Hate Speech Detection on Social Media using Transformer-based Models

Author: Aayushi Patel  
Department of Computer Science  
SHRI RAM INSTITUTE OF SCIENCE AND TECHNOLOGY JABALPUR (Rajiv Gandhi Proudyogiki Vishwavidyalaya Bhopal MP)  
Email: - aayuship378@gmail.com

# Abstract

Hate speech has become a major challenge in online social media platforms, contributing to toxicity, misinformation, and harassment. In this paper, we propose an approach for hate speech detection using transformer-based models such as BERT and its variants. The study focuses on practical implementation in Google Colab for efficient training and evaluation. Experimental results show that transformer models outperform traditional baselines in terms of accuracy and robustness. The proposed framework demonstrates the potential of lightweight models for real-time detection of online hate speech.

# Keywords

Hate Speech Detection, Natural Language Processing, Transformers, BERT, Deep Learning, Google Colab

# 1. Introduction

The exponential growth of social media has led to an unprecedented rise in user-generated content. While this has improved communication, it has also provided a medium for spreading hate speech, cyberbullying, and online harassment. Automated detection of hate speech is essential to maintain healthy online communities.  
  
Traditional machine learning models such as Logistic Regression and SVM have shown moderate performance, but with the advent of deep learning and pre-trained transformer models, significant improvements have been achieved. This research explores transformer-based architectures for hate speech detection.

# 2. Related Work

Several researchers have used machine learning for offensive language detection. Recent work focuses on transformer-based architectures such as BERT, DistilBERT, and TinyBERT. These models leverage contextual embeddings to capture semantic meaning effectively.

# 3. Methodology

The proposed framework consists of the following stages:  
- Dataset: A labeled dataset of tweets was used (Davidson et al.).  
- Preprocessing: Cleaning, tokenization, stopword removal.  
- Models: Baseline classifiers (SVM, Logistic Regression) vs. Transformer models (BERT, TinyBERT, DistilBERT).  
- Implementation: Training and evaluation conducted in Google Colab using PyTorch and HuggingFace Transformers.

# 4. Experiments and Results

The models were evaluated on accuracy, precision, recall, and F1-score. Colab implementation allowed efficient training with GPU support.

## 4.1 Baseline Results

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall |
| Logistic Regression | 0.72 | 0.70 | 0.69 |
| SVM | 0.74 | 0.71 | 0.72 |

## 4.2 Transformer Models

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall |
| BERT | 0.82 | 0.81 | 0.80 |
| TinyBERT | 0.79 | 0.78 | 0.77 |
| DistilBERT | 0.80 | 0.79 | 0.78 |

# 5. Conclusion

This research demonstrates the effectiveness of transformer-based models for hate speech detection. Among all tested models, BERT achieved the best performance. Colab provided a simple and efficient platform for training. Future work includes integrating explainable AI techniques and deploying the model as a real-time application.

# Acknowledgment

The author would like to thank the open-source community and Google Colab for providing accessible computational resources.

# References

[1] Davidson, T., et al. Automated Hate Speech Detection and the Problem of Offensive Language. ICWSM, 2017.

[2] Devlin, J., et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT, 2019.

[3] Jiao, X., et al. TinyBERT: Distilling BERT for Natural Language Understanding. EMNLP, 2020.

[4] Sanh, V., et al. DistilBERT: A distilled version of BERT. arXiv:1910.01108, 2019.