**Internship/Project II (Comp – 8977)**

**Fall 2023 term**



**UNIVERSITY OF WINDSOR**

**Project Report: Sentiment Analysis of Twitter Data**

**Submitted by:** Jayandran Girisala 110094016 Varshitha Kambham 110087679

Prajwal Saleena Prakash 110093111

**Project Supervisor:** Dr. Prashanth Cheluvasai Ranga

**Course Instructor:** Dr. Kalyani Selvarajah

**Table of contents**

1.[Abstract…………………………………………………………………………………………………………………………………………3](#_TOC_250017)

2.[Introduction…………………………………………………………………………………………………………………………………..3](#_TOC_250016)

3.[Background of Twitter Sentiment Analysis…………………………………………………………………………..………..3](#_TOC_250015)

* 1. [Motivation and need for Twitter Sentiment Analysis…..………………………………………….……….4](#_TOC_250014)

3.2 Factors driving Twitter Sentiment Analysis………………………………………………………………………4

4.[Literature Review………………………………………………………………………………………………………………..………..5](#_TOC_250013)

5.[Project model and approach………………………………………………………………………………………………………….5](#_TOC_250009)

5.1.[Data Sets…………………………………………………………………………………………………………………………6](#_TOC_250007)

5.2.[Data Pre-Processing……………………………………………………………………..…………………………………6](#_TOC_250007)

5.3.Model Building and Evaluation……………………………………..………………………………………….……..8

5.4.Visual Representation ……………………………………………………………………………………………………9

6.[Results and Output screens…………………………………………………………………………..…………………………….10](#_TOC_250002)

7.[Conclusion and Future Scope…………………………………………………………………………..………………………….1](#_TOC_250001)9

8.[References………………………………………………………………………………………………………..………………………..21](#_TOC_250000)

9.GitHub Link………………………………………………………………………………………………………………………………….22

# 1.Abstract:

With an emphasis on data analysis from Twitter, this project report investigates the area of sentiment analysis in social media. The work focuses on using a combination of machine learning algorithms and natural language processing techniques to classify tweets into numerous sentiment classifications, including the recognition of hate speech. This comprehensive approach employs a range of models, such as logistic regression, support vector machines, random forest, and XGBoost, in addition to extensive preprocessing and feature extraction. Apart from providing an understanding of the sentiment distribution in social media conversations, the study significantly advances the ongoing research on digital communication analysis, hate speech identification, and the iPhone and MacBook.

# 2.Introduction:

The objective of the Twitter Sentiment Analysis project is to use natural language processing and machine learning techniques to automatically identify the sentiment of tweets. The project's objective is to provide relevant public opinion data that corporations, governments, individuals, and social media content moderators may use.

# 3.Background of Twitter Sentiment Analysis:

Social media platforms such as Twitter store large amounts of unstructured text data, necessitating a range of methods and techniques for effective analysis. Among them, the Bag of Words (BoW) technique represents text data by tabulating word frequencies in texts; it ignores word order, yet it nevertheless performs very well for basic sentiment analysis tasks. Furthermore, the TF-IDF (Term Frequency-Inverse Document Frequency) measure offers valuable information by assessing a word's relevance in a document in relation to a set of documents. Finding key phrases in tweets for sentiment analysis is made much easier with this. In addition, this project employs a range of machine learning models, including Naive Bayes, Decision Trees, Support Vector Machines, and Neural Networks, for sentiment categorization. The pivotal process of model tuning enhances their performance and promotes better generalization to new data, facilitating more accurate sentiment analysis.

## 3.1Motivation and need for Twitter sentiment analysis:

The goal of this project is to leverage sentiment analysis's capabilities to offer a multitude of benefits. It first aids businesses in comprehending the thoughts and perspectives of their clientele, allowing them to enhance their products and services and make informed decisions. Additionally, it makes it easier for government organizations to understand what the public thinks about a range of policies and subjects, which promotes greater public participation and engagement in the political process. The purpose of this project is to give people the tools they need to monitor their online presence, assess their reputation, and learn what the public thinks of the content they have produced. Additionally, it enhances social media content monitoring and promotes a safer online environment by promptly identifying potentially harmful or abusive tweets.

3.2 Many **factors** can drive Twitter sentiment analysis, including:

1. Business intelligence: Businesses can utilize Twitter sentiment research to learn more about the tastes and opinions of their customers, which can help them improve their product development, customer service, and marketing tactics.
2. Reputation management: Businesses may keep an eye on how consumers are feeling about their brands and products on Twitter to safeguard their online image and respond quickly to any negative comments.
3. Political analysis: Sentiment analysis on Twitter can assist politicians and policymakers make educated judgements by revealing the public's opinions on political matters.
4. Customer service: Businesses can increase overall customer happiness by promptly identifying and resolving consumer complaints or concerns by utilizing sentiment analysis.
5. Brand monitoring: Social media monitoring tools may find mentions of a product, service, or brand, allowing businesses to engage with customers, monitor the reputation of their brands, and track competitors.

# 4.Literature Review:

Sentimental analysis has several uses in a variety of fields, including business, where it is used to gather customer feedback on items so that organizations may learn from online evaluations and comments.

Understanding the material, summarizing it, and assigning weight to its good, negative, or neutral aspects are possible solutions for longer texts. The extractive and abstractive methods are the two main strategies for extracting text summaries. The extractive approach creates a summary by taking words and word groups out of the source text. A summary that is more like what a person would provide would be produced using an abstractive technique that seeks to acquire an internal language representation.

# 5.Project model and approach:

For this project, we used Twitter to extract tweets, cleaned and pre-processed the retrieved data, used TextBlob to analyse sentiment, created a word cloud to show the most frequently used words, and represented positive and negative phrases.

Tweets are short text messages with a character count of up to 280 that make up most of the data on Twitter. A tweet can contain text, material such as images, videos, and GIFs, hashtags, mentions, and links.

Metadata from Twitter accounts for users includes the time and date of the tweet, their location, their user ID, how many likes and retweets they have received, and how many followers and followers they have. Twitter data can be used for a variety of purposes, such as sentiment analysis, social network analysis, topic modelling, and trend analysis.

It is important to remember that Twitter data is a sample of the public and may not fully represent the views of the general public. It's possible that Twitter users are not representative of the public because they tend to be younger and more politically engaged than the general population. As such, it is imperative to exercise caution when extrapolating conclusions or making generalisations from Twitter data.

Sentiment data is information from Twitter that has been classified as either good, negative, or neutral according to the content of tweets. This data may be used to train machine learning models that will analyse sentiment in newly collected, unlabeled Twitter data. A ready-to-use solution for sentiment analysis on Twitter is provided by pre-trained models like TextBlob. These models can analyse sentiment on new Twitter data without the requirement for manual labelling because they have been trained on big datasets.

## 5.1Data Sets:

## Hate\_Speech\_Detection

## iPhone Customer Reviews

## MacBook Customer Reviews

## 5.2Data Pre-Processing:

The data preprocessing in the provided code is a critical phase where the raw tweet data is cleaned and transformed into a format suitable for analysis. Here's a detailed breakdown of each step in the preprocessing workflow:

**Loading the Data:**

The code begins by loading the training and testing datasets using Pandas.

**Combining Datasets:**

The training and testing datasets are combined into one dataset (combi) for consistent preprocessing.

**Removing Twitter Handles (@user):**

A custom function remove pattern is defined to remove Twitter handles (text starting with '@') using regular expressions.

This function is applied to each tweet in the combined dataset.

**Removing Special Characters, Numbers, and Punctuations:**

All characters in the tweets that are not alphabets or hashtags are replaced with spaces. This is achieved with combi['tidy\_tweet'].str.replace("[^a-zA-Z#]", " ").

**Removing Short Words:**

Short words, typically less than 4 characters, are removed as they are generally not significant for analysis.

**Tokenization:**

Tweets are split into individual words (tokens) using the split() method.

**Stemming:**

Each token is stemmed to its root form using NLTK's PorterStemmer. Stemming helps in reducing the complexity by converting different forms of a word to a base form.

**Rejoining Tokens:**

The stemmed tokens are then joined back to form the cleaned tweet strings.

**Word Cloud Visualization:**

Word clouds are generated for all tweets, as well as separately for positive and negative tweets, to visualize the most frequent words in each category.

**Extracting Hashtags:**

A function hashtag extract is defined to extract hashtags from the tweets.

This function is used to separately extract hashtags from positive and negative tweets.

**Feature Extraction using Bag of Words and TF-IDF:**

The cleaned tweets are transformed into numerical features using both Count Vectorizer (Bag of Words) and TfidfVectorizer (TF-IDF).

These numerical features are used for training the machine learning models.

**Sentiment Analysis using TextBlob:**

The TextBlob library is used to classify the sentiment of tweets in both the training and testing datasets. This part is more focused on sentiment polarity determination (positive, negative, neutral).

Each of these steps systematically prepares the raw tweet data, making it suitable for the subsequent sentiment analysis task. The preprocessing removes irrelevant information (like handles and special characters), normalizes the text (through stemming), and converts the text into a numerical format that can be fed into machine learning models for classification.

**5.3Model Building and Evaluation :**

**Introduction**

The focus of our work is covered in this section, where we develop and assess machine learning models to categorise tweets as either hate speech or non-hate speech. We used four different algorithms: XGBoost, Random Forest, Support Vector Machine (SVM), and Logistic Regression, to see which one worked best for our dataset. For the final two datasets, we applied the same strategy.

**Model Selection Rationale:**

* Logistic Regression: It is renowned for being easy to use and effective at solving binary classification issues. It acts as a reference model that can be used to compare more intricate models.
* Support Vector Machine (SVM): It works well in high-dimensional spaces, which makes it appropriate for text classification with a large number of characteristics (words).
* Random Forest: A group technique with several decision tree classifiers. It lowers the chance of overfitting by providing resilience and a strong capacity for generalization.
* XGBoost: A gradient boosting framework-based ensemble machine learning technique based on decision trees. renowned for its effectiveness and quickness, especially with structured datasets.

**Preprocessing for Model Input:**

Each model required the input data to be transformed into a numerical format. This was achieved through two primary methods:

* Bag-of-Words (BoW): Text is converted into fixed-length vectors by counting the occurrences of each word.
* TF-IDF (Term Frequency-Inverse Document Frequency): It evaluates a word's significance inside the corpus by taking into account its frequency throughout the dataset as a whole, not just in a particular document.

**Training and Hyperparameter Tuning:**

* Splitting the Data: The dataset was split into training and validation sets to enable model evaluation.
* Model Training: Each model was trained on the BoW and TF-IDF feature sets. For XGBoost, further tuning was conducted to optimize parameters like max\_depth, min\_child\_weight, and eta.
* Hyperparameter Tuning for XGBoost: A grid search approach was used to iteratively explore combinations of parameters to find the most effective model settings.

**Model Evaluation Metrics:**

* F1 Score: Chosen for its balance between precision and recall, particularly important in datasets with imbalanced classes.
* Confusion Matrix: Used to understand the type of errors (false positives and false negatives) each model was making.
* ROC Curve and AUC: For assessing the models' performance in distinguishing between the classes.

**Comparative Analysis and Findings:**

* Performance Overview: Each model's performance was evaluated based on the F1 score, confusion matrix, and AUC from the ROC curve.
* Best Performing Model: XGBoost emerged as the top-performing model. Its superiority can be attributed to its ability to handle the sparse nature of text data and efficiently manage the trade-off between bias and variance.
* Insights from Evaluation: The evaluation highlighted the strengths and weaknesses of each model. While Logistic Regression and SVM offered baseline performances, Random Forest and XGBoost showed improved capability in handling the complexity and nuances of the dataset.

## 5.4Visual representation:

## WordCloud

The most frequently occurring positive and negative words are highlighted and displayed in a larger font size than the less frequently occurring ones in a word cloud, a visual representation of text data. Text analysis often uses word clouds to find the most important words or subjects rapidly in a given text.

One can create a word cloud in a variety of ways. Before anything further, the text data is cleaned and processed to remove any unwanted characters, such as stop words or punctuation. A frequency distribution of words is then produced by counting the instances of each word in the text. Next, the most often used terms are selected to be included in the word cloud, and the less frequently used terms are removed.

The word cloud can then be made using the list of terms that appear frequently. This is typically accomplished via software or an internet tool that takes the word list and produces a visual representation of the text. The most frequently used terms in a word cloud are often displayed with a larger font size and a more prominent position than the less frequently used words. The frequency of a word in the text is taken into consideration by an algorithm that sets the font size and position of each word.

Because word clouds provide a rapid and easy way to identify the most important phrases or ideas in a document, they are a popular approach of visualising text data. Data journalism, social media analysis, and marketing are just a few of the fields that commonly use them. Word clouds can oversimplify text data and fail to properly capture the complexity of underlying ideas or moods, thus it's important to be aware of their limitations.Making a word cloud is an easy and efficient approach to see the most frequent terms in a document.

* **Frequency Chart**
* **Pie Chart**
* **Confusion Matrix**
* **ROC Curve Plot**

# 6.Results and Output screens:

# Dataset 1 : Hate Speech Detection

# Positive Words using Word Cloud

# A screenshot of a computer Description automatically generated

# Negative Words Using Word Cloud

# A screenshot of a computer screen Description automatically generated

# Frequency Words for the Positive Words

# A screen shot of a computer Description automatically generated

# Frequency Words for the Positive Words

# A graph of different colored bars Description automatically generated with medium confidence

# Training Dataset Visualization

# A pie chart with numbers and a few negatives Description automatically generated with medium confidence

# Testing Dataset visualization

# A pie chart with different colored circles Description automatically generated

# Confusion Matrix for XGBoost\_BoW.

# A blue and white graph Description automatically generated

# Confusion Matrix for XGBoost\_TF-IDF.

# A graph with blue squares and numbers Description automatically generated

# ROC Curve for XGBoost

# A graph of a function Description automatically generated with medium confidence

# Confusion Matrix and ROC Curve for Random Forest Model

# A screenshot of a graph Description automatically generatedA line graph with a blue line Description automatically generated with medium confidence

# Confusion Matrix and ROC Curve for SVM

# A blue and white graph Description automatically generatedA graph with a line and a point Description automatically generated with medium confidence

# Confusion Matrix and ROC Curve for Logistic Regression

# A blue squares with white text Description automatically generatedA line graph with a blue and orange line Description automatically generated

Overall Sentiment Score For Hate\_Speech\_Detection: 0.011764705882352941

# Dataset 2 : iPhone Customer Reviews

Overall Sentiment Score For iPhone Customer reviews: 0.5044

A screenshot of a white text

Description automatically generated

# Dataset 3: MacBook Customer Reviews

Overall Sentiment Score For iPhone Customer reviews: 0.3016

A screenshot of a computer

Description automatically generated

# 7.Conclusion:

In summary, the sentiment analysis research showcased here marks a noteworthy advancement in the application of natural language processing (NLP) and machine learning to comprehend social media discourse, particularly on Twitter. By carefully cleaning, tokenizing, and stemming tweet datasets, the project successfully prepared them for analysis. The robustness of these algorithms in classifying tweets into sentiment categories was demonstrated by the implementation of techniques like XGBoost, Random Forest, Support Vector Machines (SVM), and Logistic Regression. XGBoost, however, displayed the best categorization out of the four. The creation of word clouds improved the interpretability of the results by providing an easy-to-understand visual representation of the most frequently used terms across several sentiment categories. The integration of these methodologies showcased the potential of machine learning in distilling meaningful insights from vast and noisy social media data, providing a snapshot of public sentiment that is invaluable for various applications, from market analysis to socio-political studies.

The research effectively handled the difficulties that come with sentiment analysis, including the nuances of human language and contextual ambiguity. The application of feature extraction methods such as TF-IDF and Bag of Words improved the model's comprehension and categorization of textual material. The collective results highlight the increasing applicability and efficacy of automated sentiment analysis in the current digital era, where social media platforms serve as vast stores of user opinions and sentiment.

# Future Scope:

In the long run, the initiative creates multiple opportunities for improvement and growth. Incorporating more complex NLP techniques, like transformer models (like BERT) and word embeddings (like Word2Vec or GloVe), is one immediate topic of future effort. In particular, these techniques are better at handling sarcasm, idioms, and nuanced expressions of sentiment than older models, which results in more accurate sentiment classification.

Using this approach with real-time data is another exciting avenue. Businesses, politicians, and researchers would have quick access to public sentiment data if a system capable of processing and analyzing tweets in real-time were developed. Considering the varied language environment of social media, extending the model to incorporate multilingual analysis will also make it more inclusive and globally relevant.

Finally, investigating neural network topologies and deep learning strategies may improve the model's functionality even more. To handle sequential data like text more effectively, this may entail experimenting with alternative layers and activation functions or using recurrent neural networks (RNNs) and long short-term memory networks (LSTMs).

To sum up, the initiative creates the groundwork for sophisticated sentiment analysis applications and paves the way for increasingly intricate, real-time, and multilingual studies. With its success, sentiment analysis powered by AI may soon become a vital tool for comprehending and utilising the massive amounts of data produced in our increasingly digital environment.

# 8.References:

1. S. Kiritchenko, X. Zhu, and S. M. Mohammad, ‘‘Sentiment analysis of short informal texts,’’ J. Artif. Intell. Res., vol. 50, no. 1, pp. 723–762, 2014.
2. X. Fang and J. Zhan, ‘‘Sentiment analysis using product review data,’’ J. Big Data, vol. 2, no. 1, pp. 1–14, 2015.
3. Chinese Text Sentiment Analysis Based on Extended Sentiment Dictio- nary published in 2019 by GuixianXu ,Ziheng Yu , Haishen Yao , Fan Li , YuetingMeng and Xu Wu.
4. R. Wang et al., ‘‘Research of text sentiment classification based on improved semantic comprehension,’’ Comput. Sci., vol. 44, no. 11A, pp. 92–97, 2017.
5. Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), pp.1093-1113.
6. Prabowo, R. and Thelwall, M., 2009. Sentiment analysis: A combined approach. *Journal of Informetrics*, *3*(2), pp.143-157.
7. Sahayak, V., Shete, V. and Pathan, A., 2015. Sentiment analysis on twitter data. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, *2*(1), pp.178-183.
8. Go, A., Huang, L. and Bhayani, R., 2009. Twitter sentiment analysis. *Entropy*, *17*, p.252.
9. Bhuta, S., Doshi, A., Doshi, U. and Narvekar, M., 2014, February. A review of techniques for sentiment analysis of twitter data. In *2014 International conference on issues and challenges in intelligent computing techniques (ICICT)* (pp. 583-591). IEEE.
10. Diyasa, I.G.S.M., Mandenni, N.M.I.M., Fachrurrozi, M.I., Pradika, S.I., Manab, K.R.N. and Sasmita, N.R., 2021, May. Twitter Sentiment Analysis as an Evaluation and Service Base On Python Textblob. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1125, No. 1, p. 012034). IOP Publishing.

**9.GitHub Link:** <https://github.com/prajwalsaleena/FINAL_MAC_PROJECT>