Addressing Class Imbalance in Hate Speech Detection Using Data Augmentation, Dowsnampling and Class Weighting

Parijat Chakraborty
Department of Statistics
University of Michigan
Email: cparijat@umich.edu

***Abstract*—Hate speech detection is a critical task in natural language processing, often hindered by imbalanced datasets. This project addresses the issue using the `tdavidson/hate_speech_offensive` dataset and BERT. A combination of contextual data augmentation, majority class downsampling, and class weighting significantly improves the model's performance. Experimental results demonstrate superior performance with a reduced bias toward the majority class compared to initial attempts without preprocessing.**

I. Introduction

Hate speech detection is an increasingly significant challenge in natural language processing (NLP), playing a crucial role in mitigating online abuse. This task has gained even more relevance in the current post-election phase, marked by heightened tensions and growing unrest due to ongoing conflicts and wars. However, the scarcity of hate speech data poses a major challenge, as the majority of online comments are non-hateful. Additionally, distinguishing between explicit hate speech and comments containing offensive language remains a nuanced problem.

The [Fac] dataset from Hugging Face exemplifies this challenge, with 76% of samples labeled as offensive, 19% as neutral, and only 5% as hate speech, highlighting a severe imbalance.

This project addresses the imbalance using a combination of data augmentation, majority class downsampling, and class weighting. Recent advancements in hate speech detection have focused on addressing data imbalance and enhancing model performance through various strategies. [Jah+24] conducted a comprehensive study on natural language processing (NLP) data augmentation techniques for hate speech detection, evaluating methods such as back-translation and BERT-based contextual augmentation. Their findings indicate that BERT-based contextual synonym replacement offers greater sentence diversity, albeit with higher label alteration rates. In another study, [Mna+22] proposed BERT-based ensemble learning models to improve hate speech detection. By combining BERT with deep learning models like Bi-LSTM and Bi-GRU, their method demonstrated the effectiveness of ensemble methods in this domain. Additionally, [LXV20] introduced Dager, a generation-based data augmentation method aimed at enhancing offensive language detection. Dager extracts lexical features of specific classes to guide a conditional generator built on GPT-2, producing augmented data that improves F-1 score with minimal use of the original dataset.

These studies underscore the importance of integrating data augmentation to mitigate class imbalance and enhance the robustness of hate speech detection models. Our project aims to address this aspect of imbalanced data using simpler techniques and only the BERT model. We demonstrate significant improvements over the initial proposal metrics, underscoring the effectiveness of the proposed methodology while using simple data preprocessing techniques.

II. Methodology

A. Dataset and Preprocessing

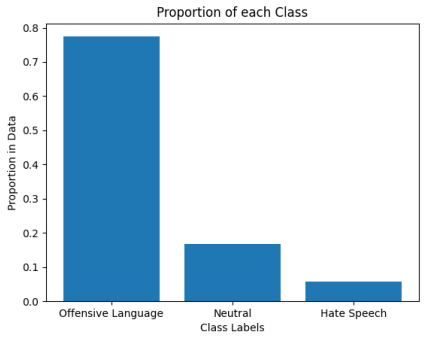
The dataset was split into 80% training and 20% testing subsets. Offensive samples were downsampled to 10,000, while hate and neutral samples were augmented to 4,000 each using `ContextualWordEmbsAug()` from the `nlpaug` library [Ma]. Thee dataset contains tweet data so it has urls and userids included in the tweets. We clean the data by removing this to ensure higher classification quality and faster processing. Fig. 1 shows the barplots of class distribution before and after preprocessing.

B. Model Training

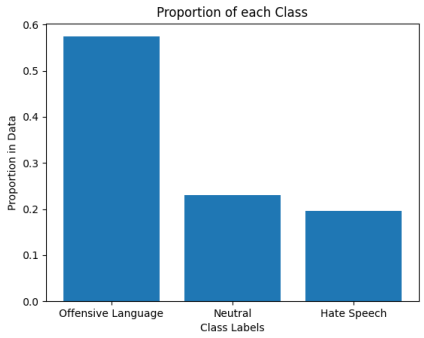
The fine-tuned BERT model incorporated class weights during training, proportional to the inverse of class frequencies. This strategy mitigates the dominance of the offensive class and promotes equal treatment across classes. To reduce overfitting further we tune several parameters of BERT: increase `weight_decay`, `hidden_dropout_prob` slightly and decrease `learning_rate` to $4e-5$ from the usual $5e-5$.

C. Evaluation Metrics

Performance was evaluated using a confusion matrix and standard classification metrics, including accuracy, precision, recall, and F1-score.



(a) Class distribution before preprocessing.



(b) Class distribution after preprocessing.

Fig. 1: Comparison of class distributions before and after preprocessing.

III. Results

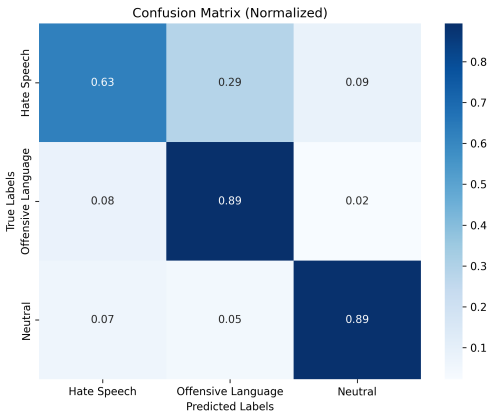
Fig. 2a presents the final confusion matrix after implementing the described methodologies. Compared to the proposal results (Fig. 2b), the model shows reduced bias, particularly for the detection of hate speech. This is very close to the confusion matrix given in [Dav+17], the first paper for this dataset. This goes to show that given proper preprocessing, a transformer-based model like BERT can function equally well as classical methods.

IV. Conclusion

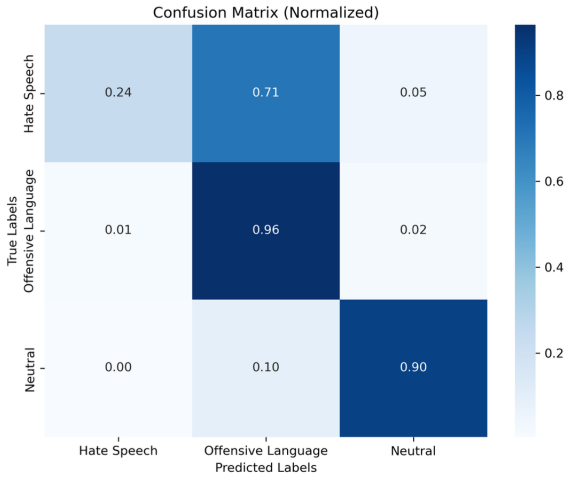
This project demonstrates the effectiveness of addressing data imbalance using contextual augmentation, downsampling, and class weighting. The proposed approach achieves sufficiently good generalization across classes compared to training without preprocessing. Future work will explore the integration of additional data augmentation techniques and model ensembles to further enhance performance.

References

[Dav+17] Thomas Davidson et al. "Automated hate speech detection and the problem of offensive language". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 11. 1. 2017, pp. 512–515.



(a) Confusion matrix after implementing proposed methodology.



(b) Confusion matrix from the initial proposal.

Fig. 2: Comparison of before and after implementing Proposed Methodology

[Fac] Hugging Face. *tdavidson/hate_speech_offensive*. https : / / huggingface . co / datasets / tdavidson / hate_speech_offensive. Accessed: 2024-12-09.

[Jah+24] Md Saroar Jahan et al. "A Comprehensive Study on NLP Data Augmentation for Hate Speech Detection: Legacy Methods, BERT, and LLMs". In: *arXiv preprint arXiv:2404.00303* (2024).

[LXV20] Ruibo Liu, Guangxuan Xu, and Soroush Vosoughi. "Enhanced offensive language detection through data augmentation". In: *arXiv preprint arXiv:2012.02954* (2020).

[Ma] Edward Ma. *nlpaug Library*. https://github.com/makcedward/nlpaug. Accessed: 2024-12-09.

[Mna+22] Khouloud Mnassri et al. "BERT-based ensemble approaches for hate speech detection". In: *GLOBECOM 2022-2022 IEEE Global Communications Conference*. IEEE. 2022, pp. 4649–4654.