Research Article

Leveraging Large Language Models and Topic Modeling for Toxicity Classification

Haniyeh Ehsani Oskouie[1], Christina Chance[1], Claire Huang[1], Margaret Capetz[1], Elizabeth Eyeson[1], Majid Sarrafzadeh[1]

1. Department of Computer Science, University of California, Los Angeles, United States

Content moderation and toxicity classification represent critical tasks with significant social implications. However, studies have shown that major classification models exhibit tendencies to magnify or reduce biases and potentially overlook or disadvantage certain marginalized groups within their classification processes. Researchers suggest that the positionality of annotators influences the gold standard labels in which the models learned from propagate annotators' bias. To further investigate the impact of annotator positionality, we delve into fine-tuning BERTweet and HateBERT on the dataset while using topic-modeling strategies for content moderation. The results indicate that fine-tuning the models on specific topics results in a notable improvement in the F1 score of the models when compared to the predictions generated by other prominent classification models such as GPT-4, PerspectiveAPI, and RewireAPI. These findings further reveal that the state-of-the-art large language models exhibit significant limitations in accurately detecting and interpreting text toxicity contrasted with earlier methodologies. Code is available at https://github.com/aheldis/Toxicity-Classification.git.

Haniyeh Ehsani Oskouie, Christina Chance, Claire Huang, Margaret Capetz, and Elizabeth Eyeson contributed equally to this work.

I. 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

 doi.org/10.32388/GBT682

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 the identities of the annotators, lived experiences, societal / cultural norms and personal beliefs, a concept known as positionality of the annotator[1]. This subjectivity can inadvertently perpetuate stereotypes and marginalization in datasets and thus impact the performance of machine learning models. Therefore, investigating the behavior of neural networks using an unbiased dataset is fundamental to the development of reliable, fair, and effective AI systems. Additionally, with the growing recognition of large language models (LLMs), ensuring that they do not produce or amplify toxic content is crucial for user safety and platform integrity. So far, many researchers have shown the inability of these models to distinguish toxicity within text[2]. Transfer learning may be used to improve content moderation systems by leveraging pre-trained models' knowledge. This approach allows for more efficient and effective toxicity classification by fine-tuning existing models on domain-specific data, potentially reducing bias and improving performance across diverse contexts. The objective of this paper is to introduce a topic-modeling-enhanced fine-tuning approach applied to toxicity data, with the aim of achieving superior results compared to existing toxicity classification models.

II. Background

Large language models (LLMs), including the Generative Pre-trained Transformer (GPT) developed by OpenAI[3] and the Bidirectional Encoder Representations from Transformers (BERT) developed by Google[4], have shown great ability in understanding and generating human language. These models are pre-trained on extensive amounts of data and are fine-tuned and applied for specific tasks, such as content generation, translation, code development, sentiment analysis, and more[5][6][7][8][9] using both single-machine and federated learning approaches[10]. Due to the growing popularity of the LLMs, there has been an increasing concern about the performance of LLMs in understanding toxicity, which plays a crucial role in creating safer, more inclusive online spaces. Unfortunately, it has been shown that neural networks, especially LLMs, often exhibit biases caused by biases in their training data, reflecting the cultural and personal backgrounds of annotators[1]. Moreover, some studies indicate that significantly more toxic language can be generated using GPT by assigning its

persona[2]. While many studies have sought to enhance the performance of LLMs regarding toxicity detection[11]{m/12/}[12], none have examined their performance on toxicity datasets generated based on annotator positionalities. These showcase the necessity of implementing an effective approach to address the toxicity present within texts while working with LLMs.

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 studies that utilize 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[13][14].

Several works explore the use of author-pooled Latent Dirichlet Allocation (LDA) to extract discussion topics from Twitter data related to climate change[15]. Similarly, there has been a focus on comment moderation, utilizing a topic-aware model to enhance automatic moderation by incorporating semantic features from topic models[16]. In a related context, various studies delve into enhancing word embeddings with topical information for toxic content detection, showcasing the effectiveness of incorporating topic-specific data in classification tasks[17].

These studies collectively underscore the significance of toxicity classification and considering topic modeling, contextual factors, and specialized features in toxicity assessment and content moderation. Expanding on the knowledge gleaned from prior research, our objective in this study is to advance the field by utilizing a topic-modeling methodology for enhancing the performance in toxicity classification for short-form text. For this purpose, we first use LDA topic modeling[18] for topic clustering on data. We then fine-tune BERTweet[19] and HateBERT[20] on the subsets of the data generated by LDA to learn embeddings and neural representations that capture the key features influencing the classification of a toxic tweet. Ultimately, to evaluate the effectiveness of the proposed approach, we compare its results against those of existing toxicity detection models.

III. Method

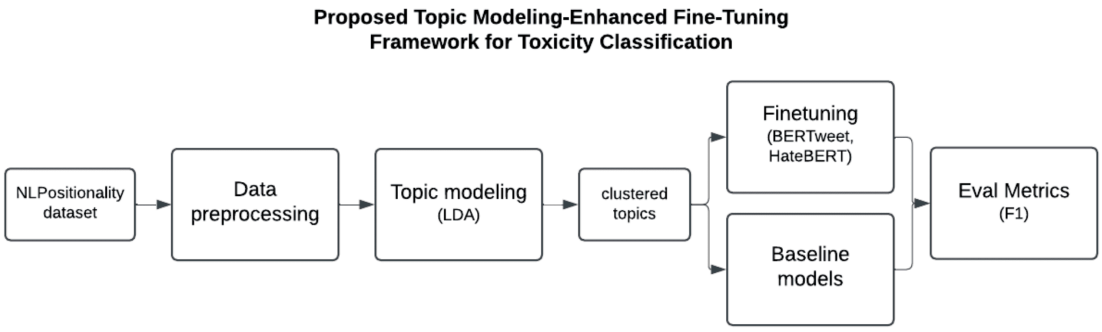


Figure 1. Diagram illustrating the proposed strategy for toxicity classification using topic modeling and fine-tuning.

An overview of our method is shown in Fig. 1. The details include:

A. Dataset

The dataset selected for our analysis is NLPositionality, a benchmark dataset consisting of labeled toxic tweets and annotator demographic metadata. This dataset is derived from[21], which introduces a framework for characterizing design biases and quantifying the positionality of natural language processing (NLP) datasets and models. By utilizing this dataset, we ensure that the analysis of toxicity in LLMs is accurate.

B. Data preprocessing

For preprocessing the data, we employ several techniques, including sentence tokenization, stop word removal, and lemmatization that enhance the quality of our input data. To handle the tokenization process for our models, we follow the same configuration settings that were established for the BERT models[19][20].

C. Topic clustering

Latent Dirichlet Analysis (LDA), a popular statistical technique[18], is performed for topic modeling on the training data and then applied to the test set. Using this method, we cluster the toxicity dataset

into 3, 6, and 10 topics. The examples of this clustering are shown in Table I. Our findings demonstrate that performing topic modeling when the number of clusters k is larger produces more insightful and expressive topics. Eventually, we use $k = 3$ for fine-tuning the models as we have a smaller dataset. Further subsetting would produce insignificant numbers that we cannot draw assumptions.

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

Table I. Topic distribution produced by Latent Dirichlet Allocation.

D. Models

For toxicity classification, we utilize two pre-trained models including BERTweet and HateBERT. 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. As BERTweet was fine-tuned on short form tweets, the goal was to leverage the model's ability to perform a task on limited context. On the other hand, HateBERT was obtained by fine-tuning the English BERT base uncased model on ToxiGen[22] data. The goal of this model was to leverage its task-specific context and its learned

ability in understand implicit toxicity to be able to generalize on more explicit examples[19][20]. These specifications make them suitable for fine-tuning with the purpose of toxicity analysis.

E. Transfer learning

The potential advantages of transfer learning 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[23]. This is particularly beneficial for HateBERT, as its pre-trained weights are already well-aligned with our task. Thus, we fine-tune the models for toxicity classification on various data splits. In this regard, we freeze all layers except for the classification head. The hyperparameters used for tranfer learning are as follows: learning rate of $5e-5$, 0 warm up steps, and 70 epochs. To ensure reliable results, each experiment is repeated five times using different manual seeds and both the mean and standard deviation of the outcomes are reported.

IV. Results

Table II presents the F1 score of the BERT models for different seeds. From the reported results, we see for BERTweet and HateBERT, fine-tuning the models on individual topics improved the F1 score compared to fine-tuning on the full dataset on average. The most significant improvement in the F1 score was for Topic 0, while the differences between the full dataset and Topic 1 and 2 were more marginal in comparison. One possible explanation may be that Topic 0 provides more distinct, consistent patterns of toxicity for the model to recognize, while Topics 1 and 2 may contain more varied, nuanced forms of toxicity. Interestingly, for HateBERT, the model fine-tuned on the entire dataset performed second to best, while for BERTweet, the model fine-tuned on the full dataset performed the worst.

Model	Data split	Seed 0	Seed 1	Seed 2	Seed 3	Seed 9	Average	Stdev
BERTweet	Topic 0	0.5588	0.5566	0.5566	0.5588	0.5588	0.5579	0.0012
	Topic 1	0.4778	0.4778	0.4778	0.4778	0.4778	0.4778	0.0000
	Topic 2	0.4659	0.4600	0.4659	0.4659	0.4600	0.4636	0.0032
	Full data	0.4610	0.4631	0.4610	0.4560	0.4610	0.4604	0.0026
HateBERT	Topic 0	0.5498	0.5498	0.5498	0.5498	0.5498	0.5498	0.0000
	Topic 1	0.4767	0.4767	0.4767	0.4767	0.4767	0.4767	0.0000
	Topic 2	0.4571	0.4572	0.4572	0.4572	0.4572	0.4572	0.0000
	Full data	0.4831	0.4765	0.4837	0.4835	0.4852	0.4824	0.0034

Table II. F1 score for BERT models with different seeds.

Due to the absence of similar datasets for toxicity detection based on annotator positionalities, we did not compare our results with any other datasets. Instead, we analyzed the performance of other baselines for toxicity detection on both the full data and its splits. These baselines include GPT-4[3], PerspectiveAPI[24], RewireAPI[25], and HateRoberta without fine-tuning[22]. For all these models, we employed the similar settings specified by[1]. The results are demonstrated in Table III. As indicated, all of these models exhibit lower performance compared to our fine-tuned models. This highlights that current toxicity detection models and large language models like GPT-4 are not effectively trained to identify toxicity in text. This indicates that, similar to our approach, they require additional training or fine-tuning on NLPositionality-like datasets to enhance their robustness to toxicity.

Model	Topic 0	Topic 1	Topic 2	Full data
PerspectiveAPI	0.3854	0.3857	0.3143	0.3636
RewireAPI	0.4386	0.4153	0.4295	0.4278
HateRoberta	0.3689	0.4216	0.3310	0.3788
GPT-4	0.3966	0.4163	0.3915	0.4054
BERTweet	0.5579	0.4778	0.4636	0.4604
HateBERT	0.5498	0.4767	0.4572	0.4824

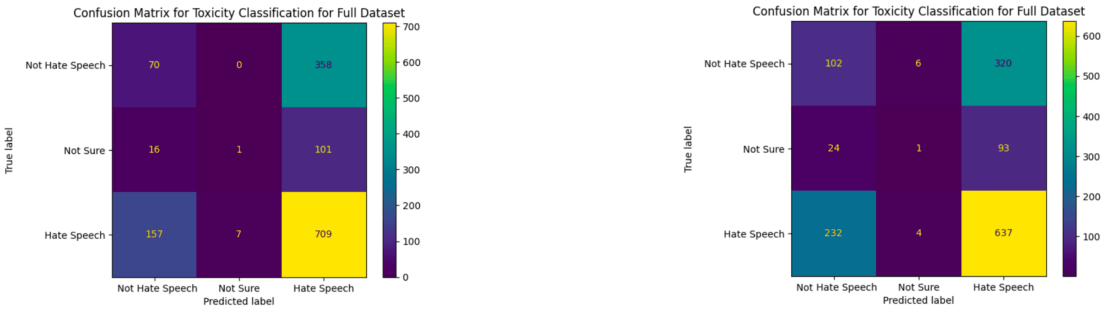
Table III. F1 score comparison between baselines and BERT models.

V. Discussion

A. Analysis

Table IV displays majority voting performed across seed runs to get predicted labels. 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.

Data subsets were further grouped by gender and ethnicity and visualized using confusion matrices as illustrated in Fig. 2. Since positionality bias propagates through the form of incorrectly labeling examples as hate speech, we were interested in the true positive rate (TPR), true negative rate (TNR), false positive rates (FPR), as well as the precision and f1 score. As seen in Fig. 2, both fine-tuned BERTweet and HateBERT had high FPRs and TNRs. For Black annotators especially, BERTweet had a high TPR and recall overall.



(a) BERTweet (b) HateBERT

Figure 2. Confusion matrices for fine-tuned BERTweet and HateBERT.

Model	Data split	Micro F1	Precision	Recall
BERTweet	Full data	0.5497	0.5497	0.5497
	Topic 0	0.5242	0.5242	0.5242
	Topic 1	0.5574	0.5574	0.5574
	Topic 2	0.5172	0.5172	0.5172
HateBert	Full data	0.5215	0.5215	0.5215
	Topic 0	0.5196	0.5196	0.5196
	Topic 1	0.5410	0.5410	0.5410
	Topic 2	0.5011	0.5011	0.5011

Table IV. Statistics for BERT models

B. Limitations

Topic modeling was not as expressive as necessary due to the variety in tweets. 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 fine-tune process may have contributed to the high error rates for that new label.

VI. 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 tackle the issue of biases introduced by human annotations in toxicity classification, which can be influenced by annotators' identities, experiences, and societal norms. Specifically, we explored the impact of fine-tuning BERTweet and HateBERT on topic-specific subsets of the NLPositionality dataset and 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 that fine-tuning the models on specific topics significantly enhances the F1 score compared to the other existing toxicity models. Future research should focus on mitigating the biases present in widely used models like GPT, as their increasing popularity raises significant concerns. Addressing these biases is crucial to ensure fair and equitable outcomes.

Appendix A. Additional results

In Table V and Table VI, we assess positionality and model alignment for different demographics based on overall f1 score as well as TPR (recall). For data subsets and demographic identities associated with higher TPR and F1 scores, this suggests a model alignment with positionality.

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

doi.org/10.32388/GBT682

Data Subset	Demographic	Micro F1	Precision	Recall
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 V. Model performance breakdown for topic and demographic subsets for Toxigen_HateBERT.

doi.org/10.32388/GBT682

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

doi.org/10.32388/GBT682

Data Subset	Demographic	Micro F1	Precision	Recall
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

Table VI. Model performance breakdown for topic and demographic subsets for BERTweet.

References

1. [a], [b], [c] *Santy S, Liang J, Le Bras R, Reinecke K, Sap M. "NLPositionality: Characterizing design biases of da tasets and models." In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Mee ting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Associ ation for Computational Linguistics; 2023. p. 9080–9102. doi:10.18653/v1/2023.acl-long.505. Available from: https://aclanthology.org/2023.acl-long.505.*
2. [a], [b] *Deshpande A, Murahari V, Rajpurohit T, Kalyan A, Narasimhan K. "Toxicity in chatgpt: Analyzing pe rsona-assigned language models." In: Bouamor H, Pino J, Bali K, editors. Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics; 2023. p. 1236–1270. doi:10.18653/v1/2023.findings-emnlp.88. Available from: https://aclanthology.org/2023.fin dings-emnlp.88.*
3. [a], [b] ***OpenAI** (2023). "ChatGPT". Large language model. Available from: https://chat.openai.com.*

4. ^Devlin J, Chang MW, Lee K, Toutanova K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: Proceedings of NAACL-HLT; 2019. p. 4171-4186.
5. ^Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019). "Language Models are Unsupervised Multitask Learners."
6. ^Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C. "m{T}5: A Massively M ultilingual Pre-trained Text-to-Text Transformer." In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. On line: Association for Computational Linguistics; 2021. p. 483-498. doi:10.18653/v1/2021.naacl-main.41. Available from: https://aclanthology.org/2021.naacl-main.41.
7. ^Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Ask ell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskev er I, Amodei D. "Language Models are Few-Shot Learners." In: Larochelle H, Ranzato M, Hadsell R, Balc an MF, Lin H, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 202 0. 33:1877-1901. Available from: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6b fcb4967418bfb8ac142f64a-Paper.pdf.
8. ^Jiang J, Wang F, Shen J, Kim S, Kim S (2024). "A Survey on Large Language Models for Code Generatio n." arXiv. Available from: https://arxiv.org/abs/2406.00515.
9. ^Zhang W, Deng Y, Liu B, Pan SJ, Bing L (2023). "Sentiment Analysis in the Era of Large Language Mod els: A Reality Check." arXiv. arXiv:2305.15005 [cs.CL].
10. ^Ghiasvand S, Yang Y, Xue Z, Alizadeh M, Zhang Z, Pedarsani R (2024). "Communication-Efficient and Tensorized Federated Fine-Tuning of Large Language Models." arXiv preprint arXiv:2410.13097. Availa ble from: https://arxiv.org/abs/2410.13097.
11. ^Zhang J, Wu Q, Xu Y, Cao C, Du Z, Psounis K (2024). "Efficient toxic content detection by bootstrapping and distilling large language models." Proceedings of the AAAI Conference on Artificial Intelligence. 38 (19): 21779–21787.
12. ^Mishra S, Chatterjee P (2024). "Exploring ChatGPT for Toxicity Detection in GitHub." In: Proceedings o f the 2024 ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results, pp. 6–10.
13. ^Adhikari R, Thapaliya S, Basnet N, Poudel S, Shakya A, Khanal B. "COVID-19-related Nepali tweets cla ssification in a low resource setting." In: Gonzalez-Hernandez G, Weissenbacher D, editors. Proceedings

of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task. Gyeongju, Republic of Korea: Association for Computational Linguistics; 2022. p. 209-215. Available from: https://aclanthology.org/2022.smm4h-1.52.

14. ^*Seeberger P, Riedhammer K. "Enhancing crisis-related tweet classification with entity-masked language modeling and multi-task learning." In: Biester L, Demszky D, Jin Z, Sachan M, Tetreault J, Wilson S, Xiao L, Zhao J, editors. Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics; 2022. p. 70–78. doi:10.18653/v1/2022.nlp4pi-1.9. Available from: https://aclanthology.org/2022.nlp4pi-1.9.*
15. ^*Dahal B, Kumar SA, Li Z (2019). "Topic modeling and sentiment analysis of global climate change tweets". Soc. Netw. Anal. Min.. 9. doi:10.1007/s13278-019-0568-8.*
16. ^*Zosa E, Shekhar R, Karan M, Purver M (2021). "Not All Comments are Equal: Insights into Comment Moderation from a Topic-Aware Model." arXiv. arXiv:2109.10033 [cs.CL].*
17. ^*Kim DY, Li X, Wang S, Zhuo Y, Lee RK (2020). "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. New York, NY, USA: Association for Computing Machinery. p. 1064–1071. doi:10.1145/3341161.3345332.*
18. a, b*Blei DM, Ng AY, Jordan MI (2003). "Latent Dirichlet Allocation". Journal of Machine Learning Research.*
19. a, b, c*Nguyen DQ, Vu T, Nguyen AT (2020). "BERTweet: A pre-trained language model for English Tweets." In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 9–14.*
20. a, b, c*Caselli T, Basile V, Mitrović J, Granitzer M. "HateBERT: Retraining BERT for abusive language detection in English." In: Mostafazadeh Davani A, Kiela D, Lambert M, Vidgen B, Prabhakaran V, Waseem Z, editors. Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021). Online: Association for Computational Linguistics; 2021. p. 17–25. doi:10.18653/v1/2021.woah-1.3. Available from: https://aclanthology.org/2021.woah-1.3.*
21. ^*Santy S, Liang J, Le Bras R, Reinecke K, Sap M. "NLPositionality: Characterizing design biases of datasets and models." In: Rogers A, Boyd-Graber J, Okazaki N, editors. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics; 2023. p. 9080–9102. doi:10.18653/v1/2023.acl-long.505. Available from: https://aclanthology.org/2023.acl-long.505.*

22. a, b *Hartvigsen T, Gabriel S, Palangi H, Sap M, Ray D, Kamar E (2022). "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.*

23. ^ *Farahani A, Pourshojae B, Rasheed K, Arabnia HR (2020). "A concise review of transfer learning." In: 2020 International Conference on Computational Science and Computational Intelligence (CSCI). 2020. p. 344–351.*

24. ^ *Google Jigsaw (2017). "Perspective API". https://www.perspectiveapi.com/. Accessed: 2024-02-02.*

25. ^ *Rewire API. 2023. Available from: https://rewire.online. Accessed: 2023-02-02.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.