How Does User Engagement Support Content Moderation? A Deep Learning-based Comparative Study

Conference Paper · August 2023

CITATIONS
4

READS
276

4 authors, including:

Kanlun Wang
University of North Carolina at Charlotte
19 PUBLICATIONS **52** CITATIONS
SEE PROFILE

Lina Zhou
China University of Geosciences
97 PUBLICATIONS **7,118** CITATIONS
SEE PROFILE

Dongsong Zhang
University of Maryland, Baltimore County
107 PUBLICATIONS **8,048** CITATIONS
SEE PROFILE

Aug 10th, 12:00 AM

How Does User Engagement Support Content Moderation? A Deep Learning-based Comparative Study

Kanlun Wang
University of North Carolina at Charlotte, kwang17@uncc.edu

Zhe Fu
University of North Carolina at Charlotte, zfu2@uncc.edu

Lina Zhou
University of North Carolina at Charlotte, lzhou8@uncc.edu

Dongsong Zhang
University of North Carolina at Charlotte, dzhang15@uncc.edu

How Does User Engagement Support Content Moderation? A Deep Learning-based Comparative Study

Completed Research Full Paper

Kanlun Wang
The University of North Carolina at Charlotte
kwang17@uncc.edu

Zhe Fu
The University of North Carolina at Charlotte
zfu2@uncc.edu

Lina Zhou
The University of North Carolina at Charlotte
lzhou8@uncc.edu

Dongsong Zhang
The University of North Carolina at Charlotte
dzhang15@uncc.edu

Abstract

Content moderation is a common intervention strategy for reviewing user-generated content on social media platforms. Engaging users in content moderation is promising for making ethical and fair moderation decisions. A few studies that have considered user engagement in content moderation have primarily focused on classifying user-generated comments, rather than leveraging the information of user engagement to make a moderation decision on user-generated posts. Moreover, how to extract information from user engagement to enhance content moderation remains unclear. To address the above-mentioned limitations, this study proposes a framework for user engagement-enhanced moderation of user-generated posts. Specifically, it incorporates the credibility and stance of user-generated content into graph learning. Our empirical evaluation shows that the models based on our proposed framework outperform the state-of-the-art deep learning models in making moderation decisions for user-generated posts. The findings of this study have implications for augmenting the moderation of social media content and for improving the safety and success of online communities.

Keywords

Content moderation, user engagement, deep learning, graph learning.

Introduction

Content moderation is a process of intervening user-generated content on social media platforms to ensure that the content complies with the platforms' policies and community standards (Gillespie 2020). Most social media platforms follow governance mechanisms, namely incorporating "structure participation in a community to facilitate cooperation and prevent abuse" (Grimmelmann 2015). Among different strategies for content moderation, this study focuses on distributed moderation, which has been widely adopted by most social media platforms (Jhaver et al. 2019). Distributed moderation involves delegating moderation tasks to a group of users, rather than relying solely on a centralized moderation team or algorithm, and may introduce biases, such as cultural bias (Gillespie et al. 2020) and political bias (Jiang et al. 2019), to the decision-making process of content moderation.

Content moderation can help improve user engagement in online communities. When users feel safe and supported in a community, they are more likely to engage positively and constructively (Jiménez Durán 2021; Liu et al. 2022). In addition, user engagement in social media refers to the extent to which users interact with a social media platform, including creating and sharing content, commenting, liking, sharing, and following other users (Shahbaznezhad et al. 2021). User engagement can in turn be valuable to content

moderation, as it allows moderators to gather feedback from their community about what content they find objectionable or inappropriate. The role of user engagement is a process and product of user interactions (Kappelman 1995), which have great potential to facilitate graph learning to form structure-based information about online discussions (e.g., commenting on a post) generated by users. The user engagement information is rich and may be influenced by other factors, such as content quality and stance (Hyland 2005). We define *user engagement* as textual content, structure-based information, and creditability (i.e., content quality) and stance of user-generated comments in the context of this study.

A few recent studies have leveraged the textual content of users' comments to build deep learning models in either making a binary moderation decision on users' comments (Pavlopoulos et al. 2017; Tan et al. 2020; Karabulut et al. 2023) or classifying types of moderators' comments to user-generated content (He et al. 2022). To the best of our knowledge, none of the previous studies has incorporated user engagement in content moderation towards user-generated posts. There is a lack of understanding of whether structure-based information of user engagement can be effective in improving content moderation. Moreover, it remains unclear to what extent structure-based information can be enhanced to improve the model performance in content moderation.

To fill the aforementioned research gaps, this study aims to answer the following research questions: RQ1) Can user engagement facilitate content moderation? RQ2) Does structure-based information of user engagement contribute to model performance? If so, how to enhance the effectiveness of structure-based information for content moderation?

This research makes three-fold contributions to the content moderation and information systems literature. First, we propose a framework for incorporating user engagement information into a model for content moderation in social media. Second, this study introduces the credibility and stance of user-generated content as factors to enhance the effectiveness of structure-based information of user engagement for content moderation. Third, this study makes a methodological contribution to data collection in support of content moderation research, which can serve as a testbed for future research.

Related Work

Content Moderation Methods

Content moderation plays a crucial role in ensuring that harmful or inappropriate content is removed from social media platforms, while also allowing for diverse opinions and viewpoints to be expressed (Gillespie 2020). Content moderation can be classified into four types, including 1) pre/proactive moderation (Coutinho and José 2017): it involves reviewing all content by human moderators and automated tools before it is posted or visible to other users, and is typically used in more sensitive or high-risk areas, such as online forums for children or political discussion groups; 2) post moderation (Coutinho and José 2017): it involves reviewing posted content before making it visible to other users. This method can be less time-intensive than pre-moderation, yet can still help prevent harmful content from being visible on a platform; 3) reactive moderation (Llansó 2020): it involves reviewing content only after it has been flagged or reported by other users. The platform relies on user reports to identify potentially harmful content first and then takes actions if the content violates community standards or terms of service; and 4) distributed moderation (Lampe and Resnick 2004): it relies on the community of power users (i.e., those users who have high reputation within an online community) to moderate the content. This can involve giving power users tools to flag or report harmful content, as well as moderating comments and other user-generated content. Among those different types of content moderation methods, distributed moderation remains dominant in use (Jhaver et al. 2019). However, a large stream of studies indicates that distributed moderation may amplify existing societal biases in decision-making, such as cultural bias (Gillespie et al. 2020) and political bias (Jiang et al. 2019).

User Engagement in Content Moderation

Social media platforms have focused on leveraging content moderation to improve user engagement on social media platforms by creating a positive, supportive, and inclusive online environment that encourages users to participate, engage, and connect in meaningful ways (Jiménez Durán 2021; Liu et al. 2022). Engaging users in content moderation outlines users' rights and helps to build a stronger and more vibrant

online community (Myers West 2018). A typical user engagement strategy for content moderation is to ask users to review, flag, and/or report inappropriate user-generated content (Llansó 2020). However, exposure to inappropriate or harmful content can have negative impacts on individuals, particularly on vulnerable populations (e.g., children). One potential solution to the problem is the design for contestability, whereby users can shape and influence the decision-making process in content moderation (Vaccaro et al. 2021). Nevertheless, how to incorporate information of user engagement in content moderation remains scarce.

Deep Learning Techniques

Deep learning techniques have been incredibly successful in content moderation as an effective and efficient solution. By extracting valuable features from user-generated content (e.g., text, images, videos), deep learning models attempt to learn the characteristics of moderated content. Since text is the most pervasive and accessible modality of social media content, a variety of deep learning models, such as Recurrent Neural Networks (RNNs) (Pavlopoulos et al. 2017) and transformers (Tan et al. 2020; He et al. 2022), have been adopted to learn the context information from the text and predict the probability of a moderation decision or the types of moderation that human moderators should make. In addition, with the rapid expansion of image content on social media platforms, the demand for efficiently moderating inappropriate images also increases dramatically. For instance, Karabulut et al. (2023) leveraged a Convolutional Neural Network (CNN)-based model to learn features from images and identify inappropriate images that need to be moderated. However, all of the aforementioned deep learning models were developed for moderating users' comments. Hence, there is a lack of model development in leveraging the information of user engagement for decision-making in moderating user-generated posts.

Method

A Deep Learning-based Framework for Content Moderation

Problem Formulation

We formulate content moderation as a binary classification problem that classifies user-generated posts as either being moderated or not. Let r denote the information of a post, which contains the content of a post (i.e., post title and/or body) and/or its associated user engagement in the form of comments, and r serves as the source of input for building classification models. The task can be formulated as learning a binary classifier for a content moderation decision:

$$y = f(\Theta, r) \tag{1}$$

where y denotes a binary classification result of a target post r being moderated, and Θ denotes the set of parameters of the classification function $f(\cdot)$. More importantly, the proposed framework integrates both user engagement information and user-generated post content to make a content moderation decision toward a user-generated post.

A Deep Learning-based Framework

We design a deep learning-based framework for classifying if content moderation of a social media post is needed or not, as shown in Figure 1. The framework consists of three components, including text encoding, representation learning, and moderation prediction. We introduce the technical details of each component next.

Text Encoding

Textual content of social media posts has been widely used in building deep learning models for content moderation. In this study, we focus on extracting textual features by converting the text of the post content and user comments into vector representations using a pre-trained BERT embedding model (Devlin et al. 2019). The power of the BERT encoder has been well demonstrated by previous NLP studies (Kaliyar et al. 2021; Minaee et al. 2021). Thus, we leverage BERT to generate the embedding vectors of each word in the post content and user comments with 128 dimensions.

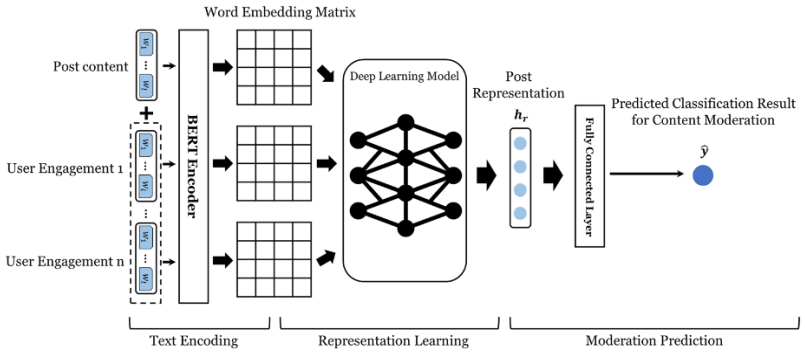

Figure 1. The Proposed Deep Learning-based Framework for Content Moderation

Representation Learning

In representation learning, we concatenate the word embeddings of post content and its comments as the input to obtain the hidden representations of post h_r. To this end, we select and adapt three state-of-the-art representation learning models, including TextCNN, BiLSTM, and RoBERTa.

- TextCNN (Kim 2014) is one of the most popular deep learning models for Natural Language Processing (NLP) tasks and has been shown to achieve state-of-the-art performances in text classification tasks. The text-CNN model uses a series of convolutional layers and pooling layers to both the local and the global context information to extract the local n-gram features and incorporate them into the post representation.
- BiLSTM is also one of the commonly used deep learning models for NLP tasks. It contains two Long Short-Term Memory (LSTM) layers: one is a left-to-right LSTM processing a word sequence from left to right, and the other one is a right-to-left LSTM layer processing the word sequence from right to left. By combining the outputs of the two layers, the BiLSTM model captures both the forward and backward context of the input text and generates a more effective post representation.
- RoBERTa (Liu et al. 2019) is one of the state-of-the-art pre-trained language models. It is a transformer-based model, which leverages the transformer layers with multi-head attention for representation learning and adopts dynamic mask learning for model training.

Moderation Prediction

We design a fully connected layer to generate a binary classification result to determine whether or not a target post should be moderated based on its latent representation h_r, as shown in Equation 2.

$$\hat{y} = \text{softmax}(W_o h_r + b_o) \tag{2}$$

Where $\hat{y}$ is the classification result of a post being moderated or not; softmax $(\cdot)$ is the softmax function; $W_o \in \mathbb{R}^{2\times d}$ denote the learnable weights; d is the dimension of the post representation h_r, and b_o is the bias for the fully connected layer. We choose the binary cross-entropy as the loss function, where T denotes the set of training instances, and $\mu\|\Delta\|$ is parameter-specific regulation hyper-parameters to prevent overfitting. By minimizing the loss value calculated by Equation 3, we tailor the model to generate classification results.

$$\mathcal{L} = \sum_{y\in T} \log(\hat{y}) + \mu\|\Delta\| \tag{3}$$

Enhancement with Structure-based Information

The primary objective of this research is to examine whether structured-based information of user engagement can facilitate moderating user-generated posts. To extract structural information, we first formulate user engagement with post r as a graph, $G_r = (V, E, W)$, where the node set V comprises the post content and its associated comments; the edge set E represents the observed interactions among the nodes in V, and W represents the weights of edge set E. We define an interaction as a comment in response to the original post or another user's comment.

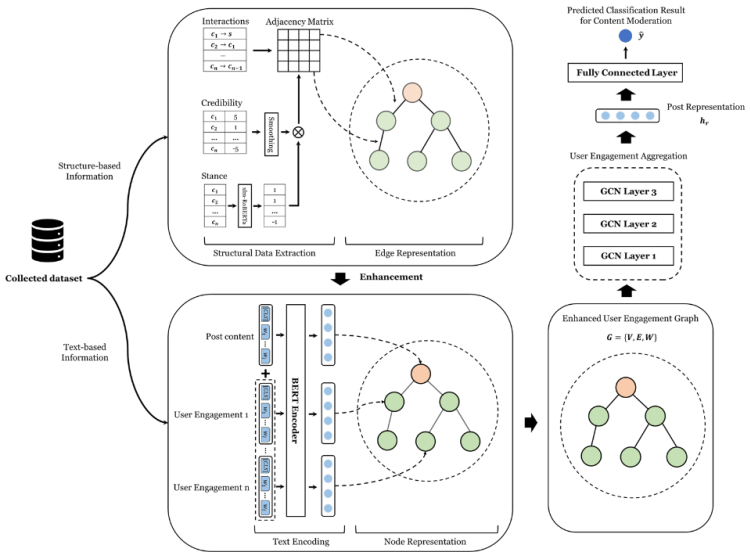

Figure 2. A Structure-based Information Enhanced Framework for Content Moderation

We leverage Graph Convolutional Network (GCN) (Kipf and Welling 2017) to learn post representations from a user engagement graph. With the enhancement of the structure-based information, we significantly extend the original framework of our proposed model to produce an enhanced framework as shown in Figure 2. We focus on introducing the new components of the enhanced framework next.

Node Representation

Given that the use of a word embedding matrix is massive and space-consuming, we add a special classification token (i.e., $[CLS]$) at the beginning of a word sequence and the final hidden state corresponding to this token is used as the aggregate sequence representation for a summary of an entire word sequence. In addition, unlike the word embedding vectors, the vectors for the $[CLS]$ tokens can be directly used to represent the post content and/or user comments for a specific post.

Edge Representation

In view that different user interactions are not equally important to the original posts/comments, we introduce weights to improve the edge representations in our graph model. Specifically, we estimate the weights of the edges based on the two influential factors of user engagement of the involved nodes: the credibility and the stance of the nodes.

Credibility: We assume that more credible user comments have a larger impact on the discussions of user engagement within a specific post. In this study, we leverage the karma score, referring to the total number of upvotes and downvotes that a user's posts and comments have received, as a proxy of the credibility of the user comment, which is shown in Equation 4:

$$W_{credibility} = \lambda + S_{Karma} \tag{4}$$

Where S_{Karma} denotes the karma score of the user comment; and λ is a smoothing factor. We introduce the smoothing factor to address the issue of missing values. The value of λ sets to be 1 in this study.

Stance: We further assume that the stances or opinions in users' comments have impacts on the discussions of user engagement within a specific post. User comments with more prominent positions should receive higher weights for the edges. In this study, we leverage the xlm-RoBERTa model (Conneau et al. 2020), which is pre-trained with the SemEval-2016 dataset (Mohammad et al. 2016), to generate the stance score W_{stance} for an edge in Graph G_r, and the score ranges from -1 ('against') to 1 ('favor').

To fuse the two types of edge weights, the credibility score $W_{credibility}$ and stance score W_{stance}, we use multiplication to derive W_{edge}, as shown in Equation 5.

$$W_{edge} = W_{credibility} \times W_{stance} \tag{5}$$

The advantage of multiplying the credibility and stance in Equation 5 is that we can use the karma score to correct the stance of user comments. For example, a child's comment with a stance of 'favor' is positive towards its parent node. If this user engagement has a negative karma score, we can infer that the stance of the user engagement is disagreed by the majority of the readers and should be corrected to 'against', which is negative. By multiplying the credibility and stance values, we can maintain or correct the stance of user engagements at the same time for edge weights generation.

User Engagement Aggregation

By aggregating the information of neighboring nodes, we can update the representations of node i in the graph G_r based on Equation 6:

$$h_i^{(k)} = W_1^{(k)} h_i^{(k-1)} + W_2^{(k)} \sum_{j \in \mathcal{N}(i)} W_{edge\ i,j} \cdot h_j^{(k-1)} + b^{(k)} \quad (6)$$

where $W_{edge\ i,j}$ denotes the edge weight from node j to node i; $\mathcal{N}(i)$ denotes a set of neighbors of node i; and $h_i^{(k-1)}$ and $h_j^{(k-1)}$ are the representations of node i and node j at the (k-1)th GCN layer, respectively. $W_1^{(k)}$ and $W_2^{(k)}$ are the learnable weights for the kth GCN layer related to node i and its neighbors, and $b^{(k)}$ is the bias for the kth GCN layer. It can be observed from Equation 6 that the representation of node i at the kth GCN layer is derived as a weighted average of its own representation and the representations of its neighbors at the (k-1)th layer. By applying the stacked GCN layers, we can enhance post representation h_r by propagating the information of user engagement that includes user comments and their associated karma scores and stances.

Evaluation

Platform and Online Community Selection

We choose Reddit as the platform for data collection because Reddit is one of the most popular social media platforms that enable social news aggregation, content rating, and discussion (Jungherr et al. 2022). The users can remain anonymous on Reddit, which allows them to naturally share informative content (Dosono and Semaan 2019). Each subreddit is a sub-online community that integrates a large fusion of content. Therefore, we select the top 10 popular subreddits from each of the four classic domains that employ different levels of content moderation, including fashion with 35% of content being moderated, health with 30%, sports with 26%, and gaming with 13%.

Data Collection and Preparation

The data collection was limited to public online communities to comply with the platform's privacy policy. Thus, the procedure did not require approval from the Institutional Review Board at the authors' institution. We leveraged a Pushshift Reddit API to scrape posts from 40 subreddits daily across four different domains from Aug 24 to October 28, 2022, resulting in 104,674 posts. Since manual moderation is time-consuming, the efficiency of content moderation depends on several factors such as the type and volume of content, the complexity of the moderation rules, and the availability of human moderators. To enhance the ecological validity of the study findings, we used a PRAW API to perform another round of data collection of the collected posts 2 months later to validate whether the post content was moderated or not. Thereafter, we used a snowballing approach to collect the corresponding comments on all the posts. The metadata includes post content, post time, comment content, comment time, karma score, etc. In view that the volume of comments associated with each post varies widely from 0 to 500 (i.e., the maximum limit restricted by the API), and the distribution of the comments associated with each post is extremely right-skewed, we set a threshold for the minimum number of comments to 2 to facilitate the extraction of graph-based structural information. On the other hand, we also set an upper bound for the number of direct comments (i.e., the comments that directly respond to the original post) to 15. The final dataset, namely *HumanMOD*[1], is publicly available, which consists of 8,511 moderated posts and another 8,511 not moderated posts that were randomly selected from the remainder of the dataset. All the posts were commented on, with a total of 148,344 comments.

[1] https://huggingface.co/datasets/SamW/HumanMOD

Evaluation Metrics and Experiment Setting

We adopt a set of widely used evaluation metrics to measure the predicted results, including accuracy, precision, recall, and F-1. Moreover, we adopt 5-fold cross-validation to evaluate model performance, with an 80/20 data split for training and testing. To evaluate the proposed user engagement-enhanced models for content moderation, we select three baseline models, including TextCNN, BiLSTM, and RoBERTa, and compare the performance of the proposed model with that of those baseline models with and without user engagement. We compare the performance of the proposed method after removing different influential factors of user engagement by conducting an ablation experiment and performing paired samples t-tests for model comparisons. In addition, we use the learning rate of 0.001, the hidden dimension of 128, and the dropout rate of 0.2 for all the models and trained all of the models using the Adam optimizer. We set the maximum length of input text to 512 tokens for a fair comparison.

Results

The Effect of User Engagement

To answer the research question RQ1, we compare the baseline models for content moderation with and without incorporating text-based user engagement information (i.e., user comments). Model performances and the results of paired sample t-tests are reported in Table 1. The t-test results show that user engagement improves the accuracy of the BiLSTM model (t=-3.495; df=4; p=.025), yet does not yield a significant effect for the TextCNN-based models (t=-.352; df=4; p=.743) and even shows a negative effect for the RoBERTa-based model (t=3.244; df=4; p=.032). The results of paired sample t-test using F1 are similar to that of using accuracy.

Models	User Engagement	Fold					MD	Sig.
		1	2	3	4	5		
TextCNN	w/o	63.56%	61.31%	62.46%	65.86%	61.28%	.004	.743
	w/	63.42%	61.63%	63.45%	62.54%	65.57%		
BiLSTM	w/o	67.67%	68.77%	68.33%	67.19%	67.83%	.066	.025
	w/	72.46%	77.91%	77.82%	67.13%	77.26%		
RoBERTa	w/o	**78.13%**	**79.38%**	**78.79%**	**78.94%**	**78.76%**	.012	.032
	w/	77.36%	77.50%	78.32%	76.76%	78.23%		

The results are reported in accuracy; the best results are in boldface; MD: mean difference.

Table 1: Model Performance Comparison Between with and without User Engagement

Our results confirm the findings of previous content moderation studies (Guo et al. 2021; Wang and Iwaihara 2021) that RoBERTa can achieve superior performance on short-text tasks that leverage post content only. More importantly, the mixed effects of user engagement across the different models suggest that the text-based information of user engagement does not necessarily improve the model's performance in content moderation. Thus, it is indeed important to investigate how to extract information from user engagement data to enhance the effectiveness of content moderation.

Content Moderation Performance

To answer the research question RQ2, we compare the performances of our proposed model and the baseline models. One unique characteristic of our proposed model is that it incorporates structure-based information of user engagement and is further enhanced by credibility and stance. The model performance is averaged across 5 folds and the results of paired sample t-test are reported in Table 2.

The t-test results show that our proposed model outperforms all the baseline models, in terms of accuracy, precision, and F1, at least at a significance level of .05. In addition, our model also outperforms the baseline models in terms of recall (p<.01 or p<.05) with a few exceptions. Our findings indicate that the enhanced structure-based information of user engagement contributes to improved model performance. In addition, the recall values are generally lower than precisions, revealing a slight tendency of false negative prediction. In other words, the models are more likely to predict the posts moderated by human moderators as non-moderated rather than the other way around. It suggests that human moderators may be subject to decision-making biases, which also partially explains the mixed effects on recall.

Models	User Engagement	Accuracy	Precision	Recall	F1
TextCNN	w/o	62.89%***	65.19%***	56.74%*	60.07%**
	w/	63.32%***	64.53%***	60.52%**	62.11%***
BiLSTM	w/o	67.96%***	68.31%***	67.45%**	67.72%***
	w/	74.52%*	76.02%*	71.99%	73.85%*
RoBERTa	w/o	78.80%**	79.69%***	**77.31%**	78.48%*
	w/	77.63%**	80.90%**	72.38%*	76.39%**
Proposed	w/	**81.70%**	**85.08%**	76.90%	**80.78%**

[a]: mean difference; ***: p<.001; **: p<.01; *: p<.05; the best results are in boldface, and the best baselines are underlined.

Table 2: Performance Comparison of Proposed and Baseline Models

Ablation Experiment

To gain a deeper understanding of the impacts of the different influential factors of user engagement of our proposed model on model performance in content moderation, we conduct an ablation experiment. The results averaged across 5 folds are plotted in Figure 3.

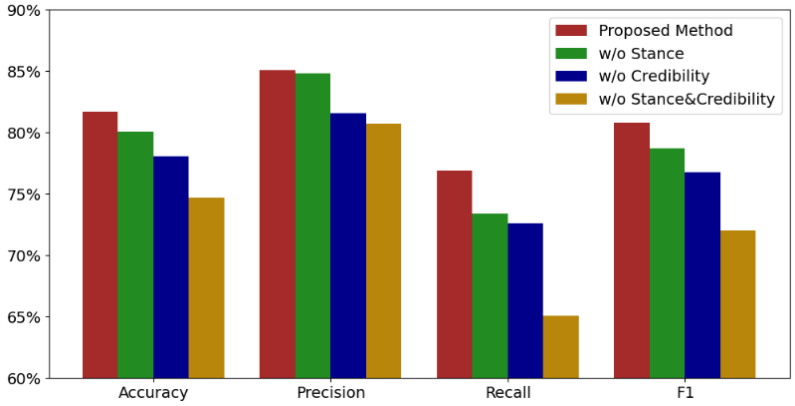

Figure 3: Ablation Experiment Results of the Proposed Method

The results show that incorporating the structure-based information of user engagement in terms of both credibility and stance achieves the best performances across all the metrics, including accuracy, precision, recall, and F1. After ablating stance, the model performance drops by 1.96% in accuracy (t=6.48; df =4; p=.003), 0.35% in precision (t=.607; df =4; p=.576), 4.55% in the recall (t=3.317; df=4; p=.029), and 2.60% in F1 (t=4.9; df=4; p=.008); after ablating credibility, the model performance drops by 4.41% in accuracy (t=12.584; df =4; p<.001), 4.11% in precision (t=4.286; df =4; p=.013), 5.59% in recall (t=3.378; df =4; p=.028), and 4.95% in F1 (t=8.189; df =4; p=.001); after ablating both content stance and credibility, the model performance drops 8.57% in accuracy (t=20.129; df =4; p<.001), 5.17% in precision (t=3.281; df =4; p=.03), 15.34% in recall (t=7.512; df =4; p=.002), and 10.89% in F1 (t=14.705; df =4; p<.001). In a nutshell, both content stance and credibility contribute to model performance in content moderation positively based on the structure-based information of user engagement. Moreover, credibility contributes more to the model's performance compared with its stance counterpart.

Conclusion

It is critical to ensure that social media users can engage in productive and respectful online interactions without being subjected to harassment, hate speech, or other forms of harmful content. However, content moderation on social media platforms is a complex and ongoing challenge, as the volume of content on these platforms is enormous and constantly growing. The findings of this study suggest that the structural information of user engagement with the enhancement of content credibility and stance can effectively improve the model performance in content moderation.

This research can be extended in multiple directions. First, this study collected user-generated content from four different domains, we envision that our proposed method can be generalizable to different domains or

online communities. Second, we enhance the structure-based information of user engagement with credibility and stance in this study. It opens up opportunities to examine other types of information that may enhance the measurement of the structure of user engagement. Third, we treat different online communities as a whole in model training and evaluation. It would be interesting to explore whether fine-tuning the general models to domain-specific ones may further improve model performance.

Acknowledgements

This work is partially supported by a Truist Research Grant and an SDS summer seed grant (#2022002). Any opinions, findings or recommendations expressed here are those of the authors and are not necessarily those of the sponsors of this research.

REFERENCES

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. 2020. Unsupervised Cross-Lingual Representation Learning at Scale, arXiv. (https://doi.org/10.48550/arXiv.1911.02116).

Coutinho, P., and José, R. 2017. "Moderation Techniques for User-Generated Content in Place-Based Communication," in 2017 12th Iberian Conference on Information Systems and Technologies (CISTI), , June, pp. 1–6. (https://doi.org/10.23919/CISTI.2017.7975786).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, June, pp. 4171–4186. (https://doi.org/10.18653/v1/N19-1423).

Dosono, B., and Semaan, B. 2019. "Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, New York, NY, USA: Association for Computing Machinery, May 2, pp. 1–13. (https://doi.org/10.1145/3290605.3300372).

Gillespie, T. 2020. "Content Moderation, AI, and the Question of Scale," Big Data & Society (7:2), SAGE Publications Ltd, p. 2053951720943234. (https://doi.org/10.1177/2053951720943234).

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernandez, A., Roberts, S. T., Sinnreich, A., and West, S. M. 2020. "Expanding the Debate about Content Moderation : Scholarly Research Agendas for the Coming Policy Debates," Internet Policy Review (9:4), Alexander von Humboldt Institute for Internet and Society, Article number: 4 1-29.

Grimmelmann, J. 2015. "The Virtues of Moderation," Yale Journal of Law and Technology. (https://openyls.law.yale.edu/handle/20.500.13051/7798).

Guo, Z., Zhu, L., and Han, L. 2021. "Research on Short Text Classification Based on RoBERTa-TextRCNN," in 2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI), , September, pp. 845–849. (https://doi.org/10.1109/CISAI54367.2021.00171).

He, Q., Hong, K., and Raghu, T. S. 2022. The Effects of Machine-Powered Platform Governance: An Empirical Study of Content Moderation. (http://hdl.handle.net/10125/80064).

Hyland, K. 2005. "Stance and Engagement: A Model of Interaction in Academic Discourse," Discourse Studies (7:2), SAGE Publications, pp. 173–192. (https://doi.org/10.1177/1461445605050365).

Jhaver, S., Bruckman, A., and Gilbert, E. 2019. "Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit," Proceedings of the ACM on Human-Computer Interaction (3:CSCW), 150:1-150:27. (https://doi.org/10.1145/3359252).

Jiang, S., Robertson, R. E., and Wilson, C. 2019. "Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation," Proceedings of the International AAAI Conference on Web and Social Media (13), pp. 278–289. (https://doi.org/10.1609/icwsm.v13i01.3229).

Jiménez Durán, R. 2021. The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter, SSRN Scholarly Paper, Rochester, NY. (https://doi.org/10.2139/ssrn.4044098).

Jungherr, A., Posegga, O., and An, J. 2022. "Populist Supporters on Reddit: A Comparison of Content and Behavioral Patterns Within Publics of Supporters of Donald Trump and Hillary Clinton," Social Science

Computer Review (40:3), SAGE Publications Inc, pp. 809–830. (https://doi.org/10.1177/0894439321996130).
Kaliyar, R. K., Goswami, A., and Narang, P. 2021. "FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach," Multimedia Tools and Applications (80:8), pp. 11765–11788. (https://doi.org/10.1007/s11042-020-10183-2).
Kappelman, L. A. 1995. "Measuring User Involvement: A Diffusion of Innovation Perspective," ACM SIGMIS Database: The DATABASE for Advances in Information Systems (26:2–3), pp. 65–86. (https://doi.org/10.1145/217278.217286).
Karabulut, D., Ozcinar, C., and Anbarjafari, G. 2023. "Automatic Content Moderation on Social Media," Multimedia Tools and Applications (82:3), pp. 4439–4463. (https://doi.org/10.1007/s11042-022-11968-3).
Kim, Y. 2014. "Convolutional Neural Networks for Sentence Classification," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar: Association for Computational Linguistics, October, pp. 1746–1751. (https://doi.org/10.3115/v1/D14-1181).
Kipf, T. N., and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks, arXiv. (https://doi.org/10.48550/arXiv.1609.02907).
Lampe, C., and Resnick, P. 2004. "Slash(Dot) and Burn: Distributed Moderation in a Large Online Conversation Space," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '04, New York, NY, USA: Association for Computing Machinery, April 25, pp. 543–550. (https://doi.org/10.1145/985692.985761).
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (https://openreview.net/forum?id=SyxS0T4tvS).
Liu, Y., Yildirim, P., and Zhang, Z. J. 2022. "Implications of Revenue Models and Technology for Content Moderation Strategies," Marketing Science (41:4), INFORMS, pp. 831–847. (https://doi.org/10.1287/mksc.2022.1361).
Llansó, E. J. 2020. "No Amount of 'AI' in Content Moderation Will Solve Filtering's Prior-Restraint Problem," Big Data & Society (7:1), SAGE Publications Ltd, p. 2053951720920686. (https://doi.org/10.1177/2053951720920686).
Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. 2021. "Deep Learning--Based Text Classification: A Comprehensive Review," ACM Computing Surveys (54:3), 62:1-62:40. (https://doi.org/10.1145/3439726).
Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. 2016. "SemEval-2016 Task 6: Detecting Stance in Tweets," in Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California: Association for Computational Linguistics, June, pp. 31–41. (https://doi.org/10.18653/v1/S16-1003).
Myers West, S. 2018. "Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms," New Media & Society (20:11), pp. 4366–4383. (https://doi.org/10.1177/1461444818773059).
Pavlopoulos, J., Malakasiotis, P., and Androutsopoulos, I. 2017. "Deeper Attention to Abusive User Content Moderation," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark: Association for Computational Linguistics, September, pp. 1125–1135. (https://doi.org/10.18653/v1/D17-1117).
Shahbaznezhad, H., Dolan, R., and Rashidirad, M. 2021. "The Role of Social Media Content Format and Platform in Users' Engagement Behavior," Journal of Interactive Marketing (53), pp. 47–65. (https://doi.org/10.1016/j.intmar.2020.05.001).
Tan, F., Hu, Y., Hu, C., Li, K., and Yen, K. 2020. "TNT: Text Normalization Based Pre-Training of Transformers for Content Moderation," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online: Association for Computational Linguistics, November, pp. 4735–4741. (https://doi.org/10.18653/v1/2020.emnlp-main.383).
Vaccaro, K., Xiao, Z., Hamilton, K., and Karahalios, K. 2021. "Contestability For Content Moderation," Proceedings of the ACM on Human-Computer Interaction (5:CSCW2), 318:1-318:28. (https://doi.org/10.1145/3476059).
Wang, X., and Iwaihara, M. 2021. "Integrating RoBERTa Fine-Tuning and User Writing Styles for Authorship Attribution of Short Texts," in Web and Big Data, Lecture Notes in Computer Science, L. H. U, M. Spaniol, Y. Sakurai, and J. Chen (eds.), Cham: Springer International Publishing, pp. 413–421. (https://doi.org/10.1007/978-3-030-85896-4_32).