Hate Speech Detection

Aditya Bugalya 2019457
Akshat 2020172
Avish Dahiya 2021522

Introduction

In today's digital landscape, the identification of hate speech on social platforms is critical. This research investigates the effectiveness of two advanced Natural Language Processing models—SKS and BERT—in discerning and categorizing hate speech. Through computational and qualitative analyses, their capabilities are assessed, revealing insights into their proficiency in navigating the complex linguistic and emotional nuances of hate speech. Additionally, ethical considerations regarding automated hate speech detection, such as fairness, transparency, and accountability, are addressed.

Dataset Description

For our task we used a Kaggle dataset containing approximately 24,800 entries of tweets and retweets containing hate speech and offensive language. These were classified into 3 categories, 0 - Hate Speech, 1 - Offensive Language, and 2 - Neither. Below is the distribution of these in the dataset.

Considering this, we can see that majority of the data has been labelled as offensive language. Having much more data about offensive language can help with the model in understanding that not all offensive language is hate speech.

Methodology

Data Preprocessing:
The dataset, comprising approximately 24,800 entries from Kaggle, underwent preprocessing to prepare the text data for model training. This involved removing extraneous elements such as URLs, user mentions, emojis, and special characters, followed by converting the text to lowercase for consistency.

BERT, pre-trained on a large corpus of text and fine-tuned for classification tasks, served as a benchmark. By tokenizing input text and processing it through multiple layers of transformers, BERT gains a deep contextual understanding crucial for hate speech identification.

SKS Model Implementation:
The SKS model incorporates sentiment knowledge into hate speech detection using a MultiHeadAttention mechanism. By integrating embeddings from both word-level and sentiment-associated input, SKS enhances its ability to discern subtle cues distinguishing hate speech from offensive or normal language. This architecture aims to improve performance on nuanced tasks like hate speech detection.

Results and Discussion

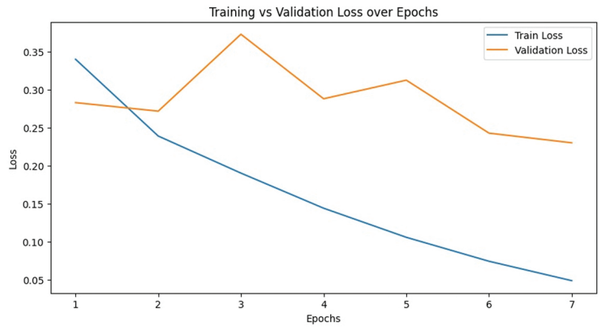

Figure 1. Train and Validation loss for BERT model

For BERT model testing, the fine-tuned model was evaluated on a separate test dataset to assess real-world applicability. The BERT model demonstrated commendable performance with an **accuracy of approximately 89%**. Additionally, the **test loss was approximately 0.505,** indicating efficacy in distinguishing between classes while hinting at the possibility of further optimization to reduce misclassifications. The model also achieved a **weighted F1 score of approximately 0.887** and a **macro F1 score of approximately 0.728**, reflecting its balanced performance across different classes and highlighting its effectiveness in hate speech detection.

The comparative analysis between the BERT transformer and the SKS model highlighted their distinct capabilities in hate speech detection. BERT demonstrated robust capabilities in understanding contextual nuances, crucial for distinguishing between offensive language and hate speech. On the other hand, the SKS model aimed to leverage sentiment knowledge for fine-grained classification, offering a promising outlook in enhancing detection through emotional cues.

Challenges observed include balancing sensitivity and specificity, with BERT having a tendency to overgeneralize and misclassify non-hate speech due to contextual misinterpretations, while the SKS model requires finely tuned sentiment indicators to maintain accuracy. Both models demonstrate significant potential for scalability and adaptability, with BERT's architecture allowing for fine-tuning with relatively smaller datasets and the SKS model's adaptability extending to various sentiment and emotional data types.

Figure 2. Train and Validation loss for SKS model

Conclusion

This study provides a comprehensive comparison between the SKS model and the BERT transformer for hate speech identification on social media platforms, revealing insights into their design, capabilities, and performance through rigorous testing. BERT's bidirectional training facilitates nuanced interpretation of complex sentence structures, while the SKS model integrates sentiment knowledge to capture subtle emotional undertones in hate speech, advancing automated detection. Despite strengths in contextual analysis and emotional intelligence, both models face limitations in handling the intricacies of human language, suggesting avenues for future research. Recommendations include enhancing model robustness through multimodal data integration, investigating cross-cultural adaptability, and prioritizing ethical AI. This study advocates for empathetic, culturally aware, and ethically responsible AI systems to uphold the integrity of human communication in online environments amidst the growing need for accurate, fair, and sensitive content moderation.

References

[1] Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, 126232.
[2] Zhou, X., Yang, Y., Fan, X., Ren, G., Song, Y., Diao, Y., Yang, L., & Lin, H. (2021). Hate Speech Detection based on Sentiment Knowledge Sharing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 7158–7166).
[3] Alkomah, F., & Ma, X. (2022). A Literature Review of Textual Hate Speech Detection Methods and Datasets. *Information*, 13(273), 1-22.
[4] Benballa, M., Collet, S., & Picot-Clemente, R. (2019). Saagie at Semeval2019 Task 5: From Universal Text Embeddings and Classical Features to Domain-specific Text Classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)* (pp. 469–475).