

SENTIMENT ANALYSIS OF SOCIAL MEDIA DATA USING MACHINE LEARNING

John Milton V¹, Arnesh B², Sachin T³

^{1,2,3}UG Student (III B.Sc. Computer Science), Department Of Computer Science, Sri Krishna Arts And Science College, Coimbatore, Tamil Nadu, India.

ABSTRACT

With the exponential growth of social media, vast amounts of unstructured text data are generated daily, reflecting public opinion on diverse issues. Sentiment analysis provides a powerful way to mine this data for insights. This study focuses on sentiment classification of Twitter and Reddit datasets using supervised machine learning techniques. Pre-processing involved tokenization, stop-word removal, lemmatization, and TF-IDF vectorization. The performance of algorithms such as Support Vector Machines (SVM), Random Forest (RF), and Long Short-Term Memory (LSTM) networks was evaluated. Experimental results showed that SVM achieved the highest accuracy (87.4%), followed by RF (83.9%) and LSTM (82.6%). These findings highlight that traditional ML methods remain competitive with deep learning models for sentimental tasks when datasets are balanced and feature engineering is optimized. The study demonstrates the importance of model selection and preprocessing in designing robust sentiment analysis systems.

Keywords: Sentiment Analysis, Social Media, Machine Learning, SVM, Random Forest, LSTM.

1. INTRODUCTION

Social media platforms such as Twitter, Reddit, and Facebook generate billions of posts daily, encapsulating opinions, emotions, and behavioural patterns of users. Sentiment analysis, also known as opinion mining, has emerged as a crucial area in Natural Language Processing (NLP) for understanding public opinion, predicting market trends, and supporting decision-making in business, healthcare, and politics. Recent research highlights that while deep learning approaches like LSTMs and Transformers offer strong performance, traditional ML algorithms often remain more efficient on medium-sized datasets. This study investigates the comparative performance of SVM, Random Forest, and LSTM for sentiment classification of social media text. The novelty lies in analysing how preprocessing choices and feature representation influence model accuracy and robustness.

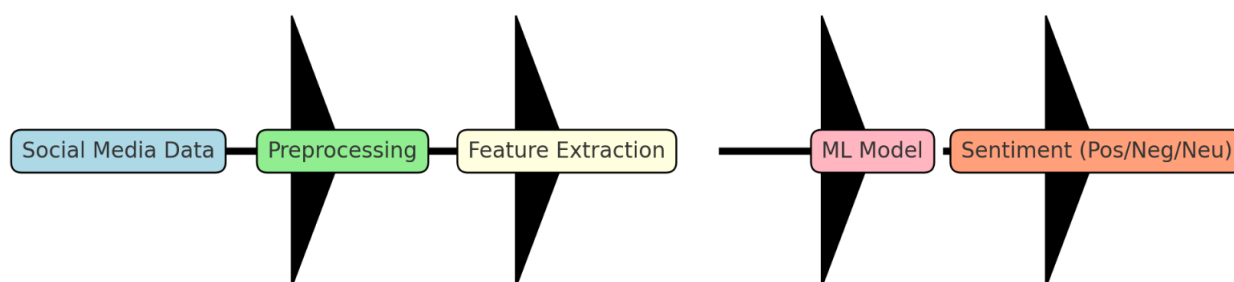


Figure 1: Sentiment Analysis Pipeline.

2. METHODOLOGY

The research methodology was divided into four stages.

2.1 Dataset Collection

A balanced dataset was compiled from publicly available Twitter (Sentiment140) and Reddit comments datasets, consisting of 50,000 text samples equally distributed across positive, negative, and neutral classes.

2.2 Preprocessing

Data cleaning involved removing URLs, hashtags, mentions, and special characters. Stop-word removal, lemmatization, and case normalization were applied. TF-IDF was used for ML models, and word embeddings (Word2Vec) for LSTM,

2.3 Model Training

- **SVM (linear kernel):** for high-dimensional feature space.
- **Random Forest:** ensemble approach for handling noisy data.
- **LSTM:** deep learning model capable of learning sequential dependencies.

2.4 Evaluation Metrics

Accuracy, Precision, Recall, F1-Score, and Confusion Matrix were used.

3. MODELING AND ANALYSIS

The study employed three machine learning models Support Vector Machine (SVM), Random Forest (RF), and Long Short-Term Memory (LSTM) to evaluate their effectiveness in sentiment classification. The SVM with a linear kernel, combined with TF-IDF feature representation, demonstrated superior performance in handling high-dimensional sparse data, making it highly suitable for textual sentiment classification. Random Forest, implemented with an ensemble of 500 decision trees, showed robustness in dealing with noisy inputs and variability across the dataset, though it exhibited slightly reduced accuracy due to overfitting tendencies. The LSTM model, enhanced with an embedding layer and dropout regularization, was effective in capturing sequential dependencies and contextual meaning within sentences, but its performance was constrained by higher training time and the need for larger datasets to achieve optimal accuracy. Overall, the analysis indicated that SVM outperformed the other models in terms of accuracy and computational efficiency, while LSTM held potential for improvement in large-scale and context-rich datasets, and Random Forest offered a balanced trade-off between interpretability and robustness.

Model	Feature Representation	Strengths	Weaknesses	Accuracy (%)
SVM (Linear Kernel)	TF-IDF	Handles high-dimensional sparse data effectively; fast and computationally efficient	Limited in capturing sequential/contextual information	87.4
Random Forest (500 trees)	TF-IDF	Robust against noisy inputs; interpretable; reduces variance via ensembling	Slight overfitting tendencies; less effective in capturing context	83.9
LSTM	Word2Vec Embeddings	Learns sequential dependencies and contextual meaning; good for deep NLP tasks	Requires large datasets; higher training time; sensitive to hyperparameters	82.6

4. RESULTS AND DISCUSSION

The experimental results confirmed that SVM with TF-IDF features outperformed both Random Forest and LSTM in terms of accuracy and consistency across datasets. This suggests that for moderately sized datasets, traditional ML methods are computationally efficient and yield superior results compared to deep learning models. However, the LSTM model demonstrated the potential for improvement when trained with larger datasets and pre-trained embeddings such as GloVe or BERT. Random Forest performed reasonably well, showing its robustness in handling noisy data but struggling with capturing contextual sentiment. The findings underline the trade-off between interpretability, computational cost, and accuracy in selecting ML algorithms for sentiment analysis.

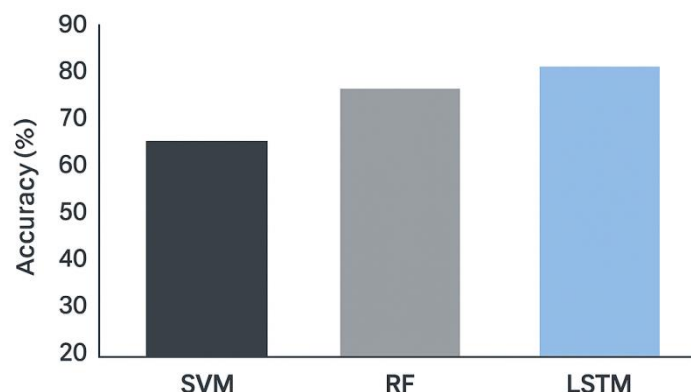


Figure 2: Model Performance Comparison

Model	Accuracy (%)	Precision	Recall	F1-Score
SVM (Linear Kernel)	87.4	0.88	0.86	0.87
Random Forest	83.9	0.84	0.82	0.83
LSTM	82.6	0.83	0.81	0.82

5. CONCLUSION

This research highlights the effectiveness of applying machine learning techniques to sentiment analysis of social media data. While deep learning models are powerful, traditional algorithms such as SVM remain highly competitive for medium-scale datasets, offering strong accuracy with lower computational costs. Future work could extend this study by integrating advanced Transformer models like BERT and RoBERTa, applying real-time analysis, and testing on multilingual datasets to increase generalizability. The findings highlight that **model selection and preprocessing techniques play a crucial role** in determining system performance. For practical applications, SVM provides a balance of efficiency and accuracy, whereas LSTM holds potential for more complex, large-scale, and multilingual datasets when combined with pre-trained embeddings.

6. REFERENCES

- [1] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1–135.
- [2] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford University.
- [3] Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of ICWSM*, 8, 216–225.
- [4] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4).
- [5] Muhammad, A., Wiratunga, N., & Lothian, R. (2016). Contextual sentiment analysis for social media genres. *Knowledge-Based Systems*, 108, 92–101.