Vibe Check: Sentiment Analysis of Tweets

Jeff Rouzel Bat-og[1], Zyrex Djewel Ganit[1], and Rainer Mayagma[1]

University of the Philippines Visayas, Miagao, Iloilo
{jabatog, zfganit, rtmayagma}@up.edu.ph

Abstract. This proposal outlines our project, focusing on sentiment analysis of tweets. We aim to leverage machine learning models to classify sentiments as neutral, positive, or negative using a dataset of tweets. The insights gained will provide valuable information about public sentiment on various issues.

Keywords: Sentiment analysis, Twitter, Machine learning, Machine learning models, Natural language processing

1 Introduction

The topic we want to study is sentiment analysis of tweets. Social media platforms, particularly Twitter, are sources of user-generated content that reflect real-time public opinions. This study is relevant as it helps understand public sentiment on various topics and trends, which is essential for businesses, policymakers, and researchers. By analyzing sentiments, we can address valuable problems such as gauging public reaction to events, products, or policies, thus aiding in informed decision-making.

This project builds upon existing research in the field of natural language processing (NLP) and machine learning. Prior studies have employed various techniques to analyze sentiments, including traditional machine learning models and more recent deep learning approaches. By implementing and comparing multiple models, we aim to contribute to the understanding of which methods are most effective for sentiment classification in tweets.

2 Related Literature

The field of sentiment analysis has garnered significant attention in recent years. Research by Pang and Lee (2008) highlights the efficacy of different machine learning algorithms in text classification tasks. Logistic regression and naive Bayes have been widely adopted due to their simplicity and effectiveness [1].

A key challenge in sentiment analysis is handling negation, which can significantly alter the meaning of a sentence. Traditional models often struggle to capture negation effectively, leading to biases and misclassification of sentiments. Studies have demonstrated that inappropriate processing of negations

can adversely affect sentiment polarity detection [2]. Furthermore, Kaddoura et al. (2021) emphasize that in dialectal Arabic, the presence of negation can drastically change the polarity of opinionated words, complicating sentiment analysis in social media contexts. Their findings indicate that treating negation improves classification accuracy, highlighting its importance in sentiment analysis tasks [3].

3 Proposed Method

We will utilize supervised learning techniques to classify sentiments based on the labeled dataset. To further improve model performance, we will implement a negation handling mechanism using the negspacy package. Negspacy allows the model to capture the effect of negation in sentences, which is crucial for sentiment analysis, as phrases like "not good" should be classified differently from "good."

The models we plan to implement are summarized in Table 1.

Model	**Description**
Logistic Regression	A baseline model known for its efficiency in binary and multi-class classification tasks.
Naive Bayes	A probabilistic model that leverages the independence assumption among features.
Decision Trees	A non-parametric model that provides interpretability in decision-making processes.
Random Forests	An ensemble method that improves classification accuracy by aggregating results from multiple decision trees.

Table 1. Overview of the proposed methods for sentiment classification.

4 Dataset

The dataset for our analysis can be found on Kaggle, containing 27,481 tweets with several columns, as shown in Table 2. The dataset link can be accessed here.

Column Name	Description
TextID	Unique identifier for each tweet.
Text	The content of the tweet.
Selected Text	A highlighted portion of the tweet relevant to sentiment analysis.
Sentiment	The labeled sentiment (neutral, positive, or negative).
Time of Tweet	The timestamp indicating when the tweet was posted.
Age of User	The age of the user who posted the tweet.
Country	The country of the user.
Population (2020)	The population of the country as of 2020.
Land Area (Km)	The total land area of the country in square kilometers.
Density (P/Km)	Population density of the country (people per square kilometer).

Table 2. Overview of the dataset columns.

5 Metrics for Evaluation

We will evaluate our models using the metrics summarized in Table 3.

Metric	Description
F1 Score	The harmonic mean of precision and recall, providing a balance between the two metrics.
Recall	The ratio of true positive predictions to the total actual positives, measuring the model's ability to identify all relevant instances.
Precision	The ratio of true positive predictions to the total predicted positives, assessing the model's accuracy in its positive predictions.
Accuracy	The overall ratio of correct predictions to the total instances.

Table 3. Overview of evaluation metrics for model performance.

These metrics are crucial for understanding model performance in multi-class classification problems and will guide our model selection process.

6 Tools and Packages

The project will be implemented using Python, with the following libraries:

Library	Description
Pandas	For data manipulation and analysis.
Scikit-learn	For implementing machine learning algorithms and model evaluation.
Matplotlib/Seaborn	For data visualization and presenting results.
NLTK/Spacy	For natural language processing tasks, including text preprocessing and feature extraction.

Table 4. Overview of tools and packages used in the project.

References

1. Pang, B., & Lee, L.: A sentimental education: Sentiment analysis using machine learning techniques. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 811-818 (2008).
2. Mukherjee, P., Badra, Y., Doppalapudi, S. M., Srinivasan, S. M., Sangwan, R. S., & Sharma, R.: Effect of Negation in Sentences on Sentiment Analysis and Polarity Detection. *The Pennsylvania State University, Great Valley*, (2023).
3. Kaddoura, S., Itani, M., & Roast, C.: Analyzing the Effect of Negation in Sentiment Polarity of Facebook Dialectal Arabic Text. *Applied Sciences*, vol. 11, no. 11, p. 4768, (2021). https://doi.org/10.3390/app11114768