Members: Jimmy Le
Course: W266 - Natural Language Processing
University of California, Berkeley

Benchmarking Transformer Models for Hate Speech Detection Directed Towards The LGBT Community on Social Media

Abstract

Social media platforms, such as X (formerly Twitter), frequently host content including hate speech targeting marginalized communities, notably the LGBT community. Detecting and mitigating this hate speech is crucial for fostering safer online environments. This paper benchmarks several transformer-based models—RoBERTa-base, HateBERT, BERTweet, and XLM-RoBERTa—in detecting LGBT-targeted hate speech using the HateXplain dataset, specifically filtered to tweets/posts targeting homosexual individuals. By comparing both general-purpose and domain-specific transformer models, we identify key factors influencing performance, such as pretraining corpora, architectural nuances, and domain-specific fine-tuning.

Our experiments demonstrate that domain-specific fine-tuning significantly enhances model performance compared to naive baselines. HateBERT achieves the highest macro F1-scores, effectively capturing subtle linguistic cues like slurs, sarcasm, and aggression due to its specialized training on offensive Reddit content. BERTweet also shows strong results due to its pretraining on extensive English Twitter data, adeptly handling social media-specific language such as slang, abbreviations, and hashtags. Conversely, general-purpose models like XLM-RoBERTa, despite robust multilingual capabilities, perform less effectively in this specialized domain.

These results illustrate the importance of targeted domain adaptation and provide practical insights for developers and researchers aiming to improve hate speech detection tools tailored specifically to the LGBT community. Our findings contribute a clear roadmap for future development of specialized hate speech detection models, emphasizing the value of tailored pretraining and fine-tuning strategies.

Introduction

Social media platforms like X (formally known as Twitter) hosts a broad spectrum of conversations, including hate speech directed toward the LGBT community. Hate speech is commonly defined as any communication that belittles a person or a group based on some characteristic such as race, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics (Nockleby, 2000). As the spread of online hate speech continues to grow, its detection on social media has gained increasing significance (Schmidt and Wiegand, 2017).

Previous studies, such as Yigezu et al. (2023) have demonstrated mixed success using general-purpose transformer architectures like RoBERTa and BERTweet; however, domain-specific fine-tuning and pretraining strategies have been shown to improve models' ability to detect targeted hate speech. HateBERT, when fine-tuned on hate speech corpora, has indicated that domain-specific training significantly improves model performance (Caselli et al. 2021). While general hate speech detection has been widely studied, there is limited research that identifies transformer models specialized in detecting hate speech targeting the LGBT community.

In this paper, we contribute to the existing body of literature by:
1. Benchmarking the performance of various models and their ability to detect hate speech directed towards LGBT community using the HateXPlain Dataset - specifically with RoBERTa-base, HateBERT, BERTweet, and XLM-RoBERTa
2. Comparing the effectiveness of these models based on their pre training strategies, and architectures
3. Identifying the most effective models for this domain and analyzing architectural and training factors that contribute to their performance.

To advance this papers' contribution, we suggest developing a model leveraging insights gained from transformer architectures with targeted fine-tuning on LGBT-specific hate speech corpora. Such a model would aim to bridge existing gaps by creating a specialized architecture that is skilled in detecting LGBT-specific hate speech. Enhancing this accuracy is a small, but meaningful step in building a specialized hate speech detector.

Objective

Benchmark performance of various models and their ability to identify hate speech directed towards LGBT community using the HateXPlain Dataset. Ultimately, the goal is to identify which models worked best, and analyze their architecture or corpus to understand why they may have performed better than others. The models will be compared via metrics such as F1-Score, Precision and Recall.

Background

Literature Review

Yigezu et al. (2023) explored transformer-based architectures for hate speech detection in online messages directed towards the Mexican Spanish-speaking LGBTQ+ population. Their study identifies common challenges such as data imbalance, order bias, and insufficient training samples which can hinder model generalization. To address these limitations, the authors advocate for preprocessing enhancements and oversampling methods to improve data representativeness. They ran two experiments - one with RoBERTa, and second with BERT which achieved an F1 score of 79.59% and 67.33% respectively which led to the authors' conclusion that transformer-based approaches are effective in identifying hate speech directed

at the LGBT+ community. From this research, we will include RoBERTa in the benchmark since it achieved better performance in addition to measuring performance with F1-Scores.

In *A System for Detecting Abusive Contents Against LGBT Community Using Deep Learning Based Transformer Models* (Manikandan et al., 2022), the authors noted that XLM-RoBERTa outperformed BERT in detecting hate speech that targeted the LGBT community for precision, recall, and f-measure. Unsupervised Cross-lingual Representation Learning at Scale (Conneau, A, et al., 2020) mentions XLM-R performed better than multi-lingual BERT which we hypothesize can be better at detecting hate speech towards our target community.

Nguyen et al., 2020 introduced BERTweet that is trained on 850 million tweets. Since the source of our data are Tweets, BERTweet can be a valuable model choice to fine-tune and benchmark.

Methodology

Dataset

This paper leverages the **HateXplain** dataset, a widely used benchmark for explainable hate speech detection introduced by Mathew et al. (2021). The dataset comprises 20,148 posts from Twitter and Gab, each annotated by three annotators who assign a classification label—normal, offensive, or hate speech—and identify the targeted group. In addition to labels and targets, HateXplain includes rationales that justify each annotation, enhancing its utility for interpretable model development. Since the goal is to benchmark each model's ability to detect hate speech towards the LGBT community, the dataset is filtered such that the targets are only homosexuals.

Baseline Models

Two baseline models are used to establish a naive performance benchmark and define a lower bound for comparison. The **Most Frequent Class** baseline uses the most common label from the training set as predictions. The **RoBERTa-Base Zero-Shot** baseline utilizes the pretrained RoBERTa model without any fine-tuning, using zero-shot classification to predict labels directly.

Fine-Tuned Models

Fine-tuning adapts pretrained models using a small set of high-quality labeled data to improve their performance on specific tasks. These models (identified below) play a key role in our benchmarking experiments.

RoBERTa (Liu et al., 2019) serves as our general-purpose baseline due to its robust performance across NLP tasks, benefiting from improved training strategies and hyperparameter optimization.

HateBERT (Caselli et al., 2021) pretrained on offensive Reddit comments, captures hate speech nuances missed by general models and is especially suited for detecting targeted language such as LGBT-related hate.

BERTweet (Nguyen et al., 2020) pretrained on approximately 850 million tweets. Traditional

transformer models are typically pretrained on formal textual sources, and fail to adequately handle Twitter's abbreviations, informal language, slang, and hashtags. Since our data is composed of Tweets, BERTweet can potentially be a better model choice.

XLM-R (Conneau et al., 2020) is a transformer-based multilingual masked language model pre-trained on text in 100 languages, which obtains state-of-the-art performance on cross-lingual classification, sequence labeling and question answering.

Hyperparameter Tuning

To optimize model training, we utilize Optuna library for hyperparameter tuning, adjusting learning rate, batch sizes, number of epochs, and weight decay. After fine-tuning and benchmarking each model on the HateXplain dataset, we analyze their architectures to understand performance differences.

Results and Discussion

model	Zero-Shot RoBERTa	Most Frequent Class
normal_precision	0.473684	0.000000
normal_recall	0.461538	0.000000
normal_f1-score	0.467532	0.000000
normal_support	78.000000	54.000000
offensive_precision	0.000000	0.000000
offensive_recall	0.000000	0.000000
offensive_f1-score	0.000000	0.000000
offensive_support	54.000000	26.000000
hatespeech_precision	0.146341	0.493671
hatespeech_recall	0.461538	1.000000
hatespeech_f1-score	0.222222	0.661017
hatespeech_support	26.000000	78.000000
accuracy	0.303797	0.493671
weighted avg_precision	0.257926	0.243711
weighted avg_recall	0.303797	0.493671
weighted avg_f1-score	0.267375	0.326325

model	RoBERTa	HateBERT	BERTweet	XLM-RoBERTa
overall_loss	0.853527	0.882770	0.860650	0.924214
eval_model_preparation_time	0.005200	0.005100	NaN	NaN
overall_f1	0.472488	0.598392	0.576210	0.484849
overall_precision	0.481040	0.614217	0.593028	0.483318
overall_recall	0.474684	0.601266	0.575949	0.493671
overall_accuracy	0.474684	0.601266	0.575949	0.493671
normal_f1	0.468085	0.586957	0.586957	0.595041
normal_precision	0.550000	0.710526	0.710526	0.537313
normal_recall	0.407407	0.500000	0.500000	0.666667
offensive_f1	0.346154	0.509804	0.436364	0.260870
offensive_precision	0.346154	0.520000	0.413793	0.300000
offensive_recall	0.346154	0.500000	0.461538	0.230769
hatespeech_f1	0.517647	0.635838	0.615385	0.483221
hatespeech_precision	0.478261	0.578947	0.571429	0.507042
hatespeech_recall	0.564103	0.705128	0.666667	0.461538
epoch	NaN	NaN	2.000000	2.000000

Figure 1: Baseline vs Fine-Tuned Models

The fine-tuned models exhibit significantly improved metrics across all classes compared to the baselines. HateBERT performs best overall with macro F1-scores above 0.59, achieving strong hatespeech F1 (0.63) and much better offensive F1 (0.51) than any baseline. RoBERTa and BERTweet also outperform zero-shot settings, with more balanced precision-recall trade-offs. For example, XLM-RoBERTa's hatespeech F1 increases from 0.22 to 0.52, and it shows offensive class recognition with 0.35 F1, which was completely missed in zero-shot.

These improvements are attributed to domain adaptation during fine-tuning, which allows models to internalize class-specific linguistic cues and context-dependent semantics that general-purpose or naive models overlook.

Model [Token Method]	Performance		
	Acc.↑	Macro F1↑	AUROC↑
CNN-GRU [LIME]	0.627	0.606	0.793
BiRNN [LIME]	0.595	0.575	0.767
BiRNN-Attn [Attn]	0.621	0.614	0.795
BiRNN-Attn [LIME]	0.621	0.614	0.795
BiRNN-**HateXplain** [Attn]	0.629	0.629	0.805
BiRNN-**HateXplain** [LIME]	0.629	0.629	0.805
BERT [Attn]	0.690	0.674	0.843
BERT [LIME]	0.690	0.674	0.843
BERT-**HateXplain** [Attn]	**0.698**	**0.687**	**0.851**
BERT-**HateXplain** [LIME]	**0.698**	**0.687**	**0.851**

Figure 2: HateXPlain benchmarks

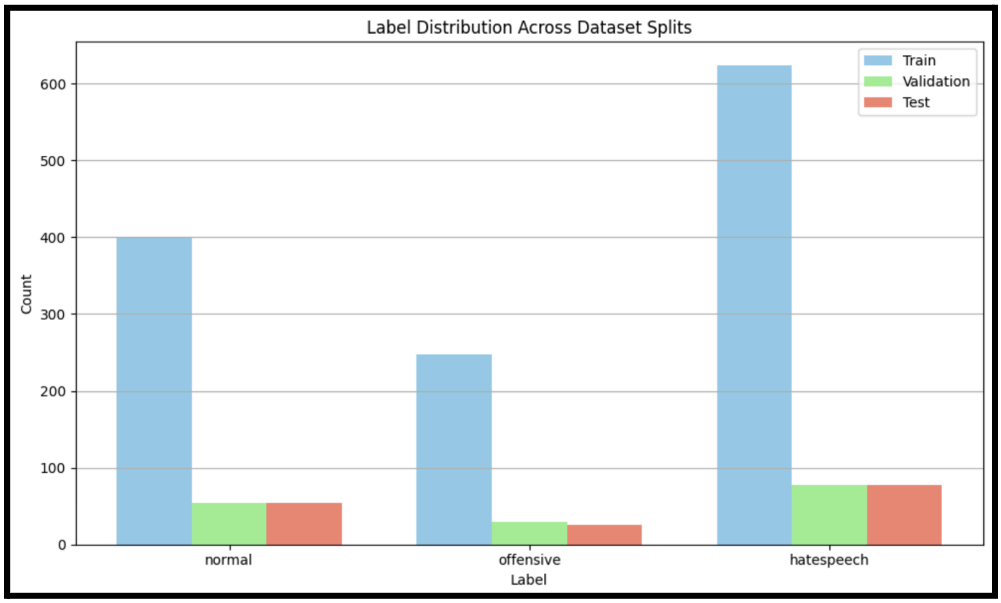
When comparing our fine-tuned models to HateXplain benchmarks, we analyze that our metrics are not as performant as BERT-HateXplain [LIME]. When comparing the Hatespeec F1-score HateBERT with a score of 0.635 is lower than their BERT-HateXplain [LIME] which means our model is still capturing hate speech well compared to the existing benchmark. This lower score in our models are discussed in the Limitations section.

Analyzing Architectures

HateBERT's superior performance stems from its domain-specific pretraining on offensive Reddit content, enabling it to capture hate-related nuances like slurs, sarcasm, and aggression. Its BERT-based architecture benefits from strong contextual encoding, while the added hate speech exposure improves task-specific sensitivity. This ultimately made it the best model choice in our selection of models.

BERTweet outperforms XLM-RoBERTa likely due to its pretraining on 850M English tweets, making it highly specialized to handle the Twitter data in our dataset. Its tokenizer handles informal language, hashtags, abbreviations, and slang typical of social media, which aligns closely with the HateXplain dataset. While XLM-RoBERTa benefits from multilingual robustness, BERTweet's domain-specific training gives it an edge in capturing the linguistic nuances of English hate speech on Twitter. Its architecture, based on RoBERTa, further supports strong contextual understanding, enhancing performance in social media-specific classification tasks.

Limitation - Class Imbalance



This paper used the original train/validation/test split provided by the HateXplain authors, which includes an imbalanced class distribution. Yigezu et al. (2023) suggested methods such as more preprocessing of data and oversampling of underrepresented classes. Due to time constraints, we did not apply such techniques to address the overrepresentation of the *hatespeech* class, which may have introduced model bias.

Limitation - Data Preprocessing

This study did not apply thorough text preprocessing, which may have affected model performance. Xu et al. (2022), who also used the HateXPlain dataset, implemented preprocessing techniques recommended by Pérez et al. (2021), including the removal of URLs, emojis, and user mentions. Incorporating similar strategies could have improved the quality of input data and potentially led to better model benchmarks.

Conclusion

This research can guide developers and researchers in selecting and fine-tuning transformer models to improve detection of hate speech targeting the LGBT community, particularly on social media platforms. It provides empirical evidence on model effectiveness and offers insights into how domain-specific training enhances classification accuracy.

Most existing hate speech models are general-purpose and lack sensitivity to the nuance and subtleties in LGBT-targeted hate speech, especially in informal, online contexts. This work attempts to addrtess that gap by benchmarking specialized transformer models, identifying which architectures and pretraining strategies are most effective for LGBT-specific hate speech detection, and providing a roadmap for future model development in this critical domain.

References

Caselli, Tommaso, et al. "HateBERT: Retraining BERT for Abusive Language Detection in English." arXiv preprint arXiv:2010.12472, 2021.

John T Nockleby. 2000. Why Internet Voting. Loy. LAL Rev., 34:1023.

Liu, Yinhan, et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv preprint arXiv:1907.11692, 2019.

Manikandan, D., Subramanian, M., & ShanmugaVadivel, K. (2022). A system for detecting abusive contents against LGBT community using deep learning based transformer models. *CEUR Workshop Proceedings*, 3395, 106–116

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherjee, A. (2021). *HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection*. Proceedings of the AAAI Conference on Artificial Intelligence, 35(17), 14867–14875. https://arxiv.org/abs/2012.10289

Nguyen, Dat Quoc, Thanh Vu, and Anh Tuan Nguyen. "BERTweet: A pre-trained language model for English Tweets." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (2020).

Pérez, J. M., Giudici, J. C., & Luque, F. (2021). *pysentimiento: A Python toolkit for sentiment analysis and SocialNLP tasks*.

Schmidt, A., & Wiegand, M. (2017). *A Survey on hate speech detection using Natural Language Processing.* In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (pp. 1–10). https://aclanthology.org/W17-1101/

Xu, J., & Weiss, Z. (2022). *How much hate with #china? A preliminary analysis on China-related hateful tweets two years after the COVID pandemic began*. arXiv. https://arxiv.org/abs/2211.06116arXiv

Yigezu, M. G., Kolesnikova, O., Sidorov, G., & Gelbukh, A. (2023). Transformer-based hate speech detection for multi-class and multi-label classification. *CEUR Workshop Proceedings*, 3496, 1–10.