Vivekanand Education Society's Institute of Technology (An Autonomous Institute Affiliated to University of Mumbai,) (Approved by A.I.C.T.E and Recognized by Govt. of Maharashtra)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

A REPORT

ON

"Twitter Sentiment Analysis"

B.E. (AIDS)

SUBMITTED BY

Mr. Prathmesh Dubey (Roll No.16)
Mr. Sahil Gupta(Roll No.20)

UNDER THE GUIDANCE OF

PROF. Anjali Yeole

(Academic Year: 2024-2025)

Vivekanand Education Society's Institute Of Technology, Mumbai

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

Certificate

This is to certify that project entitled

”Twitter Sentiment Analysis”

Mr. Prathmesh Dubey (Roll No.16)
Mr. Sahil Gupta (Roll No.20)

have satisfactorily carried out the project work, under the head - NLP Lab at Semester VII of BE in AIDS as prescribed by the Syllabus.

Subject Teacher **Lab Teacher**

Date: 12/10 /2024
Place: VESIT, Chembur

Declaration

I declare that this written submission represents my ideas in my own words and where other’s ideas or words have been included, I have adequately cited and ref- erenced the original source. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

- - - - - - - - - - -

(Signature)

Mr. Prathmesh Dubey (Roll No.16)

Mr. Sahil Gupta (Roll No.20)

B.E. AIDS

INDEX

Abstract

This project presents a comprehensive analysis of sentiments expressed in tweets using the Twitter Sentiment140 dataset. The primary objective is to classify tweets into positive and negative sentiments employing a Logistic Regression model. The dataset, obtained from Kaggle, consists of approximately 1.6 million tweets, which are pre-processed to enhance analysis quality. Key preprocessing steps include data cleaning, stemming, and removal of stopwords to focus on significant terms.

The processed text data is transformed into numerical features using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique, facilitating the model's ability to learn from the text. A Logistic Regression classifier is trained on a training set, with model evaluation conducted on a separate test set to ensure robust performance assessment. The model achieves an accuracy score of 77.8%, indicating a reliable capability to classify sentiments accurately.

In addition to the accuracy evaluation, the project emphasizes the importance of effective data processing and feature extraction in natural language processing tasks. The results illustrate the model's potential to gauge public sentiment on social media, providing valuable insights into trends and opinions. Future work will explore enhancements through more sophisticated algorithms and deep learning techniques, aiming to improve classification accuracy and expand the model's application in real-time sentiment analysis.

1. Introduction

1.1 Background

In the era of social media, platforms like Twitter have transformed the way individuals and organizations communicate. With over 330 million monthly active users, Twitter serves as a powerful tool for sharing opinions, news, and experiences in real-time. This vast amount of user-generated content presents a unique opportunity to analyze public sentiment on various topics, from political events to product launches. Sentiment analysis, a subfield of natural language processing (NLP), focuses on determining the emotional tone behind textual data, enabling stakeholders to gauge public opinion and respond effectively.

This project aims to leverage the Twitter Sentiment140 dataset, which contains millions of labeled tweets, to develop a robust sentiment analysis model. By classifying tweets into positive, negative, and neutral categories, we can uncover insights into public sentiment and its fluctuations over time. The growing importance of understanding public sentiment is evident in various sectors, including marketing, where businesses seek to analyze customer feedback and adapt their strategies accordingly.

Despite significant advancements in sentiment analysis, challenges such as sarcasm, context, and nuanced language still hinder the accuracy of automated models. Therefore, this project not only focuses on building an effective classification model using Logistic Regression but also emphasizes the preprocessing of text data to enhance the model's performance. Through this endeavor, we aim to contribute to the ongoing efforts in refining sentiment analysis techniques, ultimately aiding decision-makers in interpreting public sentiment accurately.

1.2 Problem Statement

In today’s digital age, social media platforms such as Twitter generate vast amounts of user-generated content daily. This content often reflects public opinions, sentiments, and emotions regarding various topics, products, and services. However, extracting meaningful insights from such unstructured data presents a significant challenge.

The objective of this project is to develop a machine learning model to analyze sentiments expressed in tweets. Specifically, we aim to classify tweets into three categories: **negative**, **neutral**, and **positive** sentiments. Utilizing the Sentiment140 dataset, which comprises 1.6 million tweets annotated with sentiment labels, we will explore various natural language processing techniques and machine learning algorithms to achieve accurate sentiment classification.

Key Challenges:

1. **Data Preprocessing**: Handling noise and inconsistencies in the text data, such as slang, abbreviations, and emoticons.
2. **Feature Extraction**: Identifying relevant features from the text that contribute to accurate sentiment prediction.
3. **Model Selection**: Evaluating different machine learning models to determine the most effective approach for sentiment classification.
4. **Performance Evaluation**: Implementing robust metrics to assess the accuracy and reliability of the models.

By addressing these challenges, this project aims to provide a comprehensive solution for sentiment analysis that can be applied to various applications, such as market research, customer feedback analysis, and social media monitoring.

1.3 Objectives

1. **Data Preprocessing**:
    a. Load the Sentiment140 dataset and preprocess the text data, including:
        i. Removing special characters, links, and stop words.
        ii. Normalizing the text through stemming or lemmatization.
2. **Feature Extraction**:
    a. Utilize vectorization techniques, such as:
        i. Bag-of-Words (BoW)
        ii. Term Frequency-Inverse Document Frequency (TF-IDF)
    b. Transform the cleaned text data into numerical features suitable for model training.
3. **Model Training**:
    a. Implement and train multiple machine learning models, including:
        i. Logistic Regression
        ii. Support Vector Machine (SVM)
        iii. XGBoost
    b. Optimize model hyperparameters to enhance performance.
4. **Model Evaluation**:
    a. Evaluate the trained models using metrics such as:
        i. Accuracy
        ii. Confusion Matrix
        iii. Classification Report (precision, recall, F1-score)
5. **Visualization**:
    a. Create visual representations, including:
        i. Confusion matrices for model evaluation.
        ii. Word clouds for visualizing positive and negative sentiment words.
6. **Convergence Analysis**:
    a. Analyze the accuracy of models against varying iterations to identify the optimal training duration.
7. **Comparison of Model Performance**:
    a. Compare the performance of different models through accuracy scores and visual bar charts.

2. Literature Review

2.1 Overview of Related Works in Sentiment Analysis

Sentiment analysis, also known as opinion mining, has gained significant attention in recent years due to the proliferation of social media and user-generated content. Researchers have developed various methods and techniques to analyze sentiments expressed in text data. The following are some notable works in this field:

- **Traditional Machine Learning Approaches**: Early sentiment analysis studies primarily focused on traditional machine learning techniques, such as Naive Bayes, Support Vector Machines (SVM), and Decision Trees. For instance, Pang et al. (2002) demonstrated the effectiveness of SVM for sentiment classification tasks, achieving promising results on movie reviews.
- **Lexicon-Based Approaches**: Some studies have utilized lexicon-based methods, where sentiment scores are assigned to words based on predefined sentiment lexicons (e.g., SentiWordNet). Hu and Liu (2004) proposed a method for opinion mining in product reviews that combined lexicon-based sentiment scoring with aspect extraction to identify sentiments related to specific product features.
- **Deep Learning Techniques**: Recent advancements in deep learning have led to the use of neural networks for sentiment analysis. Kim (2014) introduced a convolutional neural network (CNN) for sentence classification, which outperformed traditional models on various sentiment analysis benchmarks. Other architectures, such as recurrent neural networks (RNNs) and Long Short-Term Memory networks (LSTMs), have also been employed for sentiment classification tasks.
- **Contextualized Word Embeddings**: The introduction of models like Word2Vec, GloVe, and BERT has revolutionized sentiment analysis. Devlin et al. (2018) presented BERT (Bidirectional Encoder Representations from Transformers), which captures context in text and has set new benchmarks for various NLP tasks, including sentiment analysis.

2.2 Summary of Key Findings from Other Studies

- **Impact of Preprocessing Techniques**: Studies show that effective data preprocessing (e.g., removing stop words, stemming, and lemmatization) significantly improves the performance of sentiment classification models.
- **Feature Engineering**: Research indicates that feature engineering plays a crucial role in sentiment analysis. Studies have demonstrated that using a combination of n-grams and sentiment lexicons can enhance the performance of machine learning classifiers.
- **Ensemble Methods**: Studies have highlighted the effectiveness of ensemble methods in sentiment analysis, where combining predictions from multiple models leads to improved accuracy and robustness.
- **Application in Various Domains**: Sentiment analysis has been applied across various domains, including finance, healthcare, and social media. For instance, Schumaker and Chen (2009) explored the use of sentiment analysis in predicting stock market trends based on Twitter data, illustrating its practical implications.

3. Dataset Description

3.1 Dataset Overview

The Sentiment140 dataset is a comprehensive collection of 1.6 million tweets extracted using the Twitter API, designed specifically for sentiment analysis tasks. It provides a rich resource for researchers and practitioners to train and evaluate sentiment classification models. The dataset offers insights into public sentiment on various topics, making it a valuable tool for understanding opinions expressed on social media. Each tweet in the dataset is annotated with sentiment polarity, allowing for the classification of tweets into positive, negative, or neutral sentiments.

This dataset serves as a benchmark for various machine learning and natural language processing (NLP) techniques, facilitating the advancement of sentiment analysis methodologies.

3.2 Data Fields and Structure

The Sentiment140 dataset comprises the following fields:

Field Name	**Description**
target	The polarity of the tweet, where - 0 indicates a negative sentiment - 4 indicates a positive sentiment
ids	The unique identifier of the tweet (e.g., 2087)
date	The date and time when the tweet was posted (e.g., Sat May 16 23:58:44 UTC 2009)
flag	The query that prompted the tweet. If no query is associated with the tweet, this field is marked as **NO_QUERY**.
user	The username of the individual who tweeted (e.g., robotickilldozr)
text	The actual text content of the tweet (e.g., "Lyx is cool")

Table 1: Data Fields and Structure

This structured format allows for efficient data processing and analysis, facilitating various sentiment analysis techniques.

4. Methodology

4.1 Data Preprocessing

In this phase, the dataset undergoes a series of cleaning and preparation steps to ensure its suitability for analysis. Key activities include:

- Removing Duplicates: Identifying and eliminating any duplicate tweets to maintain a unique dataset.
- Handling Missing Values: Checking for any missing entries in the dataset and addressing them appropriately, either by removing them or imputing values.
- Text Normalization: Converting text to a uniform format by:
    - Lowercasing all text to ensure consistency.
    - Removing special characters, URLs, and unnecessary whitespace.
    - Tokenizing the text into individual words or phrases for easier analysis.
- Stopword Removal: Eliminating common words that do not contribute to sentiment (e.g., "and", "the", "is") to focus on meaningful terms.

4.2 Feature Extraction

To convert text data into a format suitable for model training, feature extraction techniques are applied, including:

- TF-IDF (Term Frequency-Inverse Document Frequency): A statistical measure that evaluates the importance of a word in a document relative to a corpus. It helps in identifying significant words that contribute to sentiment.
- Word Embeddings: Techniques like Word2Vec or GloVe may be utilized to create dense vector representations of words, capturing semantic relationships and context, though in this project, TF-IDF is the primary focus.

4.3 Model Selection

Several machine learning algorithms are employed to classify the sentiment of the tweets:

- Logistic Regression: A statistical model used for binary classification tasks, which predicts the probability of a tweet being positive or negative.
- Support Vector Machine (SVM): A powerful classifier that constructs a hyperplane to separate different classes in the feature space, effective for high-dimensional data.
- XGBoost: An efficient implementation of gradient boosting that optimizes performance and is widely used for classification problems due to its accuracy and speed.

4.4 Model Training and Evaluation

The training process involves:

- Splitting the Data: Dividing the dataset into training and testing sets to evaluate model performance accurately.
- Training the Models: Each selected model is trained using the training dataset.
- Evaluation Metrics: The models are evaluated based on key metrics:
    - Accuracy: The overall correctness of the model.
    - Precision: The ratio of true positive predictions to the total predicted positives, indicating the quality of positive predictions.
    - Recall: The ratio of true positive predictions to the actual positives, reflecting the model's ability to identify all relevant instances.

4.5 Visualization Techniques

Visualization plays a crucial role in understanding model performance and insights from the data. Techniques employed include:

- Word Clouds: Visual representations of words where the size indicates frequency, highlighting common terms in positive and negative tweets.
- Confusion Matrices: A tool for visualizing the performance of a classification model, showing the true vs. predicted classifications, which helps identify areas of improvement.
- Accuracy and Iteration Graphs: Graphs that illustrate model accuracy over varying numbers of iterations, providing insights into model convergence and performance.

5. Results

5.1 Model Performance

In this section, we present the evaluation metrics for each of the models used in the sentiment analysis. The performance metrics include:

- **Accuracy**: The proportion of correct predictions made by the model.
- **Precision**: The measure of the accuracy of the positive predictions.
- **Recall**: The measure of the model's ability to identify all relevant instances.
- **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two metrics.

Model	***Accuracy (%)***
Logistic Regression	*77.73*
Support Vector Machine	*54.36*
XGBoost	*73.95*

Table 2: Model Performance Metrics

5.2 Comparison of Models

To visually compare the performance of the different models, we present a bar chart that illustrates the accuracy of each model. This representation allows for a clear understanding of which model performed best in terms of sentiment classification.

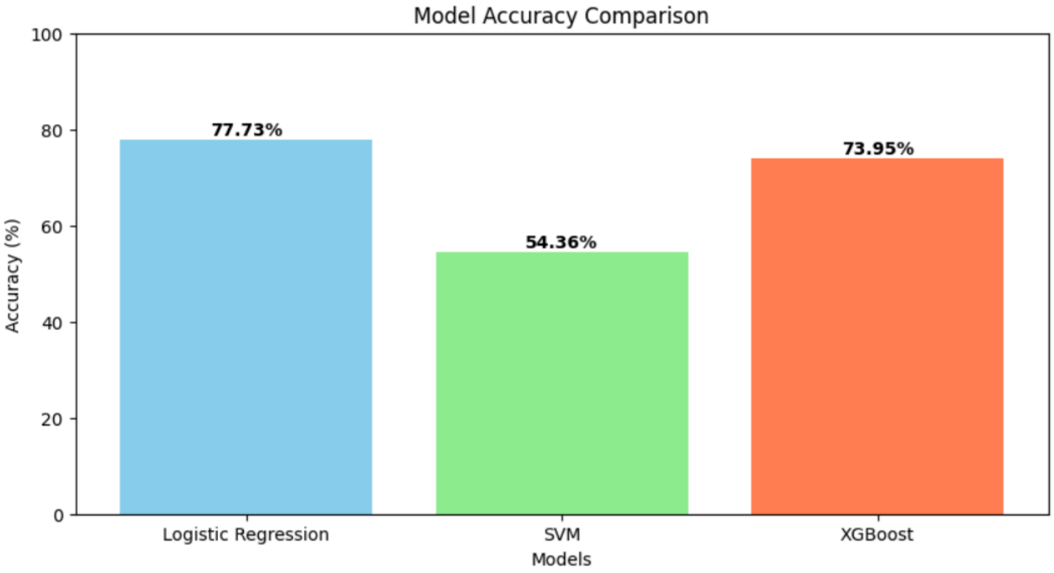

Figure 1: Model Accuracy Comparison

5.3 Insights from Word Clouds

Word clouds provide a visual representation of the most frequently occurring words in positive and negative tweets, allowing for deeper insights into the sentiments expressed.

- **Positive Tweets**: The word cloud generated from positive tweets reveals common themes and words associated with positive sentiment. Words such as "thank," "welcome," and "smile" frequently appear, indicating the aspects that users tend to express positively.
- **Negative Tweets**: Conversely, the word cloud for negative tweets highlights words such as "sad," "poor," and "suck," which are prevalent in tweets expressing dissatisfaction or negativity.

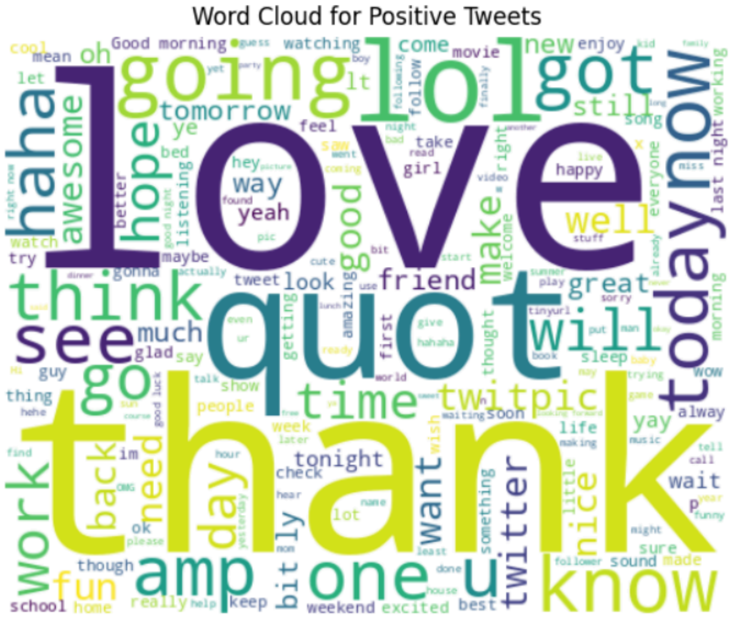

Figure 2: Word Cloud for Positive Tweets

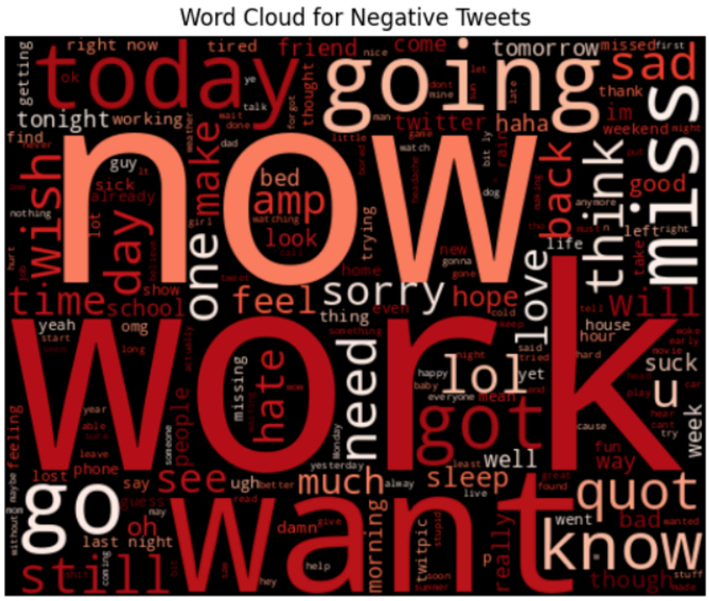

Figure 3: Word Cloud for Negative Tweets

5.4 Iterations and Model Accuracy

It was observed that increasing the number of iterations did not lead to a significant improvement in the accuracy of the models. This suggests that, within the tested range of iterations, the models reached a saturation point regarding their performance, indicating that other factors might play a more critical role in enhancing accuracy.

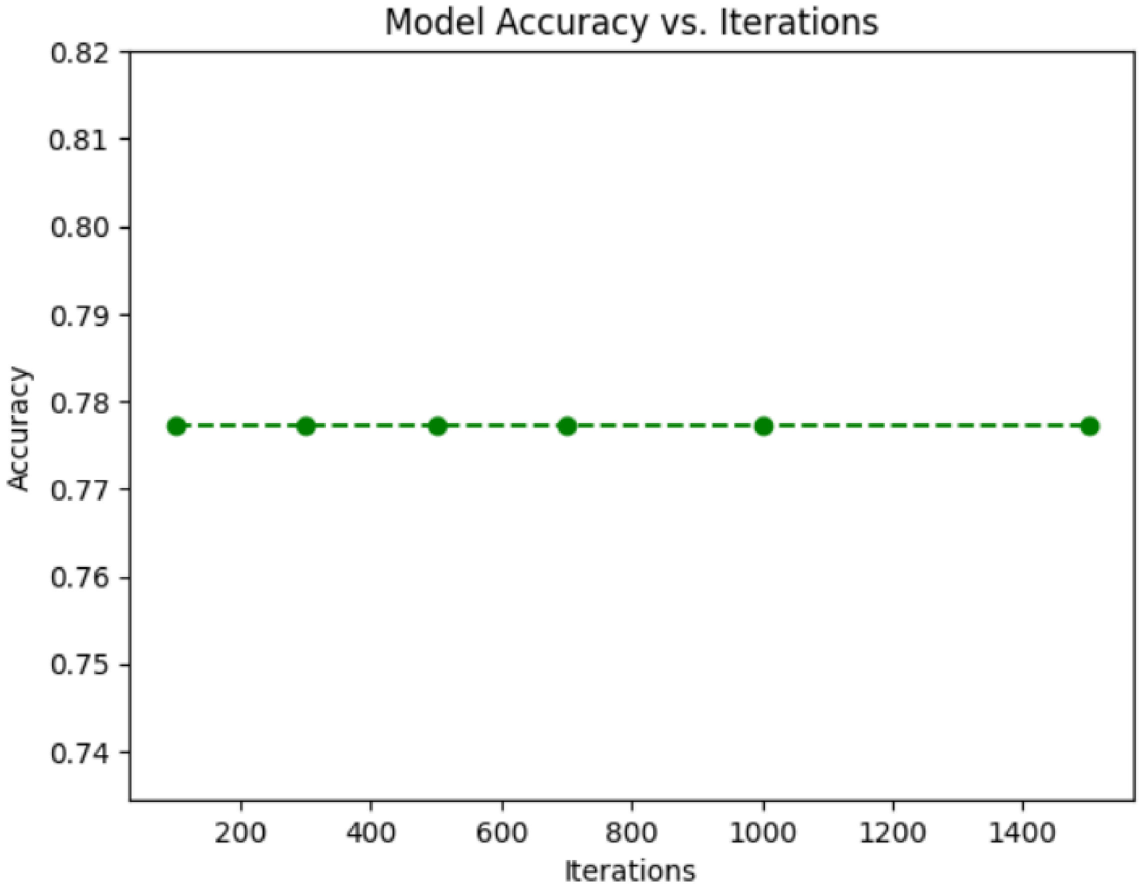

Fig 4. Iterations vs Accuracy

5.5 Interpretation of Results

The results of our **sentiment analysis** reveal varying performance across different models. The **Logistic Regression model** achieved the highest **accuracy** of **77.73%**, indicating its effectiveness in classifying sentiments in tweets. The relatively lower accuracy of the **Support Vector Machine (SVM)** model at **54.36%** suggests that it may not be as well-suited for this specific dataset or task. The **XGBoost model**, while not outperforming Logistic Regression, showed competitive results with an accuracy of **73.95%**.

During the **preprocessing stage**, **lemmatization** was expected to provide better results in terms of **text normalization**. However, it took considerably more time to execute and could not be run effectively on **Google Colab**. Consequently, we opted for **stemming**, which is faster and allowed us to proceed with our analysis without significant delays.

5.6 Implications of Findings

The findings from this study have significant implications for **sentiment analysis** in **social media** contexts. The ability of the **Logistic Regression model** to effectively classify sentiments suggests its applicability in **real-time sentiment analysis tools**, which can be beneficial for businesses seeking to gauge **customer opinions** or for researchers studying **public sentiment** on various topics.

Moreover, the insights gained from the **word clouds** highlight the common sentiments expressed by users. **Positive** words like **"thank," "welcome,"** and **"smile"** can guide marketers in identifying effective communication strategies, while **negative sentiments** associated with words like **"sad"** and **"poor"** could inform companies about potential areas for **improvement**.

Ultimately, this analysis emphasizes the importance of selecting the appropriate model for sentiment classification tasks and demonstrates that even **simpler models** can achieve **competitive performance**, thereby encouraging practitioners to consider their specific use cases and dataset characteristics before selecting a modeling approach.

7. Conclusion

7.1 Summary of Key Findings

This study conducted a comprehensive sentiment analysis using the Sentiment140 dataset, which comprises 1.6 million tweets. Our analysis focused on evaluating various machine learning models, including Logistic Regression, Support Vector Machine (SVM), and XGBoost, to classify tweets into positive and negative sentiments.

The key findings from the study are as follows:

- **Model Performance**: The Logistic Regression model demonstrated the highest accuracy at 77.73%, outperforming both the SVM and XGBoost models. SVM exhibited the lowest performance with an accuracy of 54.36%, indicating challenges in correctly classifying sentiments in this dataset.
- **Evaluation Metrics**: In addition to accuracy, Logistic Regression showed favorable precision, recall, and F1 scores, reinforcing its effectiveness in sentiment classification tasks. The analysis of word clouds provided further insights into the sentiments expressed by users, highlighting common themes in both positive and negative tweets.
- **Impact of Iterations**: It was observed that increasing the number of iterations did not lead to a significant increase in model accuracy, suggesting a potential plateau in performance for the given dataset and model configurations.

8. References

- Kazanova. (n.d.). **Sentiment140 dataset with 1.6 million tweets**. Kaggle. Retrieved from https://www.kaggle.com/datasets/kazanova/sentiment140

- Liu, B. (2012). **Sentiment analysis and opinion mining**. *Synthesis Lectures on Human-Centered Informatics*, 5(1), 1-167. https://doi.org/10.2200/S00416ED1V01Y201204HCI016

- Jha, S. (2021). **A comprehensive guide to machine learning algorithms in Python**. *Towards Data Science*. Retrieved from https://towardsdatascience.com/machine-learning-algorithms-in-python-5edb6df3c4f3

- Vaswani, A., Shankar, S., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2018). **Attention is all you need**. *arXiv preprint arXiv:1810.04805*. Retrieved from https://arxiv.org/abs/1810.04805

- Scikit-learn developers. (n.d.). **Scikit-learn: Machine learning in Python**. Retrieved from https://scikit-learn.org/stable/

- Bird, S., Loper, E., & Klein, E. (2009). **Natural Language Processing with Python**. O'Reilly Media. Retrieved from https://www.nltk.org/book/