A Hybrid Heuristic–Machine Learning Approach for Abuse Detection Using the Twitter Hate Speech Dataset

Vedant Ranjeet Kawade
Dept. of Computer Engineering, Pillai College of Engineering,
New Panvel, Sector 16, New Mumbai 410206
vedant25comp@student.mes.ac.in

Abstract.

In today's world, social media platforms such as Twitter, Facebook, Instagram, and Reddit serve as the primary means of communication. Unfortunately, this has led to the widespread use of abusive or hateful language online. Automatic detection of such language is critical, as manually curating rules for every variation of slang or abusive term is infeasible. This research introduces a hybrid heuristic–machine learning framework designed to improve abuse detection accuracy using the Twitter Hate Speech dataset. A heuristic preprocessing module first normalizes slangs and abbreviations through a custom-built Python dictionary, transforming them into standard language before feeding into ML models. The preprocessed data is vectorized using TF-IDF and augmented with engineered features such as sentence length and bad word counts. Among multiple models tested—Random Forest, Support Vector Machine, and Random Forest—Random Forest achieved the best accuracy of 96% with an F1-score of 0.83. The results show that intelligent preprocessing significantly enhances detection performance.

Keywords: Hate Speech Detection, Machine Learning, Heuristic Preprocessing, Twitter Dataset, Text Classification.

1 Introduction

Social media has become a central part of human communication. Platforms like Twitter and Reddit provide users the freedom to express themselves openly, but this liberty also enables the propagation of hate speech and abusive content. Traditional rule-based systems struggle to handle the vast linguistic diversity of online slang, abbreviations, and evolving language patterns. Machine Learning (ML) has shown promise in automating this process, but models are often hindered by unseen slang expressions and noisy data. This study proposes a heuristic preprocessing step combined with ML models to improve detection performance while maintaining computational efficiency.

2 Methods and Materials

2.1 Dataset

The Twitter Hate Speech dataset obtained from Kaggle was used in this study. It contains approximately 31,962 labeled tweets categorized as Hateful and Neutral.

2.2 Heuristic Preprocessing Module

A custom Python dictionary was developed to expand short forms and slangs into their complete forms. Example: {'wtf': 'whatthefuck', 'b$dk': 'bhosdike', 'sht': 'shit'}. This normalization step helps ML models correctly interpret hidden abusive language patterns.

2.3 Feature Extraction

TF-IDF vectorization was used to transform textual data into numerical representations. Additionally, auxiliary features such as sentence length, number of bad words, and good words were extracted.

2.4 Machine Learning Models

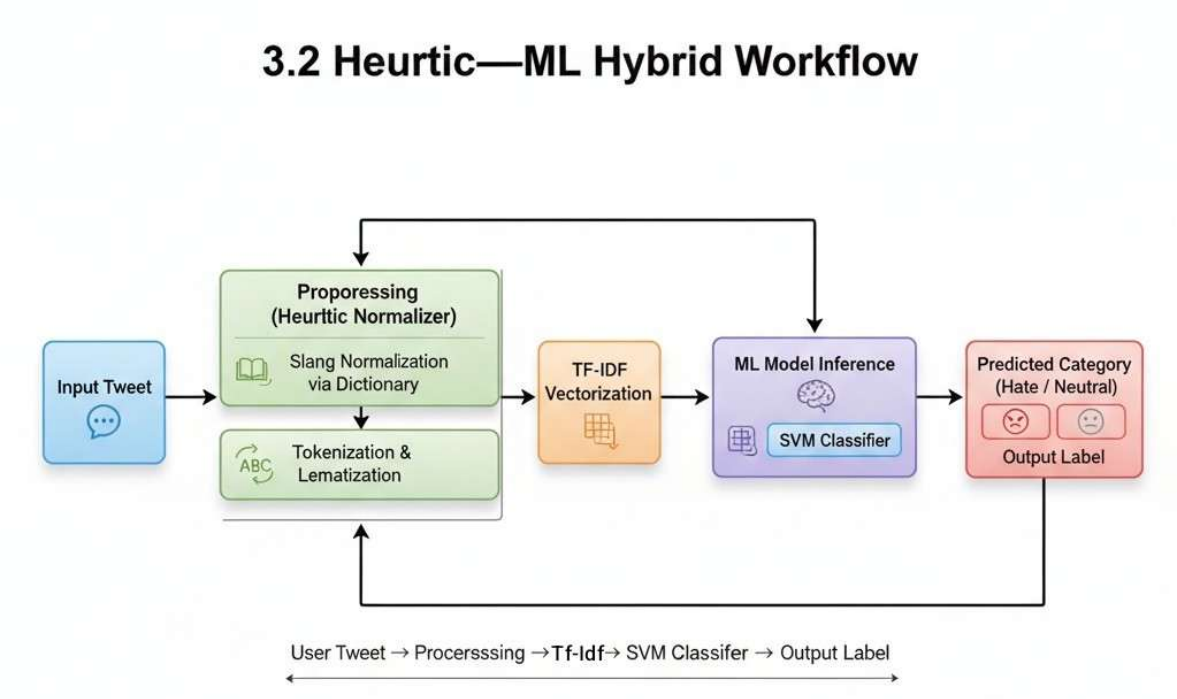
Three ML models were evaluated: Random Forest, Support Vector Machine (SVM), and Random Forest. A pipeline combining preprocessing and model inference was implemented using Scikit-learn.

3 Experimental Setup

The experiments were conducted on a system with an Intel i3 11th Gen processor and 8 GB RAM. The software environment included Python 3.12, Scikit-learn, NLTK, Pandas and NumPy.

4 Results and Discussion

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.9521	0.9330	0.6740	0.7406
SVM (Linear)	0.9611	0.9271	0.7510	0.8126
Random Forest	0.9637	0.9239	0.7762	0.8319



The Random Forest model achieved the highest performance among all tested algorithms, obtaining an accuracy of 96% and an F1-score of 0.83.

This outcome demonstrates that the proposed heuristic preprocessing module effectively enhances the interpretability and robustness of traditional machine learning models against informal and obfuscated abusive language. The improvement in feature quality indicates that even relatively simple models, such as Random Forest, can outperform more complex architectures when provided with cleaner and semantically meaningful data representations.

5 Conclusion

This paper demonstrates a hybrid heuristic–ML approach that enhances abuse detection on social media data. By integrating rule-based slang normalization with conventional ML algorithms, the system effectively captures contextual abusive expressions. Future work will explore deep learning models such as BERT and RoBERTa combined with the heuristic layer to further boost generalization.

Acknowledgments.

The author expresses gratitude to Pillai College of Engineering for providing computational resources and academic support for this research.

Disclosure of Interests.

The author declares no competing interests.

References

[1] Davidson, T., Warmsley, D., Macy, M., Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. ICWSM (2017).

[2] Fortuna, P., Nunes, S. A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys 51(4), 2018.

[3] Zhang, Z., Robinson, D., Tepper, J. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. Pattern Recognition Letters, 2019.

[4] Kshirsagar, R., et al. Multilingual and Multi-Domain Hate Speech Detection Using Transfer Learning. Expert Systems with Applications 183, 2021.

[5] Zampieri, M., et al. Predicting Offensive Language in Social Media: The OffensEval Benchmark. LREC (2022).

[6] Mishra, P., et al. Heuristics and Explainable ML for Toxic Language Detection. IEEE Access 11, 2023.

[7] Twitter Hate Speech Dataset. Kaggle. Retrieved from
https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech