Hate Speech Detection on Social Media

Noamen Syrine [1] **Neuenschwander Didier** [1]

Abstract

Hate speech detection is a critical area of research that addresses the identification and mitigation of harmful and abusive language on digital platforms. This study explores the development and evaluation of BERT model [1] to accurately detect hate speech across diverse datasets. We aim to identify various forms of hate speech, including racism, sexism, homophobia, and other discriminatory content. Our approach involves the collection of a comprehensive dataset, representing cultural contexts found online, to ensure the model's applicability and fairness. We demonstrate that our proposed model performs well in terms of accuracy and F1-score, offering a viable solution for real-time hate speech monitoring and intervention. This research contributes to the broader goal of fostering safer online communities by providing an effective tool for the early identification and management of hate speech.

Keywords: hate speech detection, transfer learning, language modeling, BERT, fine-tuning, social media

1. Introduction

More than 500 million tweets are sent daily [2], some of which is made up of hate speech that is difficult to discern manually in this vast number. Studies show that hate speech content on social media platforms like Facebook had a prevalence rate of approximately 0.05% in 2021, meaning about five out of every 10,000 posts contain hate speech [3]. This problematic behaviour is detrimental to the well-being of victims, leading to sleep problems, safety issues, increased anxiety, depressive thoughts and financial problems [4], These impacts underscore the need for automated detection systems to ensure healthier social interactions. The implementation of such filters must be handled with special care to avoid either infringing on freedom of expression or failing to protect individuals' well-being. Hate speech detection on social media is a critical yet challenging task, essential for maintaining healthy online interactions while preserving freedom of expression. Current limitations in this field include the difficulty of accurately distinguishing hate speech from non-hate speech due to the nuanced and context-dependent nature of language.

To improve the identification of hate speech on publicly available benchmark datasets, we adopted a transfer learning strategy using the pre-trained language model BERT, trained on English Wikipedia and BookCorpus. Our goal is to fine-tune DistilBERT for binary classification to effectively separate hate speech from non-hate speech. The aim of our study is to improve the performance of hate speech detection in the English language by exploiting the powerful text understanding capabilities of BERT and perform an analysis of the datasets used, as the accuracy of hate speech detection is highly dependent on the quality of the training data.

We hypothesise that:

- **BERT Transfer Learning:** Using a pre-trained BERT model will improve the accuracy of hate speech detection compared to models trained from scratch.
- **Impact on Dataset Quality:** A good understanding of the dataset used is essential as the definition of hate speech is not well defined [5].

2. Related Work

The increasing availability of large datasets has fueled the use of complex models like deep learning and graph embeddings to enhance hate speech detection in social media. Zhang et al. (2018) [6] demonstrated success with a hybrid convolutional and gated recurrent neural network (CNN-GRU), achieving improved results on most of their tested datasets. Existing work includes Huang et al. (2020) [7] where they assemble a multilingual Twitter corpus for the task of hate speech detection with inferred four author demographic factors: age, country, gender and race/ethnicity. The corpus covers five languages and they examine factors that can cause biases.

Our keyword reference is the benchmark hateXplain [8], the first benchmark for hate speech covering multiple aspects of the issue. Each post in the dataset is annotated (hate, offensive or normal) by 3 annotators. For another comparison with the BERT model detecting hate speech [9], from the official Hate Speech and Offensive Content Identification in English and Indo Languages competition. We will first fine tune BERT on the datasets used by these two references and

compare our result to theirs. Then we collect a large dataset and train on it.

3. Method

In this study, we employed the DistillBERT [10] model, a distilled version of BERT, due to its reduced memory footprint and faster inference time, which are crucial factors given our compute resources. We applied majority vote to extract one label out in the case with multiple annotators. Preprocessing consisted of data filtering and batched tokenization of the dataset. We retained only the relevant text and corresponding labels, and performed text cleaning to remove elements like usernames, and URLs that could introduce noise into the model. Fine-tuning was conducted on the training + validation set using the AdamW optimizer with weight decay 0.01 and a learning rate 1e-5. The additional hyperparameters determined through multiple experiments with early stopping. Given BERT's established robustness in capturing nuanced language patterns relevant to hate speech, we limit the training to 4 epochs. Model performance was continuously assessed using the validation set to prevent overfitting and ensure optimal training conditions. **From multi label classification to binary:** The distinction between "offensive" and "hateful" speech is often subjective and context-dependent [11] where they find that racist and homophobic tweets are more likely to be classified as hate speech but that sexist tweets are generally classified as offensive. Additionally, many hate speech datasets suffer from limited annotation diversity. Some rely on single labels from majority voting or a single expert, potentially overlooking the nuanced disagreements that arise from varying interpretations. This can lead to information loss and annotation bias. While metrics like Krippendorff's alpha attempt to quantify inter-annotator agreement, even well-known benchmarks like [8] achieve only moderate inter-agreement (0.46), underscoring the complexity of the task. This motivated our decision to adopt a binary classification approach. This approach merges "offensive" and "hateful" into a single "hateful" category, allowing us to focus on identifying the most harmful forms of online discourse while mitigating the challenges of multi-label classification. While acknowledging the nuances lost in this simplification, we believe the binary approach offers a more pragmatic and effective solution for hate speech detection in the context of our mini-project.

3.1. Case Study: HateXplain

The HateXplain dataset was collected from Twitter and Gab pots. It underwent a multi-step annotation process where annotators labeled posts as hateful, offensive, or normal, identified target communities, and provided rationales for their classifications. We began with a detailed analysis of this dataset. There are 11415 text inputs that are classified as hatespeech and 7814 as normal. We analyze the most common words that recur a lot in hateful or normal speech:

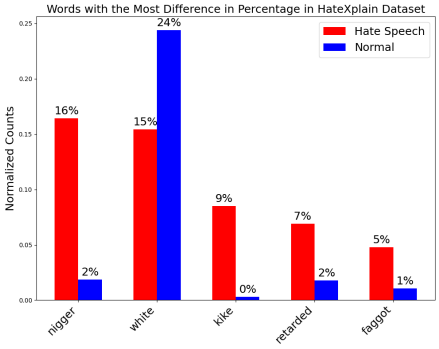

Figure 1: Top 10 Words with the Highest Percentage of Occurrence in One of the Two Classes, Ordered by the Greatest Percentage Difference Between Hate Speech and Normal Posts.

4. Validation

To assess our model's performance, we surveyed similar approaches to fine-tuning BERT for hate speech detection. Kumar et al. (2020) [9] presented a BERT-based model, and a comparison of their reported accuracy with ours is shown in Table 1 following their hyperparameters. We use their same dataeset Hasoc [12] collected for the official competition of Hate Speech and Offensive Content Identification in multiple languages. Notably, our model achieves results comparable to the work of [8] for multi label classification (hateful, offensive, normal) despite their more intricate approach: They computed the Ground truth attention where they convert each rationale into a an attention vector to reduce model-specific biases and highlight common patterns. Also, they employed LIME [13] (Local Interpretable Model-Agnostic Explanations) as explainability technique to highlight important words for classification.

	HateXplain		Hasoc	
	Reported	Ours	Reported	Ours
Accuracy	69%	69.49%	–	71.87
F1 Score	0.674	0.678	0.5031	0.71

Table 1: Comparison of our approach to other works

4.1. Curated Dataset

We observe that Hate Speech detection task is data-driven, therefore we collect a large dataset consisting of HateXplain, hasoc used before and in addition:

- Hate Speech 18 [14]: the text was extracted from a white

supremacist forum. Several subforums have split into sentences. Those sentences have been manually labelled as containing hate speech or not.

- Hate Speech Offensive [11]: They used a crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords

Dataset	#Train	#Validation	#Test	Accuracy %	F1
HateXplain	15,383	1,924	1,924	78.84	0.778
hate_speech18	6,421	2,141	2,141	90.98	0.706
hate_speech_offensive	14,869	4,957	4,957	71.049	0.69
hasoc	3,337	371	814	86.76	0.71
Total	40,010	9,391	9,836	87.637	0.870

Table 2: Summary of the curated dataset

Results on the curated dataset Fine tuning the model for binary classification of hate sepech, seperately on each dataset achieved the results 2 on their each train, validation, test sets. Training our model on the entire collected dataset yielded an overall accuracy of 87.64%:83.75% for hate speech detection and 90.2% for normal instances. F1 score is 0.87.

These results suggest that our approach, which involves training on a large and diverse dataset, contributes to a more robust model. While the model exhibits slightly higher accuracy on normal instances, the balanced F1 score indicates that the classification is not overly biased towards the majority class despite potential class imbalances in the dataset

4.2. Misclassified examples

When going through the samples that were misclassified by the model, we observed that:

1. *Labeling Inconsistencies*: During our analysis, we identified several instances where the dataset's annotations appeared inconsistent with the actual content. For example, the phrase "I hope the filthy scumbags who did this get what they deserve" was labeled as "normal" despite its clear expression of hostility. We found numerous similar examples where language was undeniably rude and arguably hateful, yet had been labeled otherwise.
2. *Model learning shortcut features*: Certain words act like trigger words that make the model automatically classify it as hateful. These features are often correlated with the target label (hateful or not hateful) but they don't actually represent the true meaning and context. To discover this, we analyzed the most frequent n-grams in both "hateful" and "normal" classes after filtering out stop words. By ranking these n-grams and cross-referencing them with their occurrences and corresponding labels in the original dataset, we identified discrepancies where frequent n-grams were associated with incorrect classifications. This highlighted instances where the model appeared to be relying on these n-grams as predictors without fully understanding the context, indicating the presence of shortcut features.

In hate speech detection, these non-robust features of the fine tuned BERT model s can manifest in a few ways:

Specific Words as Triggers: Models might learn that the presence of certain words like "nazi", "trump" almost always indicates hate speech. While these words are indeed often used in hateful contexts, they can also appear in non-hateful situations for example historical text or neutral context. Also, models might pick up on stylistic patterns commonly found in hateful content, such as excessive capitalization, punctuation, or the use of certain emojis. However, these patterns can also be used in non-hateful ways for emphasis or expression. We could have performed more intricate pre-processing or post-processing but this would change the original input. For example, the removal of emojis could cause a great loss in context.

Topic Bias: If a model is trained on data that heavily associates certain topics (e.g., politics, religion) with hate speech, it might learn to flag any mention of those topics as hateful, even if the discussion is civil or nuanced. Relying on shortcut features can lead to a few serious issues like more false positives, more false negatives and perpetuating bias. If shortcut features are correlated with certain groups or communities, the model can end up disproportionately flagging content from those groups as hateful, even if it's not. This reinforces harmful stereotypes and biases.

4.3. Mitigating these false shortcut features

More diverse and balanced training dataset, encompassing a wider range of topics and demographics, would likely enable the model to better discern the true intent behind messages, rather than relying solely on the presence of trigger words.

Adversarial Training: We could have added adversarial examples during the training with trigger words appearing more in a non hateful context. This would teach the model with carefully crafted examples that contain trigger words in non-hateful contexts. This helps expose the model's reliance on shortcuts and guides further improvement.

5. Conclusion

In conclusion, our BERT-based model demonstrates promising performance in hate speech detection, achieving results comparable to more complex state-of-the-art approaches. This effectiveness highlights the potential of simpler approaches in this domain. However, challenges remain in addressing biases and shortcut features within the training data. Curating a larger, more diverse dataset from multiple sources and high quality annotating could mitigate annotation bias and improve generalization to unseen data. Our future work would have focused more on making the model more robust by mitigating the non-robust features.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[2] G. Press, "Twitter statistics and facts," 2024. Accessed: 2024-05-22.

[3] G. Rosen, "Hate speech prevalence has dropped by almost 50% on facebook," October 17 2021. Meta, VP of Integrity.

[4] A.-D. League, "Online hate and harassment: The american experience 2021," 2021. Accessed: 2024-05-22.

[5] "What is hate speech?," 2024. Accessed: 2024-05-22.

[6] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *The Semantic Web. ESWC 2018. Lecture Notes in Computer Science*, vol. 10843, pp. 745–760, Springer, Cham, 2018.

[7] X. Huang, L. Xing, F. Dernoncourt, and M. J. Paul, "Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition," in *Proceedings of the Twelfth Language Resources and Evaluation Conference* (N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Marseille, France), pp. 1440–1448, European Language Resources Association, May 2020.

[8] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," 2022.

[9] A. Kumar and A. Jha, "Nitp-ai-nlp@hasoc-fire2020: Fine tuned bert for the hate speech and offensive content identification from social media," in *Proceedings of the FIRE 2020 - Forum for Information Retrieval Evaluation*, CEUR-WS.org, 2020.

[10] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[11] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pp. 512–515, 2017.

[12] A. Kumar, S. Saumya, and J. Singh, "Nitp-ai-nlp@hasoc-fire2020: Fine tuned bert for the hate speech and offensive content identification from social media," 12 2020.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[14] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate Speech Dataset from a White Supremacy Forum," in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, (Brussels, Belgium), pp. 11–20, Association for Computational Linguistics, Oct. 2018.