IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 4, March 2025

Sentiment Analysis on X (Formerly Twitter) using Machine Learning

Sarang Rajput, Yash Sonawane, Dipak Bhosale, Gaurav Kadam
Guru Gobind Singh Polytechnic, Nashik, India

Abstract: *Sentiment Analysis plays a crucial role in understanding public opinions on social media platforms. This paper presents a machine learning-based sentiment analysis model trained on 10,000 tweets obtained from Kaggle. Basic preprocessing steps were applied, including removal of HTML tags, conversion to lowercase, stopword removal, and abbreviation replacement. Logistic Regression achieved an accuracy of 69%, while Random Forest obtained 68%. A key observation was that Google AI's Gemini better understood sarcasm and provided insights, whereas our models predicted sentiment based on surface- level analysis. The study highlights the need for context-aware NLP models for better sentiment classification.*

Keywords: Sentiment Analysis

I. INTRODUCTION

- Social media platforms have revolutionized communication, allowing individuals to express their opinions on a vast scale. With millions of tweets posted daily, extracting meaningful insights from this data has become essential for businesses, politicians, and researchers.
- Sentiment analysis aims to classify text as positive, negative, or neutral, helping in decision- making processes. However, traditional machine learning models face challenges in accurately interpreting sentiment, especially in the presence of sarcasm and ambiguous language.
- This paper explores a machine learning-based approach to sentiment analysis on X (formerly Twitter), comparing its performance with modern AI systems such as Google's Gemini.

II. RELATED WORK

- Several studies have explored sentiment analysis using both traditional machine learning and deep learning models.
- Popular methods include Naïve Bayes, Support Vector Machines (SVM), and Decision Trees. Deep learning approaches such as Long Short-Term Memory (LSTM) networks and transformer- based models like BERT and RoBERTa have significantly improved contextual understanding.
- However, sarcasm detection remains a challenging problem. Recent advancements in large language models (LLMs) like GPT and Gemini have demonstrated superior comprehension capabilities, surpassing traditional models.

III. METHODOLOGY

3.1 Dataset

- The dataset consists of 10,000 tweets obtained from Kaggle, labeled into positive, negative, and neutral sentiments.
- These tweets were collected from various domains, including politics, sports, and entertainment, ensuring a diverse dataset.
- A key challenge in sentiment analysis is the presence of informal language, slang, and emojis, which can affect model performance.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 4, March 2025

3.2 Preprocessing

Text preprocessing is crucial for improving model accuracy. The following steps were applied:
- HTML tag removal: Eliminating unnecessary tags from tweets.
- Lowercasing: Converting all text to lowercase to ensure uniformity.
- Stopword removal: Removing commonly used words that do not contribute to sentiment.
- Abbreviation replacement: Expanding common abbreviations (e.g., 'u' to 'you').
- Tokenization: Splitting sentences into individual words.
- Lemmatization: Converting words to their base form (e.g., 'running' to 'run'). The cleaned data was then used to train machine learning models.

3.3 Feature Engineering

• We used TF-IDF (Term Frequency-Inverse Document Frequency) for feature extraction. TF-IDF represents the importance of words in a document relative to the entire dataset.
• This method helps improve classification performance by reducing noise in textual data.

3.4 Machine Learning Models

• We implemented Logistic Regression and Random Forest for sentiment classification. Hyperparameter tuning was performed using GridSearchCV to optimize model performance.
• The models were evaluated using precision, recall, and F1-score metrics.

IV. RESULTS AND DISCUSSION

4.1 Model Performance

• Logistic Regression achieved an accuracy of 69%, while Random Forest obtained 68%.
• To further evaluate the models, we analyzed precision, recall, and F1-score for each sentiment class.
• Confusion matrices revealed that both models struggled with sarcasm detection, often misclassifying sarcastic tweets as neutral.

4.2 Comparison with Google AI's Gemini

A key observation was that our models struggled to understand sarcasm, whereas Google AI's Gemini demonstrated superior contextual comprehension. Gemini was able to detect sarcasm and provide deeper insights into tweet sentiments. This highlights the limitations of traditional ML models and the need for more advanced NLP techniques.

4.3 Challenges and Limitations

1. Sarcasm Detection: Traditional ML models lack the ability to detect sarcasm effectively.
2. Data Imbalance: Neutral tweets were more prevalent in our dataset, leading to bias in classification.
3. Domain-Specific Language: Tweets from different domains require specialized sentiment models.
4. Limited Training Data: Larger datasets could improve model generalization.

V. CONCLUSION AND FUTURE WORK

This study demonstrates the potential and limitations of traditional machine learning models in sentiment analysis. Future work will explore advanced deep learning techniques, such as transformer-based models (e.g., BERT, RoBERTa) to improve contextual understanding. Additionally, fine-tuning large language models (LLMs) specifically for sarcasm detection could further enhance performance. Integrating external knowledge sources and using multi-modal analysis (text, emojis, and images) can further improve sentiment classification.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 4, March 2025

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in Advances in Neural Information Processing Systems, 2013.

[3] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval,2008.