CSE 3000 Group 6
Nikhil Ghosh (nig14014)
Jacob Goldstein (jmg21035)
Avery Hogrefe (amh21022)
Robert Cameron Stawarz (rcs20006)

Technical Report: An AI Content Moderator

Abstract

This study investigates the application of transformer-based AI models for content moderation in online platforms. We fine-tuned a DistilBERT model on a dataset of 24,783 tweets pre-classified as offensive, hate speech, or neither, achieving an F1 score of 98.0% after three epochs of training. Despite these promising metrics, our analysis reveals significant limitations in the model's ability to generalize, likely due to dataset imbalance and limitations in capturing linguistic nuance. We identified false negatives for transphobic and xenophobic content, highlighting the risks of over-reliance on automated systems. The research underscores fundamental challenges in defining offensive content across diverse contexts and adapting to evolving language norms. We conclude that while AI can enhance moderation efficiency, it should serve as an assistive tool for human moderators rather than a replacement, with AI flagging obvious violations while human judgment addresses nuanced cases and evolving social contexts. This work contributes to the broader understanding of practical and ethical considerations in developing responsible AI moderation systems.

Introduction

One potential use of AI technology is online content moderation - filtering offensive or hateful speech on platforms like social media or chat rooms. Traditionally, these spaces have relied on reporting systems, word filters, or human moderators. Our project explores an AI classifier's ability to perform this task, with emphasis on the practical and ethical flaws such systems may introduce. We are particularly concerned with how a classifier might reinforce existing biases.

Methods

We downloaded a tweets dataset (n=24,783) from Davidson et al. 2019, with offensive, hate speech, or neither labels. We engineered a binary classification feature by combining hate speech and offensive into a positive (n=20,620) and using the "neither" as negative (n=4163).

We also imported the encoder-only transformer DistilBERT, a distilled version of Google's BERT base model created by HuggingFace. Specifically, we used the uncased model, which doesn't distinguish between "hello" and "Hello", along with HuggingFace's AutoTokenizer. We configured Pytorch to minimize computational costs on our hardware, and we set device batch sizes small (8 for training, 16 for testing) with a gradient accumulation of 4 steps to minimize

device load. We used an 80/20 ratio to split the data into train and test sets. We used standard weight decay of 0.01 and warmup ratio of 0.1. We chose F1 score as our optimization metric and ran Transformers fine-tuning for 3 epochs.

Results

Table 1 shows evaluation metrics logged at every 200 steps, where each epoch was 619 steps. Accuracy, precision, and recall metrics are shown in addition to the F1 optimization metric, which started at 96.7% and ending at 98.0%,. Figure 1 plots Table 1's data and shows a flat plateau, suggesting that the high performance scores are misleading. The validation loss increased slightly, suggesting some overfitting or lack of generalization.

Step	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall	True Positives	False Negatives
200	0.156500	0.126252	0.945330	0.966440	0.987598	0.946169	3902	222
400	0.123600	0.118114	0.961872	0.976909	0.984487	0.969447	3998	126
600	0.108800	0.105777	0.955417	0.972853	0.985810	0.960233	3960	164
800	0.079300	0.113170	0.963688	0.978271	0.974038	0.982541	4052	72
1000	0.076500	0.104679	0.964495	0.978479	0.986926	0.970175	4001	123
1200	0.076100	0.102673	0.965907	0.979468	0.981495	0.977449	4031	93
1400	0.038900	0.123743	0.966109	0.979631	0.979631	0.979631	4040	84
1600	0.040900	0.129087	0.967319	0.980359	0.980359	0.980359	4043	81
1800	0.039500	0.126747	0.967319	0.980349	0.980825	0.979874	4041	83

Table 1: Model Evaluations During Training

Figure 1: Model Performance Scores over Training

Discussion

The model's inflated performance metrics may stem from a skewed dataset. With nearly 5x as many offensive tweets as neutral, can make F1 score a misleading metric. Borderline tweets and limited language diversity can also make the model overfit on superficial patterns, leading to high apparent performance but low generalized robustness.

Improving future models requires addressing both data quality and model design. While sourcing a better-balanced and more diverse dataset would be ideal, techniques like data augmentation - such as paraphrasing or synonym substitution - could help the model learn more robust semantic features beyond repeated patterns or trending language. Label smoothing or semi-supervised learning could reduce the impact of ambiguous or borderline examples. A multi-class classification scheme might also capture more nuance between offensive, neutral, or mixed-content messages. Incorporating context-aware architectures, such as models fine-tuned on conversation threads or user history, could further improve predictions by leveraging discourse-level cues. Finally, systematic error analysis and cross-validation would reveal specific weaknesses, enabling more targeted improvements.

Testing the ethical implications of a model is difficult; beyond performance, the domain of content moderation demands careful scrutiny. Our model achieved strong scores on validation data, but we identified notable errors. For example, the message “Being trans/gay is a mental illness”was labelled non-offensive - a troubling false negative that risks enabling transphobic or homophobic speech. It also misclassified “Go back to your country” as neutral. Such errors are harmful: they can alienate users, amplify hate speech, and erode platform trust. Conversely, false positives - where innocuous speech is flagged - can suppress legitimate expression and drive users toward evasive tactics, as seen on platforms like TikTok where moderation policies have spurred the rise of coded language. This indicates the inherent risk of relying on AI.

These challenges highlight a deeper issue: what exactly defines offensive content? Is discussing marginalized groups, addiction, or self-harm inherently offensive? Should criticism of racism or police brutality be censored? The definition of offensive language is subjective, context-dependent, and constantly evolving. Even a flawless model today may fail tomorrow as social norms shift. For example, the word “retarded” was originally created as a medical term before becoming pejorative (MentalHealth.com). Any effective content moderation system - AI or otherwise - must adapt to these changes. Yet many current models, including ours, lack the nuance and flexibility to navigate such linguistic and cultural shifts.

Conclusion

Based on this, we recommend that AI is used as an aiding tool for the human moderator, not as a replacement. We cannot guarantee the absence of major false positives and negatives, and we need constant adaptability to keep up with social and linguistic change. Thus, the ideal scenario is for the AI moderator to assist the human moderator, perhaps making suggestions, pointing out areas of concerns, and blocking the most clear and obvious examples of offensive language. The risks of completely relying on automated moderation means that we should use human judgment in conjunction with the power of AI.

References

Our model and training process are open-source on our GitHub repository https://github.com/uconn-cse3000-team6/content-moderation

T. Davidson, D. Bhattacharya, and I. Weber, “Racial Bias in Hate Speech and Abusive Language Detection Datasets,” *arXiv:1905.12516 [cs]*, May 2019, Accessed: Mar. 2025. [Online]. Available: https://arxiv.org/abs/1905.12516

T. Davidson, D. Warmsley, M. Macy, and I. Weber. "Automated Hate Speech Detection and the Problem of Offensive Language." ICWSM. 2017. Available: https://github.com/t-davidson/hate-speech-and-offensive-language/

“distilbert/distilbert-base-uncased · Hugging Face,” *huggingface.co*, Mar. 11, 2024. https://huggingface.co/distilbert/distilbert-base-uncased

MentalHealth.com, “History of Stigmatizing Names For Intellectual Disabilities Continued,” May 2024, Accessed: Apr. 2025. [Online]. Available: https://www.mentalhealth.com/library/history-stigmatizing-names-intellectual-disabilities-continued