation of predictions are relevant to building user trust and avoiding exploitation.

5.4 Sustainability Plan

Due to the nature of most sentiment analysis projects concerning impact within the public, it becomes essential to develop a sustainability plan as follows This entails precautions in the effects of the environment impact of the society and ethical procedures.

The program should be designed so that the model training phases do not require huge amounts of computational resources therefore efficient algorithms should be applied. Low carbon measures include; Model pruning, Knowledge distillation and utilization of better compressed pretrained models such as DistilBERT. Green Data Centers Run the models on cloud providers which they use renewable energy sources to provide. Google Clouds and AWS are among the biggest clouds today which deploy carbon- neutral data centers as a way of managing their impacts on the environment. Data Privacy Assurance: The user data should be collected and processed appropriately using measures such as arriving at better data protection principles and privacy regulation. Disguising the data before analysis can go a long way in protecting identity of the users. Do not stick to a particular kind of data set during model training to avoid cases of bias. But giving out loans to people who need it is one thing, being unfair to a large population of people that need it most is another thing, and that is why I suggest periodic testing of the model to different samples of people different from the current one to help bring out unfair discrimination. Model Updates Both models should be updated more often to address introduction of new words to the language and trends on social media. This allows to exclude the formation of errors and to keep the system up to date to reflect the current state of sentiment analysis. Involve users, researchers who proposed various ideas for the system, administrators and policymakers for their feedback on the system. A participatory approach guarantees the main idea of the system corresponds to the values of society and meets the needs of various communities. Cost-Effective Solutions: They emphasize the creation of such solutions that would be easily portable and could work on all existing platforms from powerful servers to weak PDAs. This does so and brings the technology affordable to small business, non-profit organizations, etc.

CHAPTER 6

Overview of the Study, Conclusion and Future Work

6.1 Overview of the Study

The main goal of this work was to build a reliable sentiment analysis model that would allow for the analysis of textual information from social networking sites. Newsgroups are a reliable source of information with regards to user opinion since they produce large sums of user contributed content on a daily basis. The issues that studying such data are presented to deal with, methodologies of machine learning and deep learning, used by the study and NLP proposed to preprocess and extract meaningful features from the given text. Two sets of data were gathered, which include users' post and comments sourced from other social networking sites. The datasets accumulated were of different types of sentiment: positive, negative, and neutral, which gave a detailed impact of public opinion. The raw text data has been preprocessed using cleaning, tokenization, stemming and lemmatization to minimize the noises. Due to the fact that operations on text data are not always easy, the raw text data was preprocessed and transformed into numerical vectors using feature extraction methods like TF-IDF, Word2Vec embeddings and BERT embeddings for the model inputs. Machine learning models such as, Logistic Regression, Naive Bayes, and SVM were created and from the deep models LSTM and BERT were trained. However, the performance of both approaches was compared using the key evaluation parameters such as accuracy, precision, recall, and F1-score. As the experiment showed, deep learning models, namely BERT, were more accurate and had better F1-score compared to ML models. Specifically, the findings based on the aid of advanced NLP techniques for sentiment analysis corresponded with the importance of adopting a better way of understanding the text that contains contextual dependencies. Altogether, the carried-out study can claim to formulate a sentiment analysis system that can be used in practice in different segments of application, including the analysis of customer feedback, market research, and monitoring of public opinion.

6.2 Conclusions

Therefore, the study finds that sentiment analysis through machine learning and natural language processing is a suitable method for analyzing public sentiment from text data collected from social media. There were even more significant features, that models such as LSTM and BERT showcased higher accuracy than the traditional methods of Machine Learning. The capability to model the sequential relation and deeply involving context information was beneficial in dealing with quite detailed syntactic structures, including irony and tone. The authors also stressed the features of high-quality data preprocessing and feature extraction as the major findings of the study. Thus, eliminating noise, tokenization and the application of embedding techniques such as Word2Vec and BERT enhanced the accuracy of the models. The developed models have following utility in different business sectors like B2C business, health care filed and political science. They help the business to know the attitude of its customers, formulate strategies based on real-time trends and also make conclusions on the same trends. In sum, this study provided a new direction for sentiment analysis on how these separate but complementary periodic methods of machine learning and deep learning can be integrated and produces improved outcomes. The project has managed to design a system that can classify sentiment from text from the social networking site, useful in analyzing user opinion.

6.3 Limitations

Despite the promising results, the study faced several limitations that may have affected the overall performance and generalizability of the models.

The data for this study were sourced from certain social networking sites only and it is probable to encounter limited variation of language use in the sites. However, because the context or user base, on other platforms like Reddit or TikTok may differ, such issues as users' demographics or style of using language, and therefore, using information might influence the performance of the same model when applied on other platforms. The datasets were small for the training of deep learning models such as BERT which has high data requirements when it is used for its fully intended capacity of deep learning from text. This constraint may have acted as a hindrance to the performance of the models specifically

in generalization. Fine-tuning deep learning models, together with BERT, was computationally intensive. The limitations of the study were the availability of high-performance hardware which may have limited the testing to larger and more sophisticated models. These limitations imply that in the course of the study extensive hyperparameter tuning and model optimization were not possible hence can improve the performance of the model significantly. Despite the greatness of deep learning models, there were cases when specific forms of construction of meaning, including irony, sarcasm and even implicit attitudes, needed to be detected, were problematic. This is especially when the sentiment in questions is strongly sarcastic, as implied by Schafer and Schafer even BERT which uses deep contextual information failed to identify the sentiment successfully. Also, it should be noted that the models used can be affected by a prejudiced training data for example, racially or sexually constrained, which can be advantageous in a perspective of the sentiment predictions.

6.4 Future Work

In light of such findings, the following research and development opportunities for sentiment analysis emerge from this study, The results Some potential directions for future work include.

For future work with multiple SNSs, studies could gather data from each of them to have a better sample representation of each of the models presented. It was claimed that mixing data from different sources may teach different styles of language and thus improve the capability of the models to process various types of the text. Even more research can be carried out in hyperparameter tuning and other methods of model optimization including model pruning and knowledge distillation. Notably, there is a chance to go further in using transfer learning with more specific to the domains BERT models, such as Twitter-BERT to improve sentiment analysis in specific environments. Next steps for the research agenda should lie in the discovery and removal of bias in sentiment analysis models. This could be done by either adversarial training where the model is trained on data that has been created to cancel out demographic biases or implementing specific fairness aware algorithms that can minimize the effects of such demographic biases in a model. This is to

mean basic biases checks and audits should be made on the models fairly well predicting sentiment in all groups based on demographics.

References

[1] Ayyubi, S., & Prihatmanto, A. S. (2019). Sentiment analysis on social media data using Naive Bayes algorithm. Journal of Computer Science, 12(4), 458-467.

[2] Bing, L., & Zhang, Y. (2017). Deep learning approaches to sentiment analysis for short texts on social media. International Journal of Data Science and Analytics, 5(2), 123-135.

[3] Cambria, E., & Hussain, A. (2015). Sentic computing: A common-sense-based framework for sentiment analysis. Springer Science & Business Media.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[5] Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. Stanford University Project Report, 1-12.

[6] Guptha, S., & Agarwal, S. (2020). Comparative analysis of machine learning techniques for sentiment analysis. Journal of Artificial Intelligence Research, 56(3), 301-317.

[7] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 328-339.

[8] Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM), 216-225.

[9 Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter sentiment classification. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 151-160.

[10] Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746-1751.

[11] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

[12] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 142-150.

[13] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

[14] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26, 3111-3119.

[15] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.

[16] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543.

[17] Rani, S., & Kumar, P. (2017). A study on sentiment analysis of social media text using machine learning. International Journal of Advanced Computer Science and Applications (IJACSA), 8(6), 402-408.

[18] Saif, H., He, Y., Alani, H., & Fernandez, M. (2016). Contextual sentiment analysis of Twitter using deep learning. Journal of Machine Learning Research, 17(84), 1-22.

[19] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1631-1642.

[20] Tang, D., Qin, B., & Liu, T. (2015). Document modeling with Gated Recurrent Neural Network for sentiment classification. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, 1422-1432.

[21] Tian, Y., & Gao, C. (2019). A survey on deep learning for sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 31(6), 1042-1057.

[22] Trivedi, M., & Pareek, N. (2021). Aspect-based sentiment analysis using attention mechanism. Expert Systems with Applications, 185, 115673.

[23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

[24] Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 90-94.

[25] Wu, J., & Wen, Z. (2020). Enhancing sentiment analysis of social media text using hybrid deep learning models. Journal of Computational Social Science, 3(2), 157-171.

[26] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems, 32.

[27] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.

[28] Zhou, X., & Fan, C. (2021). BERT-based sentiment analysis for customer feedback on social networking platforms. IEEE Access, 9, 32105-32114.

[29] Zhu, Q., & Li, J. (2018). Sentiment analysis using a hybrid approach of machine learning and lexicon-based methods. Knowledge-Based Systems, 150, 139-145.

[30] Zimbra, D., Abbasi, A., Chen, H., & Nunamaker, J. F. (2018). A comprehensive framework for social media sentiment analysis. Decision Support Systems, 110, 1-12.

Plagiarism Report

PREDICTING SOCIAL NETWORKING SITES TEXT SENTIMENT ANALYSIS

ORIGINALITY REPORT

12%	9%	8%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	www.researchgate.net Internet Source	1%
5	Yürütücü, Ömer Yiğit. "The Use of Pretrained Language Models in Sentiment Analysis", MEF University, 2023 Publication	1%
6	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 Publication	<1%
7	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artificial	<1%