Leveraging LLMs and Topic Modeling for Toxicity Classification

Christina Chance, Claire Huang,
Haniyeh Ehsani Oskouie, **Elizabeth Eyeson**, **Margaret Capetz**
cchance, clairehuang1, haniyehehsani, eeyeson, mcapetz17 at g.ucla.edu

Abstract

Content moderation and toxicity classification are major tasks due to social implications. However, work has shown that bias amplification and the suppression and marginalised communities occurs with major toxicity classification model. Much work suggests that the positionalitiy of annotators influence the gold standard labels in which the models learned from propagate annotators bias. To further assesses the impact of positionality during the annotation process, we explore topic modeling strategies for content moderation. We compare the results of fine-tuning on BERTweet and HateBERT with the full dataset and three separate topics generated by LDA topic modeling. We found that fine-tuning on specific topics appears to have improved the F1 score of the models.

1 Introduction

Content moderation is important to mitigating the spread of potentially harmful content like hate speech, self-harm or harassment on social media platforms. Without effective moderation, users risk being exposed to psychological harm or perpetuating harm itself. Thus, upholding civility, psychological safety and inclusivity in social media interactions depends upon robust content moderation mechanisms. This is important in our increasingly digital world.

Popular toxicity classification and moderation techniques tend to rely on human annotations due to limitations in automated labeling. However, such annotations can amplify bias due to annotaters' identities, lived experiences, societal/cultural norms, and personal beliefs influencing their annotations. This subjectivity can inadvertently perpetuate stereotypes and marginalization in datasets and thus impact the performance of machine learning models.

Our goal is to explore the following question: How does fine-tuning BERTweet and HateBERT on topic-specific subsets of the NLPositionality dataset compared to the full dataset affect model performance?

The NLPositionality Toxicity and Hate Speech consists of labeled toxic tweets and annotator demographic metadata. This dataset was derived from the paper "NLPositionality: Characterizing Design Biases of Datasets and Models" [Santy et al., 2023], which introduced a framework for characterizing design biases and quantifying the positionality of NLP datasets and models.

We propose a topic modeling-enhanced fine-tuning approach on the NLPositionality Toxicity and Hate Speech dataset. We utilize topic modeling to reveal latent themes across toxic and non-toxic content. By analyzing these themes, we found underlying patterns and topics prevalent in tweets labeled as toxic vs. non-toxic. We then fine-tuned BERTweet, a specialized BERT model for short tweet data, and HateBert, a finetuned BERT model for the task of toxicity classification, to learn embeddings and neural representations that capture the key features influencing the classification of a toxic tweet. We fine-tuned both BERTweet and HateBert on subsets of the data based on the topic clustering. We found that when fine-tuned on specific topics compared to the entire dataset, the

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

models' performance for the classification of certain topics in the dataset increased. By leveraging the capabilities of BERTweet and HateBert, we were able to extract nuanced linguistic and contextual cues that contribute to toxicity classification.

2 Literature Review

The use of short form text, such as Tweets, has been a popular area of study as the limited context forces the model to find latent cues and patterns compared to longer text in which there is more context for classification. Many works, such as [Adhikari et al., 2022] and [Seeberger and Riedhammer, 2022] who used tweets for crisis analysis in a classification setting, leverage the accessibility of tweets as well as its ability to capture cultural and social sentiment at any given time, especially used in the content moderation domain.

Content moderation and toxicity classification is a sub-field that benefits from short formed tweets due to the formulaic behavior of common toxic language and content. Several studies such as [Biraj Dahal and Li, 2019] explore the use of author-pooled Latent Dirichlet Allocation (LDA) to extract discussion topics from Twitter data related to climate change. Similarly, Zosa et al. [2021] focus on comment moderation, utilizing a topic-aware model to enhance automatic moderation by incorporating semantic features from topic models. Both studies highlight the importance of leveraging topic modeling techniques to improve content classification and moderation strategies.

In a related context, Kim et al. [2020] delve into enhancing word embeddings with topical information for toxic content detection, showcasing the effectiveness of incorporating topic-specific data in classification tasks. This approach aligns with the findings of Deshpande et al. [2023], who evaluate the toxicity of a dialogue-based Large Language Model (LLM) using specific word embeddings and lexical baselines. These studies collectively underscore the significance of considering contextual factors and specialized features in toxicity assessment and content moderation.

Furthermore, Malmasi and Zampieri [2017] contribute to the field by exploring techniques for identifying hate speech on social media platforms while differentiating it from general profanity. Their focus on establishing baseline models through supervised classification methods resonates with the overarching theme of leveraging innovative approaches to address harmful content online.

3 Methodology

We explore the integration of various techniques to perform toxicity classification for content moderation. In brief, we explored the use of topic-modeling for data pre-processing prior to fine-tuning two different models. Utilizing topic-modeling, we split the data into three separate topics. We fine-tuned the models on different sets of data, including the full data, topic 1, topic 2, and topic 3. Then, we fine-tuned on these datasets with BERTweet and HateBERT.

3.1 Data Pre-processing

For topic modeling, we employed several preprocessing techniques, including sentence tokenization, stop word removal, and lemmatization, to enhance the quality of our input data. To handle the tokenization process for BERTweet and HateBERT models, we leveraged the convenience and flexibility of the AutoTokenizer feature offered by the Hugging Face library. By utilizing AutoTokenizer, we were able to seamlessly integrate these models into our workflow and ensure optimal tokenization accuracy and efficiency.

3.2 Topic Modeling

We performed Latent Dirichlet Analysis (LDA) [Blei et al., 2003] topic modeling using the training data and then applied the results to the test set. We used the Python library Tomotopy [Lee, 2022] to cluster the toxicity dataset into 3, 6, and 10 topics using LDA. We found that performing topic modeling when K is larger produces more insightful and expressive topics, as shown in Table 1. We ultimately decided to use $k = 3$ as we have a smaller dataset. Further subsetting would produce insignificant numbers that we cannot draw assumptions.

Table 1: Topic distribution produced by Tomotopy Latent Dirichlet Allocation

	Number of Topics	
3	6	10
0 : woman people wa always never 1 : people white get issue race 2 : people muslim like make want	0 : get gay love know make 1 : people race issue make without 2 : white wa men order would 3 : people subhuman muslim white black 4 : woman people always never take 5 : people like wa ha even	0: f*ck get people old white 1 : people know muslim non thankfully 2 : like would white black dwarf 3 : history people without man black 4 : people like white woman race 5 : ret*rd make back people immigrant 6 : woman always never take idiot 7 : wa people ha time started 8 : people wa really white million 9 : woman men get people like

3.3 Models

We utilized two pre-trained models from Hugging Face library: BERTweet [Nguyen et al., 2020] and HateBERT [Hartvigsen et al., 2022]. BERTweet is a model that was trained on a more general corpus of tweets, while HateBERT was trained on a more relevant corpus to our research: hate-speech related texts. According to Hugging Face, BERTweet is the first public large-scale language model pre-trained for English Tweets. The model was trained based on the RoBERTa pre-training procedure. The corpus consists of 850M English Tweets (16B word tokens 80GB): 845M Tweets streamed from 01/2012 to 08/2019 and 5M Tweets related to the COVID-19 pandemic. As BERTweet was fine-tuned on short form tweets, the goal was to leverage the models ability to perform a task on limited context. On the other hand, HateBERT is a English pre-trained BERT model developed by the collaboration of Microsoft Research, Allen Institute for AI, and Carnegie Mellon University. The model was obtained by training the English BERT base uncased model, fine-tuned on ToxiGen data. The goal of this model was to leverage its task specific context and its learned ability of understand implicit toxicity to be able to generalize on more explicit examples..

3.4 Fine-tuning

During the fine-tuning process, we used layer-freezing. Freezing a layer in the fine-tuning process means it will not be trained, thus its weights will not be changed. This may be especially relevant for HateBERT since the learned weights are already relevant to our task. The potential benefits of layer freezing include reducing the risk of overfitting by preserving the generalization ability of the pre-trained model, saving computational resources and time, and preventing catastrophic forgetting by preserving the features learned by the pre-trained model. Thus, we fine-tuned the BERTweet and HateBERT models for toxicity classification on various data splits. For BERTweet, we froze the RoBERTa layers as we wanted to model to learn task specific information while style maintaining syntactical information. For HateBERT, we froze all layers except for the classification head since the model is already pre-trained on this task. Due to the fact that out task is multi-class classification and HateBERT is binary classification, we wanted to model to learn the final output scheme for our task. We found that freezing the hidden layers led to comparable results and significantly decreased model training time. In terms of other fine-tuning hyperparameters, we used the Adam Optimizer with a linear learning rate of 5e-5, 0 warm up steps, and 70 epochs. To ensure reliable results, each experiment was repeated five times using different manual seeds, enabling us to report both the mean and standard deviation of the outcomes.

4 Results

From the reported results, we see for BERTweet, training the model on individual topics improved the F1 score compared to training the model on the full dataset. The most significant improvement in the F1 score was from the full dataset to Topic 1, while the differences between the full dataset and Topic 2 and 3 are more marginal in comparison. For HateBERT, the Topic 1 data split had the highest F1 score, while Topic 3 had the lowest F1 score. Interestingly, for HateBERT, the model trained on the entire dataset performed second to best, while for BERTweet, the model trained on the full dataset performed the worst.

Table 2: F1 score for BERTweet with different seeds

Data split	Seed 0	Seed 1	Seed 2	Seed 3	Seed 9	Average	Stdev
Full data	0.4610	0.4631	0.4610	0.4560	0.4610	0.4604	0.0026
Topic 1	0.5588	0.5566	0.5566	0.5588	0.5588	0.5579	0.0012
Topic 2	0.4778	0.4778	0.4778	0.4778	0.4778	0.4778	0.0000
Topic 3	0.4659	0.4600	0.4659	0.4659	0.4600	0.4636	0.0032

Table 3: F1 score for HateBERT with different seeds

Data split	Seed 0	Seed 1	Seed 2	Seed 3	Seed 9	Average	Stdev
Full data	0.4831	0.4765	0.4837	0.4835	0.4852	0.4824	0.0034
Topic 1	0.5498	0.5498	0.5498	0.5498	0.5498	0.5498	0.0000
Topic 2	0.4767	0.4767	0.4767	0.4767	0.4767	0.4767	0.0000
Topic 3	0.4571	0.4572	0.4572	0.4572	0.4572	0.4572	0.0000

5 Discussion

5.1 Analysis

Majority voting was performed across seed runs to get predicted labels since the labels are categorical. Data subsets were grouped by gender and ethnicity and visualized using confusion matrices. Since positionality bias propagates through the form of incorrectly labeling examples as hate speech, we are interested in the true positive rate (TPR), true negative rate (TNR), false positive rates (FPR), as well as the precision and f1 score.

We see that both BERTweet and HateBERT have high FPR and TNR. For Black annotators especially, BERTweet has a high TPR and recall overall. It appears that the breakdown by topic did not yield notable differences in performance compared to the full dataset, suggesting that there is no one cluster that capture more latent and semantic features and information that influence model prediction. In Table 7 and Table 6, we assess positionality and model alignment for different demographics based on overall f1 score as well as TPR (recall). For data subset and demographic identifies with higher TPR and F1 scores, it suggest model alignment with positionality.

Table 4: BERTweet Statistics

Data split	Micro F1	Precision	Recall
Full data	0.549683	0.549683	0.549683
Topic 0	0.524249	0.524249	0.524249
Topic 1	0.557377	0.557377	0.557377
Topic 2	0.517162	0.517162	0.517162

5.2 Limitations

Topic modeling is not as expressive as necessary due to the variety in tweets. We utilized Tomotopy instead of Gensim as prior work had shown that Tomotopy has more expressive topics. We found that with a large number of topics, we had a more understandable grouping, but due to the size of the dataset as well as the knowledge that further subsets of the data would be too small for analysis, we decided to use a smaller number of topics. In addition, BERTweet was pre-trained on general tweets, which may not be specific enough for our downstream task of training for toxicity classification. Further, because HateBERT was pre-trained on a binary toxicity classification dataset, the inclusion of a third label for our dataset during the finetune process may have contributed to the high error rates for that new label.

Table 5: HateBERT Statistics

Data split	Micro F1	Precision	Recall
Full data	0.521494	0.521494	0.521494
Topic 0	0.519630	0.519630	0.519630
Topic 1	0.540984	0.540984	0.540984
Topic 2	0.501144	0.501144	0.501144

5.3 Future work

Our results demonstrate that models trained on the dataset split of Topic 1 stood out in terms of F1 score. In future work, we can further investigate why this is and look at the bias within the data splits performed by LDA topic modeling. Also, using the current NLPositionality dataset, we can test data splits of 6 and 10 topics to test if the topic granularity affects the F1 score and other statistics. Additionally, we can also pre-train on other datasets. Though we have already reasoned not to do this in practice because of the small data subset size due to the many topics, we can increase the corpus size and test our assumption. In general, training on a larger corpus may help improve the model. In addition, we can explore out-of-domain applications. We can test the model's generalizability to other datasets, such as data from other social media platforms beyond Twitter and Reddit.

6 Conclusion

Our work was motivated by the fact that effective content moderation is critical to limit the spread of harmful content on social media platforms. We aimed to address the problem of how human annotations in toxicity classification can introduce biases based on annotators' identities, experiences, and societal norms. Our goal was to explore the impact of fine-tuning BERTweet and HateBERT on topic-specific subsets of the NLPositionality dataset on its generalization to other platforms. We accomplished this by using topic modeling via LDA to find latent themes in toxic and non-toxic tweets. Our results demonstrate how fine-tuning on specific topics improves the F1 score of the model and we provide in-depth statistical analysis related to annotator positionality.

References

Rabin Adhikari, Safal Thapaliya, Nirajan Basnet, Samip Poudel, Aman Shakya, and Bishesh Khanal. COVID-19-related Nepali tweets classification in a low resource setting. In Graciela Gonzalez-Hernandez and Davy Weissenbacher, editors, *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 209–215, Gyeongju, Republic of Korea, October 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.smm4h-1.52.

Sathish A. P. Kumar Biraj Dahal and Zhenlong Li. Topic modeling and sentiment analysis of global climate change tweets. *Soc. Netw. Anal. Min.*, 9, 2019. doi: 10.1007/s13278-019-0568-8.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.88. URL https://aclanthology.org/2023.findings-emnlp.88.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for implicit and adversarial hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Do Yeon Kim, Xiaohang Li, Sheng Wang, Yunying Zhuo, and Roy Ka-Wei Lee. Topic enhanced word embedding for toxic content detection in q&a sites. In *Proceedings of the 2019 IEEE/ACM*

International Conference on Advances in Social Networks Analysis and Mining, ASONAM '19, page 1064–1071, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368681. doi: 10.1145/3341161.3345332. URL `https://doi.org/10.1145/3341161.3345332`.

Minchul Lee. bab2min/tomotopy: 0.12.3, July 2022. URL `https://doi.org/10.5281/zenodo.6868418`.

Shervin Malmasi and Marcos Zampieri. Detecting hate speech in social media. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_062. URL `https://doi.org/10.26615/978-954-452-049-6_062`.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, 2020.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. NLPositionality: Characterizing design biases of datasets and models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.505. URL `https://aclanthology.org/2023.acl-long.505`.

Philipp Seeberger and Korbinian Riedhammer. Enhancing crisis-related tweet classification with entity-masked language modeling and multi-task learning. In Laura Biester, Dorottya Demszky, Zhijing Jin, Mrinmaya Sachan, Joel Tetreault, Steven Wilson, Lu Xiao, and Jieyu Zhao, editors, *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 70–78, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlp4pi-1.9. URL `https://aclanthology.org/2022.nlp4pi-1.9`.

Elaine Zosa, Ravi Shekhar, Mladen Karan, and Matthew Purver. Not all comments are equal: Insights into comment moderation from a topic-aware model, 2021.

Appendix

Table 6: Model performance breakdown for topic and demographic subsets for Toxigen_HateBERT.

Data Subset	Demographic	Micro F1	Precision	Recall
full	asian	0.5031	0.5031	0.5031
topic 0	asian	0.5517	0.5517	0.5517
topic 1	asian	0.4828	0.4828	0.4828
topic 2	asian	0.4894	0.4894	0.4894
full	black	0.5385	0.5385	0.5385
topic 0	black	0.3158	0.3158	0.3158
topic 1	black	0.4286	0.4286	0.4286
topic 2	black	0.4737	0.4737	0.4737
full	latino/latina	0.5849	0.5849	0.5849
topic 0	latino/latina	0.7222	0.7222	0.7222
topic 1	latino/latina	0.55	0.55	0.55
topic 2	latino/latina	0.4667	0.4667	0.4667
full	man	0.4949	0.4949	0.4949
topic 0	man	0.5526	0.5526	0.5526
topic 1	man	0.5528	0.5528	0.5528
topic 2	man	0.4348	0.4348	0.4348
full	native american	0.6667	0.6667	0.6667
topic 0	native american	0.5	0.5	0.5
topic 1	native american	1.0	1.0	1.0
topic 2	native american	0.5	0.5	0.5
full	non-binary	0.4933	0.4933	0.4933
topic 0	non-binary	0.5455	0.5455	0.5455
topic 1	non-binary	0.4828	0.4828	0.4828
topic 2	non-binary	0.625	0.625	0.625
full	pacific islander	0.7143	0.7143	0.7143
topic 0	pacific islander	0.3333	0.3333	0.3333
topic 1	pacific islander	1.0	1.0	1.0
topic 2	pacific islander	1.0	1.0	1.0
full	white	0.4857	0.4857	0.4857
topic 0	white	0.5352	0.5352	0.5352
topic 1	white	0.5309	0.5309	0.5309
topic 2	white	0.4768	0.4768	0.4768
full	woman	0.5410	0.5410	0.5410
topic 0	woman	0.4919	0.4919	0.4919
topic 1	woman	0.5357	0.5357	0.5357
topic 2	woman	0.5053	0.5053	0.5053

Table 7: Model performance breakdown for topic and demographic subsets for BERTweet.

Data Subset	Demographic	Micro F1	Precision	Recall
full	asian	0.5399	0.5399	0.5399
topic 0	asian	0.5690	0.5690	0.5690
topic 1	asian	0.5172	0.5172	0.5172
topic 2	asian	0.5106	0.5106	0.5106
full	black	0.5962	0.5962	0.5962
topic 0	black	0.3684	0.3684	0.3684
topic 1	black	0.4286	0.4286	0.4286
topic 2	black	0.5263	0.5263	0.5263
full	latino/latina	0.5849	0.5849	0.5849
topic 0	latino/latina	0.6667	0.6667	0.6667
topic 1	latino/latina	0.6	0.6	0.6
topic 2	latino/latina	0.5333	0.5333	0.5333
full	man	0.5378	0.5378	0.5378
topic 0	man	0.5395	0.5395	0.5395
topic 1	man	0.5829	0.5829	0.5829
topic 2	man	0.4275	0.4275	0.4275
full	native american	0.75	0.75	0.75
topic 0	native american	0.5	0.5	0.5
topic 1	native american	1.0	1.0	1.0
topic 2	native american	0.5	0.5	0.5
full	non-binary	0.4933	0.4933	0.4933
topic 0	non-binary	0.4545	0.4545	0.4545
topic 1	non-binary	0.4828	0.4828	0.4828
topic 2	non-binary	0.6667	0.6667	0.6667
full	pacific islander	0.5714	0.5714	0.5714
topic 0	pacific islander	0.3333	0.3333	0.3333
topic 1	pacific islander	1.0	1.0	1.0
topic 2	pacific islander	1.0	1.0	1.0
full	white	0.5165	0.5165	0.5165
topic 0	white	0.5211	0.5211	0.5211
topic 1	white	0.5494	0.5494	0.5494
topic 2	white	0.4967	0.4967	0.4967
full	woman	0.5578	0.5578	0.5578
topic 0	woman	0.5135	0.5135	0.5135
topic 1	woman	0.5491	0.5491	0.5491
topic 2	woman	0.5372	0.5372	0.5372