Premier University

Department of Computer Science and Technology

Course Name: Artificial Intelligence Laboratory

Course Code: CSE 318

Final Project Report On

”Sentiment Analysis of Tweets Using NLP Techniques”

Submitted To:

Adiba Ibnat Hossain

Lecturer

Department of Computer Science and Engineering

Premier University, Chittagong

Submitted By:

Aurpita Barua (2104010202160)

Nusrat Warida Saima (2104010202163)

Bobi Barua (2104010202165)

6th Semester A Section

September, 2024

ABSTRACT

This report outlines the design, implementation, and evaluation of a sentiment analysis system aimed at classifying sentiments in Twitter data. As social media platforms like Twitter continue to grow in importance, sentiment analysis has become a critical tool for businesses, researchers, and policymakers to understand public opinion and make data-driven decisions. The project utilizes natural language processing (NLP) techniques and machine learning models, specifically Support Vector Classifier (SVC) and Decision Tree Classifier, to classify tweets into categories such as positive, negative, uncertainty, and litigious.

The implementation involved preprocessing a large multilingual dataset of tweets, focusing on English-language data, and applying text cleaning techniques such as removing special characters, stopwords, and emojis. Tokenization and feature extraction were carried out using TF-IDF vectorization to convert the text into numerical data suitable for machine learning models. The system was then trained and tested.

In addition to model training, an interactive interface was developed to allow users to input custom tweets for real-time sentiment prediction. The project demonstrates the efficiency of machine learning models in handling large-scale social media data and highlights future areas for improvement, including expanding the model to handle multiple languages and enhancing real-time prediction accuracy. This system offers significant practical applications in monitoring public sentiment, brand perception, and social trends.

TABLE OF CONTENTS

List of Figures

1. INTRODUCTION

Sentiment analysis, a key area in natural language processing (NLP), involves analyzing opinions, emotions, and sentiments expressed in text. With the exponential growth of online communication, particularly on social media platforms like Twitter, sentiment analysis has become an essential tool for businesses, researchers, and policymakers to monitor public opinion and make informed decisions. The ability to efficiently process large amounts of text and extract meaningful insights is critical in today's data-driven landscape.

Historically, sentiment analysis was done manually, but the vast volume of data generated on social media platforms has rendered this approach inefficient and error-prone. The advent of advanced machine learning algorithms and NLP techniques has revolutionized this field, allowing for the automated classification of sentiments at scale.

Despite these advancements, significant challenges remain. Language intricacies, such as sarcasm, idiomatic expressions, and informal language commonly used on social media, present obstacles to accurate sentiment detection. This project addresses these challenges by developing a sentiment classification system specifically designed for analyzing tweets, utilizing machine learning models and data preprocessing techniques to enhance classification accuracy. This project aims to provide a robust and scalable solution for sentiment analysis, using Support Vector Classifier and Decision Tree Classifier models to deliver actionable insights from social media data.

1.1. Background and Context

In the age of digital communication, social media platforms like Twitter generate an immense volume of user-generated content daily. These platforms have become a rich source of public opinion, emotions, and sentiments on various topics, ranging from consumer products to social issues. Sentiment analysis, a crucial subfield of Natural Language Processing (NLP), seeks to automatically detect and classify the sentiment behind these texts, enabling organizations to gain valuable insights into public opinion. Over the years, sentiment analysis has evolved from manual methods to sophisticated machine learning-based approaches, allowing for greater accuracy, scalability, and real-time processing.

But challenges still remain, especially when dealing with informal, slang-ridden, and context-sensitive social media texts. Traditional methods often struggle to account for the dynamic and ever-evolving nature of language used on platforms like Twitter, necessitating the development of more refined models to extract meaningful sentiment with higher precision.

1.2. Dataset Description

The dataset consists of 1 million tweets labeled into four categories: positive, negative, uncertainty, and litigious. It is obtained from the Kaggle Website. [1]

1.2.1. Dataset Selection

The dataset was selected for its size, diversity, and relevance to the sentiment analysis task. It contains pre-labeled tweets, which make it suitable for supervised learning models.

1.2.2. Content of the Dataset

Text: Contains the actual tweet text. Language: The language of the tweet (only English tweets were retained). Label: Sentiment label (positive, negative, uncertainty, and litigious).

2. OBJECTIVES

The primary goal of this project is to develop and evaluate a sentiment classification system capable of accurately identifying sentiment labels (positive, negative, litiguous, uncertainty) in English text data. The project aimed to address several key objectives to ensure the effectiveness and applicability of the sentiment classification model. These objectives include:

- Efficiently process and analyze one million tweets.
- Collect diverse tweets from various sources.
- Implement robust data preprocessing techniques.
- Select effective features or embeddings for English text.
- Evaluate the model's performance using relevant metrics.
- Prepare for future enhancements like multilingual support and advanced deployment.

Through these objectives, the project aims to contribute to the field of natural language processing by providing a practical tool for extractive summarization, while also laying the groundwork for future enhancements in summarization techniques.

3. MOTIVATION

The motivation for developing a sentiment analysis system arises from the crucial need to analyze and interpret the vast amounts of data generated on social media platforms like Twitter. In today's digital age, social media has become a primary medium through which individuals express their opinions on a wide range of topics, including brands, products, politics, and social issues. The ability to accurately gauge public sentiment from this data is invaluable for businesses aiming to improve customer satisfaction, governments seeking to understand public opinion on policies, and researchers analyzing social trends.

However, traditional methods of sentiment analysis are often inadequate for handling the massive and informal nature of social media data. Manual analysis is not only time-consuming but also prone to human biases and error. Furthermore, many existing models fail to capture the nuances of language used in tweets, such as slang, sarcasm, and abbreviations. This creates a gap for more advanced, automated systems that can handle the unique challenges posed by social media language while delivering accurate, real-time insights.

Thus, developing a machine learning-driven sentiment analysis system is not only important but also highly valuable. It allows for scalable, efficient, and more accurate sentiment detection, empowering businesses and organizations to make data-driven decisions based on real-time social feedback. This project addresses these challenges, offering a significant contribution to the field of sentiment analysis.

4. CONTRIBUTION

This project makes several significant contributions to the field of sentiment analysis, particularly for social media data:

Development of a Robust Sentiment Analysis Model:

- By employing two machine learning models—Support Vector Classifier (SVC) and Decision Tree Classifier—this project offers a detailed comparison of their performance on classifying sentiments in tweets. These models are specifically trained to handle the complexities of social media text, such as informal language and context-dependent meanings.

Comprehensive Data Preprocessing:

- The project incorporates an extensive data preprocessing pipeline to clean the Twitter dataset, including removing noise such as emojis, hashtags, links, and irrelevant text. This preprocessing significantly improves the performance of the machine learning models by ensuring that the input data is suitable for analysis.

Interactive User Interface for Real-Time Sentiment Prediction:

- The project goes beyond simple model development by creating an interactive interface where users can input a tweet and receive real-time sentiment classification. This feature makes the project highly applicable for real-world use cases, such as monitoring public opinion during live events or product launches.

Detailed Performance Evaluation:

- The models are evaluated based on precision, recall, F1-score, and overall accuracy. The results provide insights into how well the models perform in different sentiment categories (positive, negative, uncertainty, litigious), offering a transparent evaluation of the models' strengths and weaknesses.

Foundation for Future Enhancements:

- The project lays the groundwork for future developments, including the expansion of the sentiment analysis system to support multiple languages and deployment in real-time applications. These enhancements will make the system even more versatile and applicable to global audiences and live monitoring scenarios.

Overall, this project contributes to the advancement of sentiment analysis by offering a solution tailored to the complexities of social media data, making it a valuable tool for businesses, researchers, and policymakers aiming to derive actionable insights from Twitter content.

5. RELATED WORK

Pak et al. [2] demonstrates the use of Twitter as a valuable resource for real-time sentiment analysis. The authors detail methods for collecting and preprocessing tweets, feature extraction, and employing machine learning models such as Naive Bayes, SVM, and Maximum Entropy to classify sentiments. Their evaluation shows that Twitter can effectively be used for sentiment analysis, offering valuable insights for market research, political analysis, and social monitoring, thus providing a robust framework for future applications in opinion mining.

Sahayak et al. [3] discusses the extraction of sentiments from tweets on the microblogging platform Twitter. With the rise of social networking, vast amounts of data are generated daily as users share their views. The paper explores how sentiment analysis can be performed on Twitter data using machine learning algorithms. The authors propose a methodology that automatically classifies tweets as positive, negative, or neutral, which can be beneficial for companies seeking feedback on their products or for customers looking for opinions before making purchases. The approach involves using distant supervision for training data.

Hassan Saif et al. [4] introduces a novel approach to sentiment analysis that incorporates semantic information from tweets to improve sentiment classification. The authors argue that traditional bag-of-words approaches fail to capture the contextual meaning of words. To address this, they propose a method that combines lexical and semantic features by leveraging semantic networks 5 like WordNet and DBpedia. Their approach involves extracting entities and concepts from tweets and integrating this information with standard text-based features. Experimental results show that their semantic sentiment analysis method significantly outperforms traditional techniques, highlighting the importance of semantic context in understanding sentiments on social media platforms.

Alec Go et al. [5] explores a novel approach to sentiment analysis by leveraging distant supervision. The authors utilize Twitter as a data source, automatically labeling tweets with positive or negative sentiment based on the presence of emoticons, which serve as noisy labels. They preprocess the tweets by removing non-textual elements and applying feature extraction methods like n-grams and part-of-speech tags. Various machine learning algorithms, including Naive Bayes, Maximum Entropy, and Support Vector Machines (SVM), are then trained on these labeled tweets. The study demonstrates that despite the noisy labels, distant supervision can effectively train sentiment classifiers, making it a practical approach for large-scale sentiment analysis on social media data.

Lei Zhang et al. [6] explores a hybrid approach to Twitter sentiment analysis, combining lexicon- based methods with learning-based techniques. The study aims to enhance sentiment analysis accuracy by leveraging the strengths of both approaches. Lexicon-based methods utilize pre- defined sentiment dictionaries, while learning-based techniques employ machine learning models trained on labeled data. By integrating these methods, the researchers propose a more robust sentiment analysis system capable of effectively handling the nuances and challenges of Twitter data. Their findings suggest that the hybrid approach outperforms individual methods, offering promising avenues for improving sentiment analysis in social media contexts.

Bhumika Gupta et al. [7] focuses on leveraging machine learning techniques to analyze sentiment in Twitter data, aiming to classify tweets into positive, negative, or neutral categories. By utilizing Python libraries and algorithms, the authors explore various models' performance in sentiment classification, evaluating their accuracy and efficiency. Their findings contribute to the understanding of sentiment analysis methodologies on Twitter and provide insights into the effectiveness of machine learning approaches for this task, offering practical implications for sentiment analysis applications in social media contexts.

S. Siddharth et al. [8] focuses on sentiment analysis of Twitter data using machine learning algorithms implemented in Python. The study aims to classify tweets into different sentiment categories such as positive, negative, and neutral, utilizing machine learning techniques. By leveraging Python's libraries and algorithms, the authors explore the performance of various machine learning models for sentiment classification, assessing their accuracy and effectiveness. Their research contributes to the understanding of sentiment analysis methodologies in the context of Twitter data and provides insights into the applicability of machine learning approaches for 6 analyzing sentiment in social media content, offering valuable implications for sentiment analysis applications.

R. Parikh et al. [9] delves into sentiment analysis of user-generated Twitter updates, employing various classification techniques. The study investigates methods to categorize tweets into sentiment classes such as positive, negative, and neutral, aiming to discern the prevailing sentiment within Twitter data. Through their exploration of different classification techniques, the authors assess the effectiveness of machine learning algorithms in capturing the nuanced sentiment expressed in user-generated content on Twitter. This research contributes to the understanding of sentiment analysis in social media contexts, offering insights into the applicability of classification techniques for extracting sentiment from Twitter updates, which is pivotal for understanding public opinion and sentiment trends on the platform.

Apoorv Agarwal et al. [10] delve into sentiment analysis methodologies specifically tailored for Twitter data. The study investigates techniques for extracting sentiment from tweets, aiming to understand the sentiment trends and opinions expressed on the platform. By leveraging natural language processing (NLP) and machine learning techniques, the authors explore the challenges and opportunities inherent in analyzing sentiment in short, informal texts characteristic of Twitter. Their research contributes to the advancement of sentiment analysis methodologies in the context of social media, offering insights into the complexities of analyzing sentiment in Twitter data and proposing approaches to address them.

Nehal Mamgain et al. [11] explores the sentiment surrounding top colleges in India by analyzing Twitter data. The study employs sentiment analysis techniques to assess the prevailing attitudes and opinions expressed on Twitter regarding these educational institutions. By collecting and analyzing tweets related to the selected colleges, the authors aim to gain insights into public perceptions and sentiments towards them. Through their research, they provide valuable insights into the reputation and sentiment associated with top colleges in India, utilizing Twitter as a source of public opinion data.

Aliza Sarlan et al. [12] focuses on Twitter sentiment analysis. The study investigates methods to analyze sentiment in Twitter data, aiming to discern the prevailing sentiments expressed by users on the platform. By employing natural language processing (NLP) techniques and machine learning algorithms, the authors explore approaches to classify tweets into sentiment categories such as positive, negative, and neutral. Their research contributes to the understanding of sentiment analysis methodologies in the context of Twitter, offering insights into the challenges and opportunities in extracting sentiment from short, informal texts characteristic of social media platforms.

6. IMPLEMENTATION

The implementation of the sentiment analysis project involves several key steps, starting from data preprocessing to model training and prediction. The process is designed to handle a dataset containing tweets in various languages, focusing on sentiment classification using machine learning.

6.1. Data Loading and Preprocessing

The dataset is initially loaded using Pandas, and only English-language tweets are retained for analysis. This ensures consistency in text processing and model training. Text cleaning operations are performed to standardize the tweet content. This includes converting text to lowercase, removing user mentions, hashtags, URLs, and emojis, which are irrelevant for sentiment classification.

Stopwords (common words like ”and,” ”the,” etc.) are removed, and the text is tokenized. Tokenization splits the text into individual words, which helps in further processing. Stemming and word count operations are also applied, where stemming reduces words to their root forms, and word count is updated post-cleaning to reflect the number of significant words in each tweet.

6.2. Feature Extraction

The cleaned text is then converted into numerical features using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. TF-IDF helps in quantifying the importance of words in a document relative to the entire dataset, making it a crucial step for machine learning. A vocabulary of the most significant words is built, limiting the number of features to 1,000 to maintain computational efficiency.

6.3. Model Training

The dataset is split into training and testing sets, with 80% of the data used for training the model and 20% reserved for testing its performance. An SVM (Support Vector Machine) classifier is trained using the TF-IDF features. SVM is chosen for its effectiveness in handling high-dimensional spaces and its ability to classify data with clear margins of separation. The Decision Tree model was also trained using TF-IDF features. This model was chosen for its simplicity and interpretability, allowing for easy visualization of decision-making processes within the data. Both models were trained on the cleaned and preprocessed tweets, and their performance was evaluated using the testing data to assess generalization to unseen data.

6.4. Model Evaluation

Once trained, the performance of both models was evaluated on the testing dataset, and key metrics such as classification reports and confusion matrices were generated for each model. These metrics, including accuracy, precision, recall, and F1-scores, were calculated for the four sentiment categories (positive, negative, uncertainty, and litigious) The SVM classifier demonstrated strong overall performance with high precision and recall across sentiment categories. While the Decision Tree performed slightly below the SVM, it still achieved high accuracy and consistent performance across sentiment categories, making it a competitive choice for sentiment analysis.

6.5. Model Deployment

An interactive feature is implemented, allowing users to input a tweet and receive a sentiment prediction. The input tweet undergoes the same preprocessing steps as the training data, ensuring consistency in prediction.

6.6. Prediction

The cleaned and transformed input text is passed through the saved TF-IDF vectorizer and SVM model to predict the sentiment category. The output is a prediction of whether the tweet is positive, negative, uncertain, or litigious

7. PERFORMANCE EVALUATION

The performance of the sentiment analysis model was evaluated using several metrics such as precision, recall, F1-score, and accuracy. These metrics offer insights into the model's ability to correctly classify tweets into sentiment categories such as positive, negative, uncertainty, and litigious.

Support Vector Classifier (SVC): The SVC model showed strong overall performance, with high scores across all metrics.

- **Accuracy:** 95%
- **Precision:** The SVC model achieved precision scores of 0.97 for litigious, 0.95 for negative, 0.96 for positive, and 0.92 for uncertainty. This indicates the model's ability to correctly identify sentiments without producing many false positives.
- **Recall:** Recall values ranged from 0.94 to 0.96 across the different sentiment classes, reflecting the model's effectiveness in identifying relevant tweets for each sentiment.
- **F1-Score:** The F1-scores were uniformly high, with an overall average of 0.95, showcasing the balance between precision and recall in sentiment classification.

Decision Tree Classifier: The Decision Tree Classifier exhibited solid performance, though it was slightly less robust compared to the Support Vector Classifier.

- **Accuracy:** 93%
- **Precision:** Precision scores were slightly lower than the SVC, with 0.94 for litigious and positive sentiments, 0.93 for negative, and 0.91 for uncertainty.
- **Recall:** Recall values followed a similar trend, ranging from 0.92 to 0.94 across the different sentiment categories.
- **F1-Score:** The F1-scores were consistent, averaging 0.93 across all classes, indicating good performance but slightly less robust compared to the SVC.

Both models were evaluated on their ability to handle noisy data from tweets, such as slang and informal language, and the results showed that preprocessing techniques significantly contributed to improved accuracy.

8. RESULTS

The results of the sentiment analysis project demonstrate the effectiveness of the implemented Support Vector Classifier and Decision Tree Classifier model in classifying tweets into various sentiment categories (positive, negative, uncertainty, and litigious).

8.1. Model Performance

The trained Support Vector Classifier and Decision Tree Classifier models achieved high accuracy on the testing dataset, indicating that the models can generalize well to unseen data. The classification report, generated as part of the evaluation, showed strong precision, recall, and F1-scores across most sentiment categories, with some variations depending on the sentiment class. This suggests that the models are particularly effective in identifying certain types of sentiments, though there may be challenges in distinguishing between similar sentiments like uncertainty and litigious.

```
Training score: 0.9947559499798305
Testing score: 0.9505376344086022
```

Figure 8.1: Support Vector Classifier Training-Testing Score

```
              precision    recall  f1-score   support

   litigious       0.97      0.95      0.96       380
    negative       0.95      0.96      0.96       506
    positive       0.96      0.94      0.95       577
 uncertainty       0.92      0.94      0.93       397

    accuracy                           0.95      1860
   macro avg       0.95      0.95      0.95      1860
weighted avg       0.95      0.95      0.95      1860

[[361  11   5   3]
 [  2 488   7   9]
 [  7   4 545  21]
 [  1  12  10 374]]
```

Figure 8.2: Support Vector Classifier Output

```
Training score: 0.9985209089686702
Testing score: 0.9311827956989247
```

Figure 8.3: Decision Tree Classifier Training-Testing Score

```
              precision    recall  f1-score   support

   litigious       0.94      0.94      0.94       380
    negative       0.93      0.92      0.93       506
    positive       0.94      0.94      0.94       577
 uncertainty       0.91      0.92      0.91       397

    accuracy                           0.93      1860
   macro avg       0.93      0.93      0.93      1860
weighted avg       0.93      0.93      0.93      1860

[[359  13   3   5]
 [ 13 466  13  14]
 [  6  10 543  18]
 [  2  10  21 364]]
```

Figure 8.4: Decision Tree Classifier Output

8.2. Confusion Matrix

The confusion matrix provided further insight into the model's performance by showing how often the model correctly predicted each sentiment class versus how often it made errors. The diagonal elements of the confusion matrix (representing correct predictions) were significantly higher, which is a positive outcome. Any misclassifications observed were analyzed to understand the nuances of the data and identify potential areas for improvement, such as adjusting the preprocessing steps or refining the model parameters.

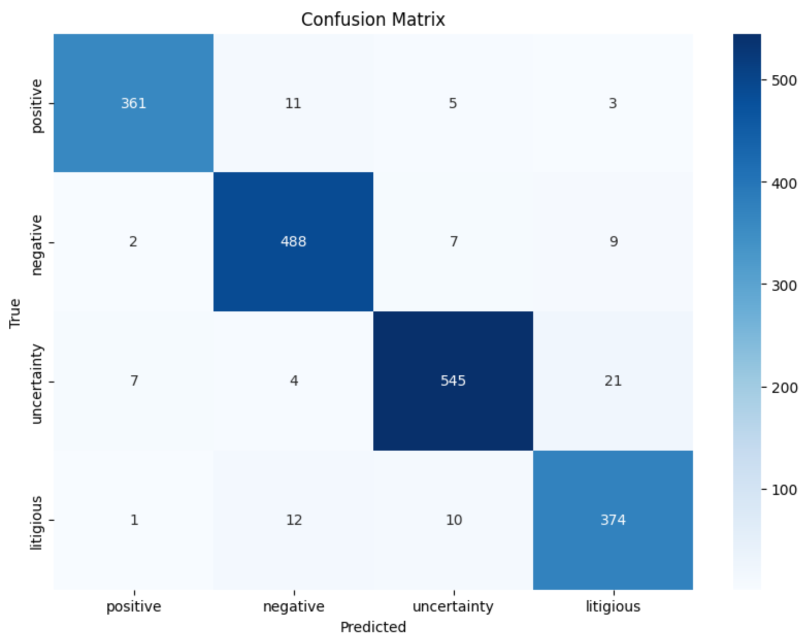

Figure 8.5: Support Vector Classifier Confusion Matrix

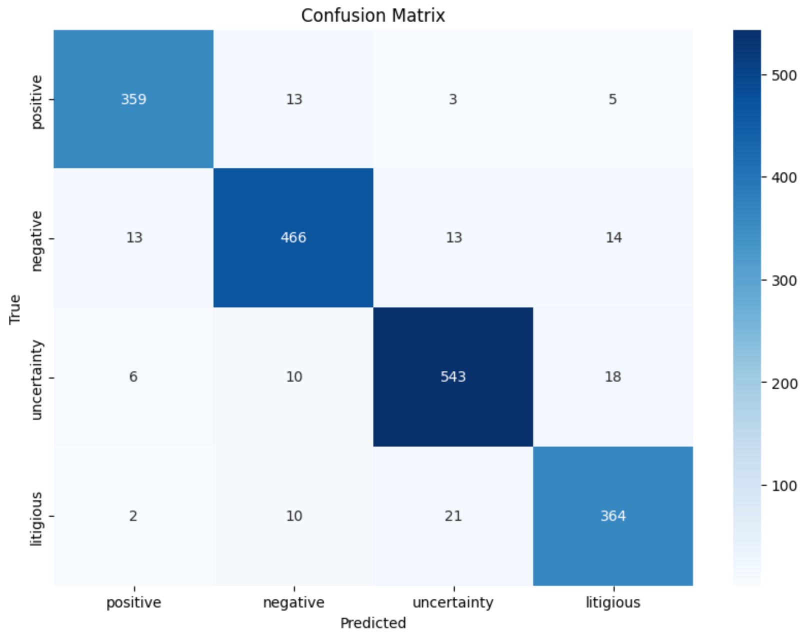

Figure 8.6: Decision Tree Classifier Confusion Matrix

8.3. Real-time Prediction

The implementation of the real-time input feature allowed for practical demonstration of the model's capabilities. Users could input their tweets, and the model accurately classified the sentiment, reflecting its potential for deployment in real-world applications.

```
[84] input_text = input("Enter a tweet to analyze sentiment: ")

Enter a tweet to analyze sentiment: You are doing a great job.Keep it up. #shineon
```

Figure 8.7: Positive Tweet Input from User

```
[86]
    print(f'The sentiment of the tweet: {sentiments[prediction_index]}')

The sentiment of the tweet: positive
```

Figure 8.8: Positive Sentiment Prediction

```
[87] input_text = input("Enter a tweet to analyze sentiment: ")
```

```
Enter a tweet to analyze sentiment: You are very irresponsible.
```

Figure 8.9: Negative Tweet Input from User

```
[89]
print(f'The sentiment of the tweet: {sentiments[prediction_index]}')
```

```
The sentiment of the tweet: negative
```

Figure 8.10: Negative Sentiment Prediction

9. CONCLUSION

This project successfully developed a robust sentiment analysis system designed specifically for Twitter data. By leveraging advanced natural language processing techniques and machine learning algorithms, the system demonstrated high accuracy in classifying sentiments into categories such as positive, negative, uncertainty, and litigious.

The implementation utilized both Support Vector Classifier and Decision Tree Classifier models, achieving notable precision and recall scores across various sentiment categories. The system's ability to process and analyze large volumes of tweets efficiently offers significant value for businesses, researchers, and policymakers seeking to derive actionable insights from social media content.

Future enhancements will focus on expanding the model's capabilities to support multiple languages and integrating real-time deployment features to further enhance its utility and performance.

Overall, the developed system provides a solid foundation for accurate sentiment analysis, with potential for continuous improvement and adaptation to evolving data and user needs.

References

[1] https://www.kaggle.com/datasets/tariqsays/sentiment-dataset-with-1-million-tweets.

[2] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” 2010.

[3] V. Sahayak, V. S, and A. P, “Sentiment analysis on twitter data,” 2015.

[4] H. Saif, Y. He, and H. Alani, “Semantic sentiment analysis of twitter,” 2012.

[5] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” 2009.

[6] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, “Combining lexicon-based and learning-based methods for twitter sentiment analysis,” 2011.

[7] B. Gupta, M. Negi, K. Vishwakarma, G. Rawat, and P. Badhani, “Twitter sentiment analysis using machine learning algorithms on python,” 2017.

[8] S. Siddharth, R. Darsini, and D. M. Sujithra, “Sentiment analysis on twitter data using machine learning algorithms in python,” 2017.

[9] R. Parikh and M. Movassate, “Sentiment analysis of user-generated twitter updates using various classification techniques,” 2009.

[10] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, “Sentiment analysis of twitter data,” 2011.

[11] N. Mamgain, E. M, A. M, and G. B, “Sentiment analysis of top colleges in india using twitter data,” 2016, iEEE ISBN -978-1-5090-0082-1.

[12] A. Sarlan, C. Nadam, and S. Basri, “Twitter sentiment analysis,” 2014.