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 textual data into emotional categories such as sadness, anger, and happiness, the study aims to accurately identify emotional tones and provide actionable insights. These findings could inform public sentiment analysis on various issues and support applications like mental health monitoring and user experience enhancement.

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

1 Introduction

The project will focus on Sentiment Analysis, where we classify the statements (e.g., tweets) into the emotional tone they convey, such as sadness, anger, happiness, or other emotions. By using machine learning models such as Naive Bayes, Logistic Regression, Random Forest, and SVM, the study aims to identify the emotional tone of given textual data accurately, and label based on what was predicted.

Relevant literature includes studies on text classification and neural networks for Natural Language Processing (NLP). Research in this field often highlights the balance between interpretability (Naive Bayes, Logistic Regression) and predictive power (Random Forest, SVM). Additionally, while the project's primary objective is to determine the emotion associated with the data, this classification lays the foundation for further analysis or applications, such as understanding emotional trends or serving as a component in larger systems like mental health monitoring or user experience tools.

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]. A study of Nandwani and Verma (2021) emphasized the importance of sentiment analysis and emotion detection in processing data from social media, where users express emotions and opinions freely. Sentiment analysis identifies the polarity of text (positive, negative, or neutral), while emotion detection delves into specific emotional states like joy or anger. These techniques, supported by NLP, are applied across business, healthcare, and education to enhance customer feedback analysis, monitor mental health, and improve teaching methods. Methods range from lexicon-based to machine learning and deep learning approaches, each with unique strengths and challenges, including managing context and ambiguity in language [4].

3 Proposed Method

We propose using Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machine (SVM) as our classification algorithms. Naive Bayes is computationally efficient and works well with smaller datasets. Logistic Regression provides a balance between simplicity and predictive power, while Random Forest, leverages ensemble learning to capture complex relationships in the given data. Lastly, SVM shines in higher dimensional spaces effective for text classification tasks.

Steps include:

- Preprocessing text (tokenization, stopword removal)
- Feature extraction using TF-IDF and N-grams
- Model training with both algorithms
- Evaluation and comparison of models using accuracy, precision, and recall

We selected these methods based on their performance in prior sentiment analysis studies.

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

4 Dataset

We will use a publicly available dataset from Kaggle (e.g., Emotions Analysis Dataset), which contains emotion-labeled sentences. The data includes thousands of samples categorized into sadness, surprise, joy, love, fear, anger emotion. The dataset link can be accessed here.

Model	Description
Naive Bayes	A probabilistic model that applies Bayes' theorem with an assumption of feature independence.
Logistic Regression	A statistical model used for binary and multi-class classification, effective for linearly separable data.
Random Forest	An ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.
SVM	A supervised learning model that uses hyperplanes to separate data points into classes, effective in high-dimensional spaces.

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

5 Metrics for Evaluation

We will use the following metrics to evaluate the performance of our models based on Table 1:

Metric	Description
Accuracy	Measures the overall correctness of predictions.
Precision and Recall	To handle class imbalance and measure the model's performance for each emotion.
F1 Score	The harmonic mean of precision and recall, providing a balance between the two metrics.
Confusion Matrix	To visualize the performance of each category.

Table 2. 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	For natural language processing tasks, including text preprocessing and feature extraction.

Table 3. 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
4. Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining, 11(81). https://doi.org/10.1007/s13278-021-00776-6