**Project-**

**Sentiment Analysis of Tweets**

# ABSTRACT

As social media networks become more prevalent, the need to gauge public opinion, trends, and consumer perspectives from social media platforms such as Twitter has equally increased. The present work undertakes a sentiment analysis of Twitter’s microblogging service, identifying sentiments as either ‘positive’ or ‘negative’. These strong and explicit messages are meant to be expressed across different platforms, and in this case, we focus on Twitter. Our method addresses and overcomes this issue through extensive preprocessing, which includes splitting sentences into tokens, eliminating unnecessary words, and stemming, leading to division and manipulation of textual content into multiple parts or a bag of words. Each of the models is tested for effectiveness on accuracy and F1-score which aims at achieving this on the training dataset as well as on the validation dataset. This aspiration has seemed to have been met as the various models employed have shown to be consistent with the research aims and objectives of this work. This work successfully demonstrates how Gensim-based models could be used for sentiment analysis on social media posts, allowing brand and public opinion assessments about specific aspects of the society.

# INTRODUCTION

Social media platforms allow people to reach a broader audience and circulate their ideas. Although there are many advantages to this, it has also paved the way for the use of hateful or offensive speech. This project seeks to address this concern through the development of a text classification model that analyses tweets and classifies them into either negative (contains racist or sexist tweets) or positive/neutral. The effort is looking to reduce harmful online materials by flagging and automatically detecting them.

In this case, the project employs the text classification techniques in a sequential manner. Initially, the dataset which contains thousands of tweets is stripped off unwanted characters, stop words and non-alphabetic texts. Afterward, a number of natural language processing (NLP) methods including tokenization, stemming, and hashtag extraction are used to facilitate the text analysis process. We also employ exploratory data analysis (EDA) to gain insights into the dataset. Such tasks include observing word counts and distributions in visual formats, generating word clouds, and analysing hashtag usage to explore the popular themes or words used in negative and positive sentiments. These processes provide a good basis for the construction of an efficient and effective classification model.

# UNIQUE SELLING POINTS

## Word2Vec Embeddings for Contextual Understanding:

Using Gensim’s Word2Vec embeddings enhances your model’s ability to capture word meanings based on context. Unlike simpler bag-of-words or lexicon-based approaches, this allows for nuanced understanding of tweet sentiment, capturing complex expressions, slang, and varying tones typical of Twitter language.

## Effective Preprocessing Pipeline for Social Media Texts:

The project’s preprocessing steps (tokenization, stemming, stopword removal, and hashtag extraction) are tailored to Twitter's unique text structure. This pipeline enables the model to handle informal language, hashtags, and other Twitter-specific elements effectively, which is often challenging in sentiment analysis. Also inclusion of exploratory data analysis (EDA), such as histograms of tweet lengths and hashtag trends, makes the project more understandable and readable.

## Balanced Approach to Sentiment Classification:

By using F1-score as a key performance metric, your project emphasizes balanced accuracy between positive and negative tweet classifications, making it suitable for applications where both precision and recall are critical (e.g., brand reputation monitoring).

## Use of Multiple Machine Learning Models:

Instead of relying on a single model, the project evaluates several algorithms, including Random Forest, Logistic Regression and Decision Tree. This ensures robust comparisons and allows the selection of the most effective model for the task. It is highly adaptable for the real-world deployment.

# OBJECTIVES

1. **To develop a robust sentiment analysis model** that accurately classifies tweets as positive or negative, utilizing Gensim’s Word2Vec for word embeddings and various machine learning algorithms.
2. **To preprocess and prepare Twitter data** by implementing NLP techniques such as tokenization, stopword removal, stemming, and hashtag extraction, ensuring the data is suitable for machine learning model training.
3. **To design an adaptable sentiment analysis pipeline** that can be fine-tuned for various domains or specific sentiment analysis applications, such as brand monitoring, opinion tracking, or social issue analysis.
4. **To evaluate the scalability and efficiency of the model pipeline** for potential real-time sentiment tracking on social media platforms, taking into account the high volume of tweets.

# PROBLEM STATEMENT

Social media platforms have seen a rise in harmful content, such as racist and sexist comments, which negatively impact online communities. Manual content moderation is time-consuming, necessitating automated solutions for identifying and flagging offensive content.

This project aims to develop a text classification model that can automatically classify tweets as either positive/neutral or negative (racist/sexist). The model will help in automating content moderation, ensuring faster identification of harmful content and promoting safer online environments.

# METHODOLOGY

This report provides an overview of a comprehensive workflow designed for text classification, specifically focused on processing and analyzing tweets. The procedure is structured into multiple steps, each with a defined purpose to ensure efficient data processing and model training.

## Data Loading

The workflow begins with the loading of the train and test datasets. This foundational step ensures that the necessary data is available for the subsequent analysis and model training phases.

## Data Preprocessing

Data preprocessing is critical to enhance the quality and suitability of the data for model training. This step includes:

* Checking for Null Values: Ensuring data completeness by identifying and handling missing data.
* Data Cleaning: Removing special characters, tokenizing text, and applying stemming to normalize the data.
* Adding Tweet Length Column: This additional feature helps analyze the text length's impact on classification outcomes.

## Exploratory Data Analysis (EDA)

EDA is conducted to gain insights into the data and identify patterns or anomalies. This involves:

* Visualizing Data Distribution: Comparing the distribution of positive and negative labels to understand the dataset's balance.
* Visualizing Word Frequencies: Creating bar plots to observe the most frequent words in the data.
* Generating Word Clouds: Developing visual word clouds for positive and negative tweets to highlight the most commonly used terms.
* Hashtag Analysis: Extracting and analyzing hashtags to assess their influence and trends across positive and negative tweets.

## Feature Extraction

The feature extraction process converts the raw text data into a numerical format that machine learning models can process. This includes:

-CountVectorizer (BoW): Using the Bag of Words approach to create numeric representations of the text.

-Word2Vec Embeddings: Generating word embeddings to capture the semantic relationships between words.

## Data Splitting

The dataset is split into training and validation sets to enable model training and evaluation. This step ensures that the model's performance can be assessed on unseen data, providing a realistic measure of its effectiveness.

## Feature Scaling

Standardizing the features ensures that all variables contribute equally to the model, preventing any one feature from dominating due to scale differences.

## Model Training and Evaluation

Multiple machine learning models are trained to determine the most suitable one for the classification task. The models used in this process include:

* Random Forest

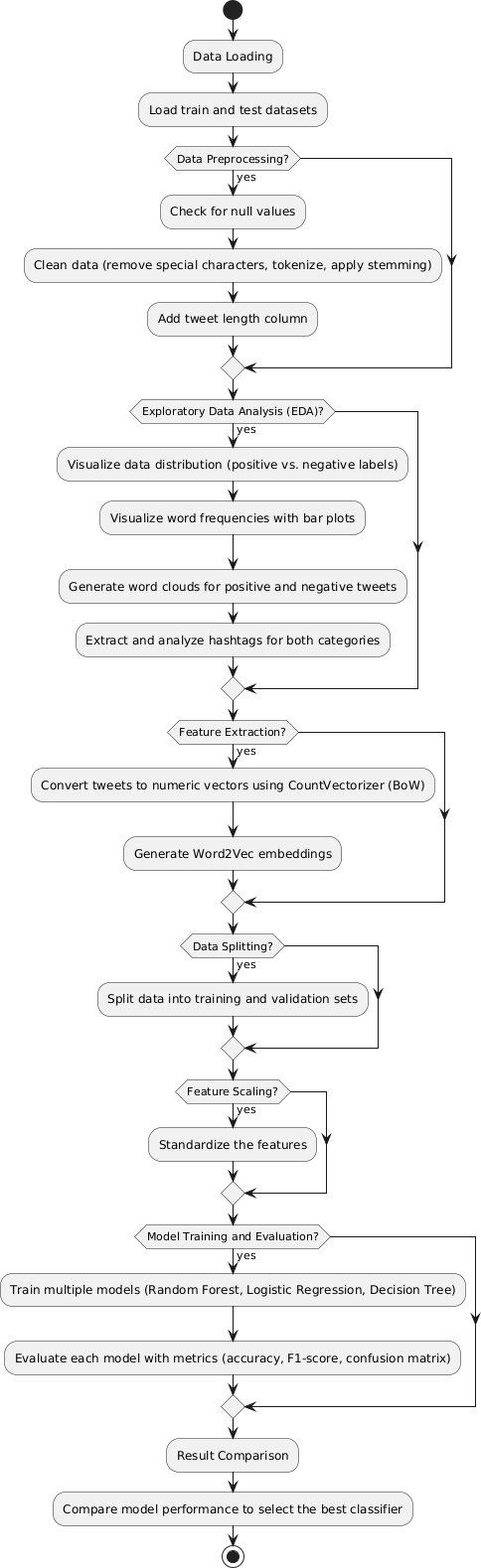
-Logistic Regression

-Decision Tree

Each model is evaluated using metrics such as accuracy, F1-score, and confusion matrix to provide a detailed understanding of its performance.

## Result Comparison

Finally, the results from all trained models are compared to identify the best-performing classifier. This step is crucial for selecting the model that offers the highest predictive accuracy and robustness.

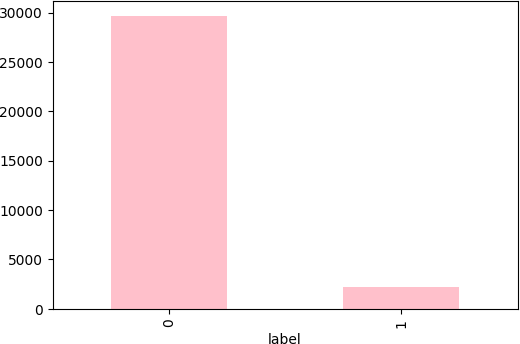


# EXPERIMENTAL RESULTS

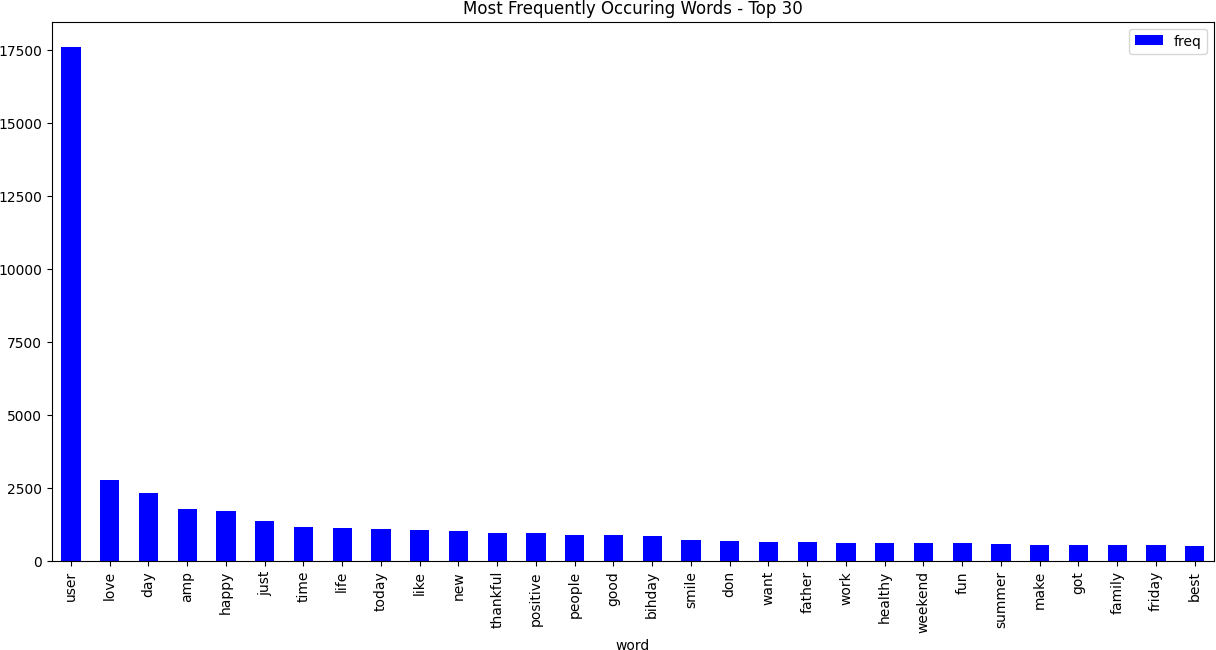
## Datasets:

https://drive.google.com/drive/folders/1\_KICz1aIOX5KmjJjYPiayU2Il4Qrx QVq?usp=drive\_link

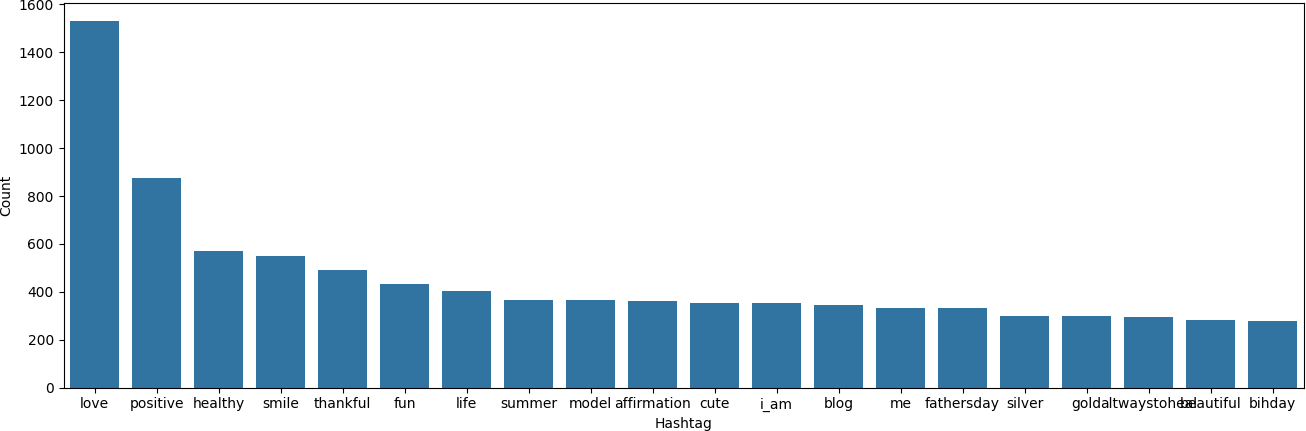
## Graphs:



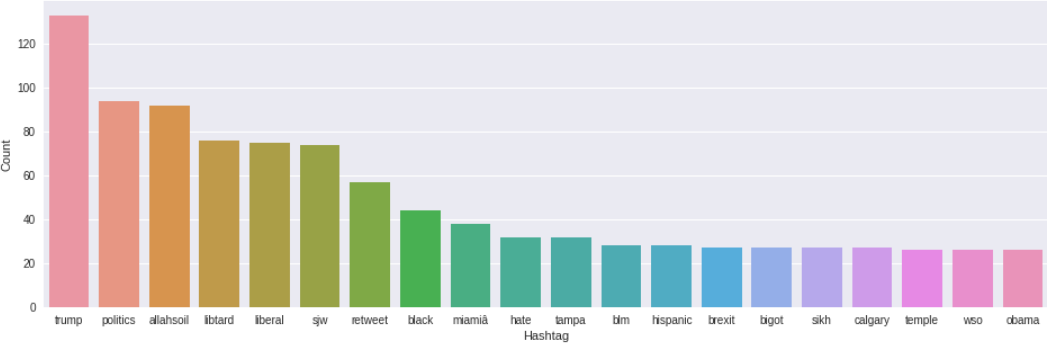
* + 1. **Fig 1.1: Imbalanced data**



## Fig 1.2: Most Frequently Occuring Words



* + 1. **Fig 1.3: Positive hashtags**



## Fig 1.4: Negative hashtags

* 1. **Results:** Accuracy of the following models is-
     1. Random Forest : 95.14%
     2. Logistic Regression: 94.05%
     3. Decision Tree: 93.25%

# CONCLUSION

This project successfully developed a sentiment analysis model for Twitter data using NLP techniques and machine learning. Through a structured approach that included comprehensive data preprocessing, feature extraction with Word2Vec embeddings, and experimentation with multiple classifiers, we achieved a robust model capable of accurately classifying tweets as positive or negative. Among the models tested, the random forrest classifier demonstrated the highest accuracy and F1-score, indicating its strong capability in distinguishing tweet sentiments. The integration of Word2Vec allowed the model to understand context and nuances in language, making it particularly suitable for handling the informal and expressive nature of social media texts. The project results suggest that this approach is effective for sentiment analysis tasks and has potential applications in fields such as brand monitoring, public opinion tracking, and customer feedback analysis on social platforms.

# FUTURE SCOPE

## Real-Time Sentiment Analysis:

* The current model can be extended to analyze tweets in real-time, enabling applications in social media monitoring for brands, governments, and organizations. By adapting the model to stream tweets and analyze them continuously, stakeholders can gain insights into live trends and emerging topics.

## Integration with Visualization Dashboards:

* Developing an interactive dashboard for visualizing real-time sentiment trends and key topics would enhance the practical utility of the model. Such dashboards could include metrics like sentiment trends over time, frequently discussed topics, and geo-tagged sentiment analysis for location-based insights.

## Multi-Language Support:

* Adding support for multiple languages would enable the model to analyze sentiments from a diverse set of users globally. Training on multilingual datasets and leveraging embeddings such as multilingual BERT could extend the model’s usability beyond English, capturing sentiment from non-English-speaking regions.

# REFERENCES

1. https://scikit-learn.org/1.5/modules/generated/sklearn.linear\_model.Logistic Regression.html
2. <https://scikit-learn.org/1.5/modules/tree.html>
3. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomF orestClassifier.html