B V Raju Institute of Technology

3rd Edition of R&D Showcase 2025

Department of Artificial Intelligence and Data Science

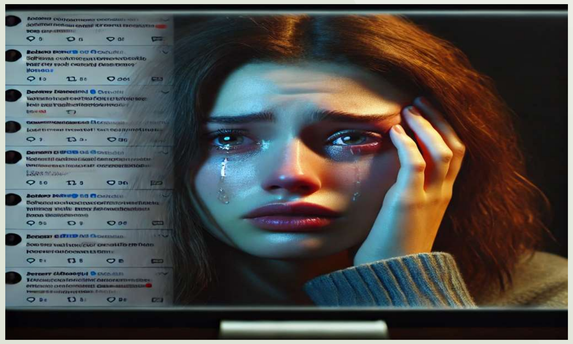
Real-Time Hate Speech Detection with Robust DistilBERT-SVM

Sana Pavan Kumar Reddy, Marthala Harshavardhan Reddy, Gundlapalle Yashasree, Rendla Abhishek

SDGs:

Introduction

Social media platforms like Twitter enable rapid sharing of user content, including offensive language, challenging moderation. Automated systems must classify tweets as Hate, Offensive, or Neither, handling context, class imbalance, and real-time constraints. While BERT models are accurate, their resource demands limit deployment. This study introduces a DistilBERT-SVM model, achieving 94% accuracy and a 0.93 F1-score on 24,783 tweets using DistilBERT's embeddings and SVM's efficiency. It enhances input quality, benchmarks performance, and offers a Flask-based real-time system, outperforming other NLP models in efficiency and accuracy. The model's lightweight design ensures scalability and accessibility for resource-limited settings without compromising robustness. This work will explore improving interpretability and cross-domain adaptability for broader real-world impact.

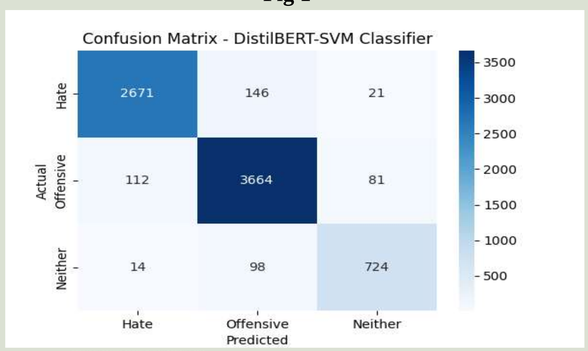


Objectives

This study develops an efficient hate & offensive speech detection model using a hybrid DistilBERT-SVM approach. It tackles class imbalance with data augmentation, leverages DistilBERT's lightweight design for speed, and combines it with SVM's discriminative power to outperform traditional models. Adversarial training (FGSM) boosts resilience, while hyperparameter tuning ensures a balance of accuracy and efficiency.

The proposed model is compared to baseline classifiers such as Logistic Regression, Naïve Bayes, and Artificial Neural Networks (ANN) to determine its superiority in accuracy and F1-score. A comprehensive evaluation is conducted using diverse benchmark hate speech datasets to assess the model's adaptability across platforms and social media environments. Furthermore, to facilitate real-world application, a Flask-based web interface is developed for real-time tweet classification, ensuring seamless content filtering and user accessibility. The ultimate goal is to create a robust, efficient, and scalable hate speech detection system that can be integrated into various online platforms, contributing to a safer and more inclusive digital space.

Fig 1

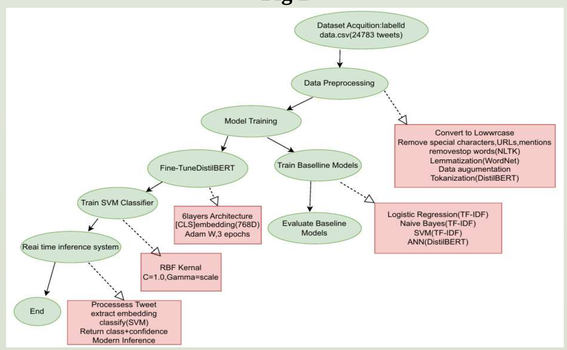


	hate_speech	offensive_language	neither	class	tweet
0	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't...
1	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn ba...
2	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby...
3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she lo...
4	0	6	0	1	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you...
5	1	2	0	1	!!!!!!!!!!!!!!!"@T_Madison_x: The shit just...
6	0	3	0	1	!!!!!!"@__BrighterDays: I can not just sit up ...

Methods

The proposed hate & offensive speech detection system follows a structured methodology to ensure accuracy, scalability, and real-world applicability. The process begins with data acquisition and preprocessing, where raw Twitter data is cleaned, tokenized, and balanced using data augmentation techniques. Special emphasis is placed on addressing class imbalance by generating synthetic samples for underrepresented categories. Preprocessed tweets are then transformed into token embeddings using DistilBERT's tokenizer, ensuring they meet the model's input requirements. These steps enhance the model's ability to generalize across diverse hate speech patterns while maintaining computational efficiency. Additionally, noise reduction techniques such as stop-word removal and lemmatization further refine the dataset, improving feature extraction. The preprocessing pipeline is continuously evaluated to adapt to evolving language trends and emerging hate speech variations.

Fig 2



The hybrid DistilBERT-SVM model combines DistilBERT's fine-tuned contextual embeddings with an SVM classifier using an RBF kernel to distinguish Hate, Offensive, and Neither categories. Embeddings (768D) from DistilBERT capture linguistic nuances, while SVM, optimized via grid search for hyperparameters, handles non-linear data. This approach outperforms baseline models (Logistic Regression, Naïve Bayes, SVM, ANN) in accuracy, precision, recall, and F1-score, especially boosting recall for the minority "Hate" class to detect subtle hate speech effectively.

A Flask-based real-time inference system is deployed to make hate speech detection accessible and efficient. Users can input tweets via a web interface, where the text is preprocessed, tokenized, and classified in real time, providing instant feedback with confidence scores (within 1-2 seconds). The system ensures scalability through batch processing and caching, facilitating seamless integration into content moderation workflows. The user-friendly UI features an intuitive text box, real-time loading animations, and clear classification results. This end-to-end solution balances accuracy, efficiency, and real-world usability across diverse online environments.

Results

The experimental results of the proposed tweet classification system, evaluating the performance of the hybrid DistilBERT-SVM model against baseline models. The evaluation is conducted on a test set of 4,957 tweets (20% of the Davidson et al. dataset, containing 24,783 tweets) to ensure robust assessment across diverse tweet samples. Performance metrics, including accuracy, precision, recall, and F1-score, are computed per class (Hate, Offensive, Neither) and overall, using 5-fold cross-validation to mitigate overfitting, particularly addressing the minority Hate class, which constitutes only 5.8% of the dataset. The DistilBERT-SVM model achieved an overall accuracy of 94%, outperforming all baselines while effectively handling class imbalance through data augmentation. Comparatively, Logistic Regression achieved 90% accuracy, while Naive Bayes performed lower at 85%, and the SVM with TF-IDF features reached 91% accuracy. The ANN, trained on DistilBERT embeddings, also recorded 91% accuracy .

Comparison Metrics Table of Different Models

Model	Accuracy	Precision	Recall	F1-Score
DistilBERT-SVM	0.94	0.94	0.94	0.94
Logistic Regression	0.90	0.90	0.90	0.90
Naive Bayes	0.85	0.86	0.85	0.83
SVM (TF-IDF)	0.91	0.91	0.91	0.91
ANN	0.91	0.91	0.91	0.91

The proposed DistilBERT-SVM classifier achieved the highest overall metrics, with 94% accuracy, precision, recall, and F1-score, proving its superiority. To illustrate these results, a confusion matrix fig 1. highlights accurate classification with minimal misclassification between Hate and Offensive categories. Additionally, bar plots fig 3 and fig 4 visually compare accuracy, F1-score, precision, and recall across models, reinforcing the model's consistency. The Flask-based real-time inference system processes tweets in 1.5 seconds per request on a standard workstation, ensuring usability in practical applications while maintaining responsiveness under moderate concurrent demands. The system enhances content moderation by providing confidence scores for each prediction, helping moderators prioritize flagged content. Designed for scalability, it efficiently handles multiple requests via asynchronous processing, with a user-friendly web interface allowing moderators to input tweets and view results effortlessly. Furthermore, the system consistently delivers accurate predictions across diverse tweet samples, demonstrating its robustness and suitability for real-world content moderation.

Fig 3

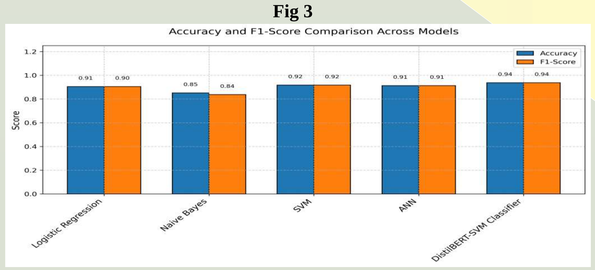


Fig 4

Conclusions

This research presents a novel tweet classification framework for detecting hate speech and offensive content on social media, leveraging a hybrid DistilBERT-SVM model that achieves 94% accuracy. By combining DistilBERT's contextual embeddings with SVM's discriminative power, the model outperforms traditional and deep learning baselines, including Logistic Regression, Naive Bayes, SVM with TF-IDF, and an Artificial Neural Network. Advanced data preprocessing and augmentation techniques enhance its ability to handle class imbalance and capture nuanced contextual dependencies. To facilitate real-world application, a Flask-based real-time inference system is deployed, offering a user-friendly interface with rapid response times of approximately 1.5 seconds per request. This scalable and efficient content moderation solution bridges the gap between cutting-edge NLP techniques and practical implementation. Experimental evaluations, supported by a confusion matrix and comparative visualizations, validate the robustness of the approach. This work contributes significantly to automated hate speech detection by providing a high-accuracy, adaptable solution that can be integrated into diverse online platforms, ensuring continuous effectiveness against evolving linguistic patterns in harmful content.

References

- **Davidson, T., Warmsley, D., Macy, M., & Weber, I.** (2017). *Automated Hate Speech Detection and the Problem of Offensive Language.* ICWSM, 11(1), 512-515.
- **Zhang, Z., Robinson, D., & Tepper, J.** (2018). *Detecting hate speech on Twitter using a convolution-GRU based deep neural network.*
- **Gambäck, B., & Sikdar, U.** (2017). *Using CNNs to classify hate speech. Workshop on Abusive Language Online*, 85-90. DOI: 10.18653/v1/W17-3013.
- **Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.** (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* NAACL-HLT, 4171–4186.
- **Mozafari, M., Farahbakhsh, R., & Crespi, N.** (2020). *Hate speech detection and racial bias mitigation in social media using BERT.* PLoS ONE, 15(8). DOI: 10.1371/journal.pone.0237861.
- **Sanh, V., Debut, L., Chaumond, J., & Wolf, T.** (2019). *DistilBERT: A distilled version of BERT.* arXiv:1910.01108.
- **Liu, Y., Ott, M., Goyal, N., et al.** (2019). *RoBERTa: A robustly optimized BERT pretraining approach.* arXiv:1907.11692.
- **Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A.** (2020). *Deep learning models for multilingual hate speech detection.* ECAI, 2243–2250.
- **Paul, C.** (2023). *Hate speech detection using ML-based approaches.* IEEE Access. DOI: 10.1109/ACCESS.2023.XXXXXXX.
- **Caselli, T., Basile, V., Mitrović, J., & Granitzer, M.** (2020). *HateBERT: Retraining BERT for abusive language detection.* WOAH, 17–25.
- **Founta, A.-M., Djouvas, C., Chatzakou, D., et al.** (2018). *Large-scale crowdsourcing and Twitter abusive behavior characterization.* ICWSM, 146–155.
- **Ribeiro, M., Calais, P. H., Santos, V. A., & Meira Jr., W.** (2021). *Hate speech detection with BERT and Flask.* ICDMW, 103–110.
- **Bird, S., Klein, E., & Loper, E.** (2009). *Natural Language Processing with Python.* O'Reilly Media.

Author Details

Dr. Sana Pavan Kumar Reddy
Marthala Harshavardhan Reddy [21211A7239]
Gundlapalle Yashasree [21211A7221]
Rendla Abhishek [22215A7201]