Investigating Toxicity Across Multiple Reddit Communities, Users, and Moderators

Hind Almerekhi
Supervised by Bernard J. Jansen and co-supervised by Haewoon Kwak
Hamad Bin Khalifa University
Doha, Qatar
hialmerekhi@mail.hbku.edu.qa

ABSTRACT

Online platforms like Reddit enable users to build communities and converse about diverse topics and interests. However, with the increasing number of users that post disturbing comments containing profanity, harassment, and hate speech, otherwise known as *toxic comments*. Moderators often struggle with managing the safety of discussions in online communities. To address these issues, we need to detect toxic comments and the root causes of toxicity in discussion threads, i.e., toxicity triggers. Additionally, we need to investigate the toxic posting behavior of users to understand how it differs across online communities and consolidate our findings with moderators from Reddit. In this work, we present our approach, which builds on state-of-the-art methods of toxic comment and toxicity trigger detection. Lastly, we present our research findings of investigating toxicity across users and moderators on Reddit.

CCS CONCEPTS

• **Human-centered computing → Social media**; • **Computing methodologies → Supervised learning by classification**.

KEYWORDS

Reddit, toxicity, trigger detection, discussion threads, online communities

ACM Reference Format:
Hind Almerekhi and Supervised by Bernard J. Jansen and co-supervised by Haewoon Kwak. 2020. Investigating Toxicity Across Multiple Reddit Communities, Users, and Moderators. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3366424.3382091

1 PROBLEM

It is no surprise that the web drastically changed the way that users communicate with each other. Before the explosion of online communication, users interacted limitedly with their social circle using primitive means of communication. Interestingly, the original goal of web interactions was to facilitate interaction among friends. However, users took advantage of online platforms to communicate with an unlimited number of users not only to share information– but to share opinions, debate, and argue about different topics of interest [26]. With the change in the communication medium came a radical change in user's behavior towards each other [18]. This change caused a disturbing rise in online harassment, cyberbullying, hatefulness, and cyber threats, known as 'toxicity' [23]. This behavior was advocated by users that realized the power of online social media, where users assumed anonymous identities and refrained from filtering their comments, leading to many uncensored and offensive conversations.

This newfound freedom in online engagement attracted unsolicited comments in online conversations; such comments endanger the user's experience and harm them due to constant aggravation. When comments contain sentences that say things like "b!tch were not talking about that so I hope you get fuck!ing raped u fuck!ing whore"; they are considered by most users to be extremely hateful, offensive, and, most importantly, very toxic. Unfortunately, such comments are not rare in online conversations, and their existence impacts the community's courtesy and overall user experience. Pew Research Center stated in their report on online harassment that in 2017, 41% of Americans were victims of online harassment, and around 66% witnessed harassment behaviors directed at others [10].

As online social media and communication platforms realized the need to fight toxicity, they employed a set of rules and guidelines that users must follow to use their services. One of the most effective techniques involves a combination of human moderators and systems that monitor and remove toxic posts and penalize users. Systems and moderators often keep a list of bad words and expressions to filter out any comments using regular expressions. The problem with this approach is that the increasing complexity of online interactions makes it very difficult to monitor all comments effectively. This problem suggests that there is a need for sophisticated solutions that can accurately detect toxic comments and their causes in online conversations [23].

Problem Statement: the task of detecting toxicity in online content always suffers from inherent issues related to subjectivity and bias in the data and detection techniques [4]. To address some of these issues, we proposed a method for detecting the causes of toxicity (i.e., toxicity triggers) in online conversations [3]. Besides toxicity triggers, we have to investigate the toxic posting behavior of users and how it affects moderators and their moderation decisions across multiple online communities. The overall goal of this study, as implied by the problem statement, calls for solutions that combine toxic comment detection [11] and user behavioral modeling [20] to investigate toxicity in Reddit communities. The details of our approach in solving the research problems are in Section 3.

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3382091

2 STATE OF THE ART

The presence of toxicity in online platforms unfavorably affects content consumption, users, and moderators. Studies that investigate toxicity in online content focus on the detection of toxic comments [11, 23] without accounting for causes or triggers of toxicity. As for user-based studies, studies like [7] expand on toxic comment detection to include user's abusive behavior across multiple cross-domain communities. However, such studies do not account for different communities within the same domain, like the subreddits on Reddit. Moreover, moderation studies [8, 16] focus on moderation practices and the effect of disciplinary decisions on users. Nevertheless, such studies do not consider the toxicity and abuse projected towards moderators and how moderation decisions contribute towards that toxicity. In the following subsections, we briefly describe some of the studies that helped us shape our research design and approach towards solving our research problems.

2.1 Toxicity Detection

In toxic comments detection, the general goal is to classify comments as either toxic or non-toxic, like the study of Nobata et al. [23], which uses syntactic, linguistic, distributional semantics (embedding), and n-gram features to detect abusive comments. The outcomes of the research show that by combining all features, the built SVM model achieves an AUC of 0.9055. Another work that deploys machine learning models to detect toxicity was by Wulczyn et al. [28], where the problem targets personal attacks at different levels of toxicity. Similar to [23], the researchers extracted n-gram and semantic features to perform the detection. The study showed that the types of personal attacks on Wikipedia were not a result of a small number of malignant users, nor was it the outcome of anonymous commenters. Around 30% of the attacks came from registered users with more than 100 contributions per user [28].

As for open-source toxic comment detection solutions, Google offers the Perspective API [1]. The goal behind the API is to provide platform owners and researchers with a tool that detects toxic comments from conversations. The models used in the API were built using machine learning techniques and relied on an extensive collection labeled by crowdworkers [28]. To improve the performance of the Perspective API models, Kaggle launched a competition for toxic comments classification, where the work of Georgakopoulos et al. [11] showed that using CNN improved on toxic comments detection. On the other hand, the research that we discussed earlier by Wulczyn et al. [28] used Multi-Layer Perceptron (MLP), which is another type of neural networks to detect personal attacks. The research showed that compared to Logistic Regression, MLP had higher AUC of 0.9659 with character n-gram features [28].

Hate speech, which is a type of toxicity, was the focus of Badjatiya et al. [5], where they attempted to use a variety of deep learning techniques to detect hate on Twitter. Using a collection of 16,000 tweets, the researchers compared the performance of CNN and Long Short Term Memory (LSTM) networks with bag-of-words, character n-grams, and TF-IDF features. The evaluation compared the performance of the deep learning methods with classifiers like SVMs, Random Forest, Gradient Boosted Decision Trees (GBDTs), and logistic regression. The results showed that deep learning methods (LSTM in particular) outperform character and word-based methods by around 18 F_1 points [5].

2.2 Toxic Behavior

One of the major challenges that come as a natural result of toxicity is toxic behavior. Whenever toxicity occurs online, it is usually associated with toxic behavior that stems from an illicit act committed by users. This problem is prevalent in online multiplayer games, where users collaborate in a social setting to accomplish a common goal (i.e., win trophies or complete tasks). To combat abusive behavior, companies that develop online multiplayer games with social capabilities must deal with users that exhibit toxic behavior [6, 17]. Despite the rules and regulations that developers employ in gaming platforms, there is still a need to report misconduct to protect other players. This approach happens when the offended player reports the abusive behavior. The problem with reporting offense is that there is no way to check if the report corresponds to an actual abusive behavior or not. To tackle this problem, Balci and Salah [6] developed a set of tools to validate if offenses that contain verbal aggression happened in games or not. The study was on a popular social game called Okey, where the goal was to identify the characteristics of abusive players. The outcomes of the research show that the developed system can identify abusive players with an accuracy of 85%. In the cases of severe abuse, the system performs at a higher accuracy in confining abusive players [6].

Neto et al. [22] introduced another study on toxic behavior in collaborative online games. The popular battle game League of Legends (LoL) exhibits toxic behavior when players refrain from socializing with teammates, which leads to conflict among players. The research focused on the engagement patterns among players during games and their association with toxic behavior. Findings show indeed different communication patterns during toxic behavior, and they are associated with the level of toxicity and gaming performance. The study adds metrics for measuring the performance of players along with the level of toxic contamination that computes the negativity of toxic behavior. Outcomes of the research show that stress and low game performance contribute to toxic behavior [22].

2.3 Online Moderation

In some online platforms, users build communities to interact with each other and discuss topics of interest. A crucial challenge in this type of setting is moderation. Seering et al. [25] proposed three processes to explain how communities evolve with the aid of their moderator's engagement. The processes explain how users become moderators; the types of actions, responses, and tasks assigned to them; and the rules that foster the development of communities. The study focused on the contributions of algorithmic and non-algorithmic tools to support moderators in developing online communities. When it comes to content moderation, scholars investigated the effect of displaying moderation rules on the perceived bias in news [29], and the spread of harassment in online discussions. Both studies showed that displaying moderation rules increase the adherence to regulations, increase positive participation in discussions, and potentially reduce bias in the perception of science-related news. As for the moderation rules that govern

[1] https://www.perspectiveapi.com/

online communities, Gibson [12] studied two communities with similar size and topic but with different moderation rules. The study showed that when the community is labeled as safe space, moderators and users remove more posts, signifying a higher level of self-censorship. Moreover, the language used in the safe space community focuses mostly on positive emotion, like leisure activities. On the other hand, the free space community showed higher rates of negative emotions, mostly related to anger associated with work, fatality, and financial concerns. The findings of the study suggest that different governance rules across communities are essential at preserving democracy in online communities.

In another vein of studies on the effectiveness of different moderation styles, Matzat and Rooks [21] experimented with direct and indirect moderation in online health communities on Yahoo!. Their findings showed that users prefer indirect moderation that provides incentives for positive interaction as opposed to direct moderation that penalizes users for their contributions. Similarly, Lander [19] found that in online learning communities, implicit moderation strategies are more effective than explicit practices. Additionally, engagement plays a crucial role in providing users with different viewpoints and opportunities, as opposed to introducing instructions to users. To understand the relationship between collaborative moderation efforts and their ability to resolve conflicts in online communities, Hauser et al. [14] developed an agent-based simulation model of conflicts in firm online communities. The study found that the characteristics of participants and the social structure of online communities play a vital role in effectively handling firestorms in social media. By adapting to individual and community-level characteristics, firms can adjust collaborative moderation practices to handle conflicts effectively.

3 METHODOLOGY

To perform our investigation of toxicity on a wide spectrum of communities, users, and moderators, we split our problem into smaller sub-problems that target toxicity, users, and moderators separately. First, we begin with defining the requirements of the study and the outcomes of each sub-problem. Then, we analyze all the available resources, which include any data, techniques, algorithms, and design approaches to choose the best options for our sub-problems. After studying the available options and possible approaches, we design solutions to every sub-problem based on predefined requirements and specific research questions. Lastly, we connect the outcomes of the sub-problems to cover all the aspects of investigating toxicity on Reddit.

4 PROPOSED APPROACH

In the following subsections, we detail our approach for investigating toxicity from different angles.

4.1 Detecting Toxicity and Toxicity Triggers

The first sub-problem pertains to identifying toxic content and toxicity triggers across multiple communities. To address this problem, we aimed at answering the following research question:
research question: *Can we predict toxicity triggers in discussion threads?.*
Since our main focus is detecting causes of toxicity, we should first identify toxic comments in discussion threads to detect their triggers. Due to our specific requirements for detecting toxicity triggers, we could not find a labeled dataset from Reddit. Therefore, we constructed a labeled collection of comments for the task of toxic comment detection. We relied on Figure Eight crowdsourcing platform (formerly known as CrowdFlower) to label a random sample of 10,100 comments from r/AskReddit, one of the largest communities on Reddit. Then, we used neural networks to build a Long Short Term Memory (LSTM) model with pre-trained GloVe word embeddings [24] to predict the toxicity of comments from the top 10 subreddits on Reddit. We used the results of the prediction to construct discussion threads with parent and child comments to detect toxicity triggers. Additionally, we computed trigger-specific features and used the same LSTM model to detect toxicity triggers [3]. Our investigation of the literature showed other features, like troll detection [9], that can improve the trigger detection model. Hence, as future work, we plan to:
DT.1 Expand our approach to incorporate more communities on Reddit.
DT.2 Incorporate other features to the toxicity trigger detection model to improve the performance.

4.2 Predicting Usesrs Toxic Posting Behavior

After studying the toxicity of comments and detecting their triggers, we formulate our second sub-question by focusing on the toxicity of users based on their postings across multiple communities. To solve this problem, we propose the following research question:
research question: *How can we detect and judge the toxicity of users that post in different communities?*
To address this problem, we use the same labeled dataset that we built previously to detect the toxicity of user's comments across multiple communities. Then, we use the prediction results to identify and judge toxic and non-toxic users. Within this line of work, we plan to:
UT.1 Include posting behavior patterns in the study of user's toxicity across both submissions and comments in Reddit.

4.3 Moderation of Reddit Communities

Moderators hold a vital role in Reddit's ecosystem. They are considered users with special privileges that enable them to monitor the safety of communities and prevent toxicity. At times, moderators stand at the receiving end of harassment, hate, and toxicity. Therefore, in the third sub-question of our study, we propose the following research question:
research question: *Can the moderation style and aspects of moderators predict if moderators get harassed?*
To get insights from moderators on Reddit, we designed a web-based survey and gathered the responses of 1,818 moderators from different communities on Reddit. Then, we analyzed their responses to get insights on the varying moderation styles and their characteristics. The insights from this work motivated us to plan the following:
MR.1 Conduct voice-based interviews with volunteer moderators from varying communities.
MR.2 Code the responses of moderators and extract additional insights to incorporate with the survey results.

MR.3 Identify the causes of harassment on moderators in Reddit and link it to toxic user behavior and toxic content.

5 RESULTS

In the following subsections, we present our detailed results that target each sub-question and present the evaluation method of each proposed solution within the frame of each sub-study.

5.1 Detecting Toxicity and Toxicity Triggers

Previously, we mentioned that in order to detect toxicity triggers, we proposed additional features to the neural network model, which include **topical shift** and **sentiment shift**. Since emotional and topical changes in discussions could be indicators of hateful comments [27], we can incorporate topical and emotional shifts as features to predict toxicity triggers. First, to measure topical shift, we computed the similarity between non-toxic parent comments and each child comment in the discussion thread using the cosine similarity between their vector representations. Then, we used k-means clustering to determine if the comments were on-topic or off-topic. By constructing two clusters that denote on-topic and off-topic comments, we considered the smallest centroid of clusters to be an indicator of comments that exhibit topical shift [27]. As for sentiment shift, we used the AFINN's lexicon [13] to score each comment's sentiment. Then, similar to topical shift, we used K-means clustering to detect the shift in sentiment. To detect toxicity triggers, we used the LSTM neural network mentioned in the previous section. Initially, we detected toxicity triggers with GloVe word embeddings. Then, we added topical and sentiment shift features to the model. Lastly, we combined topical and sentiment shift features with GloVe embeddings. The achieved average accuracy of the model, given in Table 2, was 82.5%, which shows a 4% improvement over the baseline model. This result indicates that topical and sentiment shift features improve the detection of toxicity triggers [2].

Table 1: Performance of the LSTM models. Sent.=sentiment

Features	ROC-AUC	Accuracy	Macro F_1
GloVe (Baseline)	0.87	78.5%	0.78
GloVe+Topic	0.88	79.1%	0.79
GloVe+Sent.	0.90	81.9%	0.82
GloVe+Topic+Sent.	**0.91**	**82.5%**	**0.83**

To characterize toxicity triggers, we investigated 270,320 toxicity triggers predicted from the top 10 subreddits and compared them with non-triggers in terms of the frequency of appearing keywords. When we examined the top 10 most frequent words from each class [2], we found that trigger comments contain controversial or provocative terms like *tax*, *vote*, and *Israel*. While non-trigger comments contain fairly mild words like *thank*, *help*, and *quest*.

5.2 Predicting Users' Toxic Posting Behavior

To predict the toxic posting behavior of users, we used an improved version of the bidirectional LSTM model with GloVe features. Moreover, we incorporated two levels of toxicity that indicate highly-toxic comments and slightly-toxic comments. Since our full dataset includes all submissions and comments from the top 100 subreddits, our final dataset consisted of 14,741,169 submissions (i.e., posts) and 226,005,117 comments. To predict the toxicity of submissions, we concatenated the title of the submission and the body (if it exists), while the comment text was used for performing the prediction. Running the prediction model on the entire submissions collection resulted in a total of 190,220 highly-toxic submissions (1.29%) and 515,868 slightly-toxic submissions (3.50%). As for the comments collection, the model predicted 11,556,260 (5.11%) highly-toxic comments and 18,402,307 (8.14%) slightly-toxic comments.

Table 2: Performance of the neural network models in terms of accuracy, macro F_1, and ROC-AUC score*

Neural network models	ROC-AUC	Accuracy	Macro F1
GRU+fasttext	0.94	87.6%	0.88
GRU+GloVe	0.96	90.0%	0.90
Improved LSTM+GloVe	**0.99**	**97.1%**	**0.97**

* The results are obtained by taking the average of 10 random runs.

5.3 Moderation of Reddit Communities

To study the impact of moderation style on the targeted harassment of moderators, we analyzed 1,818 responses to the moderation survey [1]. Then, we converted the textual responses to numerical values for the regression analysis. Afterward, we build an ordinary least square regression (OLS) model to study the influences of the practices of a moderator, which are represented as multiple independent variables on the level of harassment against the moderator as the dependent variable. The results of the regression model are given in Table 3. We find six significant variables ($p < .05$) in Table 3. While only one variable, harassmentModsStop, decreases the level of harassment against a moderator, other variables increase the harassment.

6 CONCLUSION AND FUTURE WORK

In this work, we described our approach in investigating toxicity in multiple Reddit communities, users, and moderators. Our main research goal is to detect toxic content and toxicity triggers, use it to detect user's toxic posting behavior, and discover insights from moderators to detect targeted harassment attacks by users towards Reddit moderators. To achieve our goal, we presented a method for detecting toxicity and toxicity triggers in online discussions on Reddit [2] to solve the problem of toxic content in online communities. We built an LSTM neural network model [30] to judge the toxicity of users across multiple Reddit communities. We designed a web-based survey on Reddit moderation practices [12] and harassment against moderators to discover the relationship between certain moderation practices and targeted moderator harassment. The current plan for this research is to focus on the moderation side and to extend the toxicity triggers study to include more subreddits and incorporate additional features for detecting toxicity triggers. Currently, we are working on the extension of our toxicity triggers study by extracting discussions from the top 100 subreddits in Reddit. We plan to improve our toxicity detection model

Table 3: The effects of the independent variables on the harassment of moderators.

	coef.	std. err.	*p*-value
importanceToxicity	0.0824	0.024	***
harassmentModsStop	-0.1416	0.022	***
KnowledgeOfRules	0.2441	0.034	***
RulesFairness	-0.0037	0.049	
StrictnessOfRules	-0.0266	0.028	
UsersViolation	0.1827	0.029	***
RateOfSubToxicity	0.2365	0.029	
ModSubCount	0.0266	0.023	
ToxicitySimilarity	-0.0132	0.021	
ToolsUsage	0.0801	0.051	
ToolsHelpfulness	0.1104	0.038	**
Gender	-0.0474	0.027	*
Age	0.1257	0.032	***
R-squared	0.741		

Note: *** : p < 0.01, ** : p < 0.05, *: p < 0.10.

by incorporating features that detect spam [15] and trolling [9] to facilitate the detection of toxicity triggers.

ACKNOWLEDGMENTS

I would like to thank my supervisor Dr. Jim Jansen and my co-supervisor Dr. Haewoon Kwak for their valuable feedback and continuous guidance.

REFERENCES

[1] Hind Almerekhi, Haewoon Kwak, , and Bernard J. Jansen. 2020. Statistical Modeling of Harassment against Reddit Moderators. In *Companion Proceedings of the Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE.

[2] Hind Almerekhi, Haewoon Kwak, Bernard J. Jansen, and Joni Salminen. 2019. Detecting Toxicity Triggers in Online Discussions. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media* (Hof, Germany) *(HT '19)*. Association for Computing Machinery, New York, NY, USA, 291–292.

[3] Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J. Jansen. 2020. Are These Comments Triggering? Predicting Triggers of Toxicity in Online Discussions. In *Proceedings of the 29th International Conference on World Wide Web* (Taipei, Taiwan) *(WWW '20)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE.

[4] Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1100–1105.

[5] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 759–760.

[6] Koray Balci and Albert Ali Salah. 2015. Automatic analysis and identification of verbal aggression and abusive behaviors for online social games. *Computers in Human Behavior* 53 (2015), 517–526.

[7] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3175–3187.

[8] Jonathan Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 184–195.

[9] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities.

[10] Maeve Duggan. 2017. Online harassment 2017. (2017).

[11] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. 2018. Convolutional Neural Networks for Toxic Comment Classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence* (Patras, Greece) *(SETN '18)*. Association for Computing Machinery, New York, NY, USA, Article Article 35, 6 pages.

[12] Anna Gibson. 2019. Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces. *Social Media + Society* 5, 1 (2019), 2056305119832588.

[13] Lars Kai Hansen, Adam Arvidsson, Finn Aarup Nielsen, Elanor Colleoni, and Michael Etter. 2011. Good Friends, Bad News - Affect and Virality in Twitter. In *Future Information Technology*, James J. Park, Laurence T. Yang, and Changhoon Lee (Eds.). Springer, Berlin, Heidelberg, 34–43.

[14] Florian Hauser, Julia Hautz, Katja Hutter, and Johann Füller. 2017. Firestorms: Modeling conflict diffusion and management strategies in online communities. *The Journal of Strategic Information Systems* 26, 4 (2017), 285 – 321.

[15] Xia Hu, Jiliang Tang, Yanchao Zhang, and Huan Liu. 2013. Social spammer detection in microblogging. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

[16] Prerna Juneja, Deepika Rama Subramanian, and Tanushree Mitra. 2020. Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices. *Proc. ACM Hum.-Comput. Interact.* 4, GROUP, Article Article 17 (Jan. 2020), 35 pages.

[17] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3739–3748.

[18] Cheng-Yu Lai and Chia-Hua Tsai. 2016. Cyberbullying in the Social Networking Sites: An Online Disinhibition Effect Perspective. In *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016* (Union, NJ, USA) *(MISNC, SI, DS 2016)*. Association for Computing Machinery, New York, NY, USA, Article Article 4, 6 pages.

[19] Jo Lander. 2015. Building community in online discussion: A case study of moderator strategies. *Linguistics and Education* 29 (2015), 107 – 120.

[20] Chenyi Lei, Shouling Ji, and Zhao Li. 2019. TiSSA: A Time Slice Self-Attention Approach for Modeling Sequential User Behaviors. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2964–2970.

[21] U. Matzat and G. Rooks. 2014. Styles of moderation in online health and support communities: An experimental comparison of their acceptance and effectiveness. *Computers in Human Behavior* 36 (2014), 65 – 75.

[22] Joaquim A. M. Neto, Kazuki M. Yokoyama, and Karin Becker. 2017. Studying Toxic Behavior Influence and Player Chat in an Online Video Game. In *Proceedings of the International Conference on Web Intelligence* (Leipzig, Germany) *(WI '17)*. Association for Computing Machinery, New York, NY, USA, 26–33.

[23] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web* (Montréal, Québec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 145–153.

[24] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[25] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443.

[26] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media.

[27] K. Topal, M. Koyuturk, and G. Ozsoyoglu. 2016. Emotion -and area-driven topic shift analysis in social media discussions. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 510–518.

[28] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1391–1399.

[29] Sara K. Yeo, Leona Yi-Fan Su, Dietram A. Scheufele, Dominique Brossard, Michael A. Xenos, and Elizabeth A. Corley. 2019. The effect of comment moderation on perceived bias in science news. *Information, Communication & Society* 22, 1 (2019), 129–146.

[30] Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 3485–3495.