Detecting Toxicity Triggers in Online Discussions

Hind Almerekhi
Hamad Bin Khalifa University
Doha, Qatar
hialmerekhi@mail.hbku.edu.qa

Haewoon Kwak
Qatar Computing Research Institute, HBKU
Doha, Qatar
haewoon@acm.org

Bernard J. Jansen
Qatar Computing Research Institute, HBKU
Doha, Qatar
jjansen@acm.org

Joni Salminen
Qatar Computing Research Institute, HBKU
Doha, Qatar
jsalminen@hbku.edu.qa

ABSTRACT

Despite the considerable interest in the detection of toxic comments, there has been little research investigating the causes – i.e., triggers – of toxicity. In this work, we first propose a formal definition of triggers of toxicity in online communities. We proceed to build an LSTM neural network model using textual features of comments, and then, based on a comprehensive review of previous literature, we incorporate topical and sentiment shift in interactions as features. Our model achieves an average accuracy of 82.5% of detecting toxicity triggers from diverse Reddit communities.

CCS CONCEPTS

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

KEYWORDS

Reddit; toxicity; trigger detection; neural networks; social media

ACM Reference Format:
Hind Almerekhi, Haewoon Kwak, Bernard J. Jansen, and Joni Salminen. 2019. Detecting Toxicity Triggers in Online Discussions. In *30th ACM Conference on Hypertext & Social Media (HT '19), September 17–20, 2019, Hof, Germany.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3342220.3344933

1 INTRODUCTION

Online social media platforms, such as Facebook and YouTube, enable users to establish communities of shared topics of interest [3]. However, as users engage in online discussions, communities face the challenge of monitoring the civility of discussions. This challenge is often related to controlling toxic comments that contain rudeness or harassment towards various targets [3]. The presence of toxicity makes it more difficult for users to interact and freely express their opinions, thus contaminating online discussions and creating unpleasant user experiences.

However, there has been surprisingly little research into what causes or sparks toxic online discussions. Detecting these toxicity *triggers* is vital because online toxicity has a contagious nature [1], and thus, one toxic comment has the potential to attract more toxic comments quickly. To cut such vicious cycles short, or even prevent them from starting, detection of toxicity triggers is a worthy and practically impactful research goal. To achieve this goal, more understanding of the *initiation* of toxic conversations via identification of toxicity trigger is needed.

In this work, we detect toxicity triggers (i.e., the starting points) leading to toxic online discussion threads in Reddit.

We define toxicity triggers in the context of online discussions as comments that incur direct toxic responses. As these triggers of toxicity may differ by the community and by topic due to different norms and uses of language [5], we study diverse communities in Reddit. We pose two research questions for characterizing and predicting toxicity triggers: *RQ1:* What are the characteristics of potential toxicity triggers in online discussions? *RQ2:* Can we predict toxicity triggers based on their characteristics? To address these research questions, we begin by analyzing an extensive collection of more than 104 million comments collected from Reddit during a period spanning nearly two years.

2 DATA COLLECTION AND PREPROCESSING

In this study, we focused on the ten subreddits with the highest number of subscribers[1]. For each subreddit, we retrieved all the comments posted between January 2016 and August 2017 using Pushshift's public Reddit collection[2]. Then, we constructed discussion threads using the ID and parent ID of the corresponding comment. We ended up with an extensive collection of comments and discussions from the top 10 subreddits on Reddit.

2.1 Toxicity Detection

As the toxicity of child comments defines toxicity triggers, we first need to identify toxic comments. While there are several toxic comment datasets from other online communities, as we mentioned earlier, toxic comments may differ across the communities due to different norms. We thus build a toxic comment dataset for Reddit by ourselves. We used Figure Eight[3] to collect labels for 10,100 randomly sampled comments from AskReddit. Then, we built a Long Short Term Memory (LSTM) neural network model using pre-trained word embeddings from GloVe [2]. To evaluate

HT '19, September 17–20, 2019, Hof, Germany

ACM ISBN 978-1-4503-6885-8/19/09.
https://doi.org/10.1145/3342220.3344933

[1] http://redditlist.com, retrieved on 21 August 2017.
[2] https://files.pushshift.io/reddit/comments/
[3] https://www.figure-eight.com

the performance of the model, we computed the average scores of five random runs. The results showed that the average ROC-AUC was 0.98, the average model accuracy was 93.7%, and the average F_1 score was 0.94.

To address RQ1, we predicted the toxicity of comments in the years 2016 and 2017 and then computed the percentage of toxic and non-toxic comments in each subreddit. The results show that the subreddit videos was the most toxic in the year 2016, while the subreddit "worldnews" is the most toxic in the year 2017. These findings indicate that toxicity is prevalent in Reddit communities; therefore, detecting toxicity triggers is a considerable problem.

3 DETECTING TOXICITY TRIGGERS

3.1 Shift Features

In addition to word embeddings, we also consider topical and sentiment shift in comments as features. Changes in the topic or the overall emotion of a discussion could tell us something about the comments that might trigger toxicity [4]. This insight motivated us to investigate topical and emotional shift as features to predict toxicity triggers.

Topical Shift: We measured the topical similarity between the non-toxic parent comment and each child comment that came after it in the discussion thread by computing 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 [4].

Sentiment Shift: To study the sentiment of each comment within the discussion thread, we used AFINN's lexicon [1] to score each comment's sentiment. Then, similar to topical shift, we used K-means clustering to detect sentiment shift.

3.2 Toxicity Trigger Detection

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 1, 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.

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 better understand toxicity triggers, we studied 270,320 toxicity triggers predicted from the top 10 subreddits and compared them with non-triggers in terms of the frequency of appearing keywords. By examining the top 10 most frequent words from each class, we can tell that trigger comments typically contain controversial or provocative terms like *tax*, *vote*, and *Israel*. While non-trigger comments contain fairly mild words like *thank*, *help*, and *quest*. Furthermore, we noticed that most of the words in the triggers are political. This finding indicates that handling toxicity triggers is the key to better political discussions online.

Table 2: The 10-most frequent words in toxicity triggers and non-triggers

Trigger	muslim, israel, kill, europe, support, tax, vote, law, president, woman
Non-trigger	film, quest, ask, answer, system, story, best, thank, help, character

4 DISCUSSION AND CONCLUSION

By manually examining incorrectly classified toxicity triggers, we found that classification errors are usually caused by a) the lack of context in the conversation thread, and b) incorrect toxic-comment classification results, which makes it difficult to detect if the parent comment triggers toxicity or not. These observations shed light on some of the challenges associated with toxicity trigger detection and open areas for future work, like incorporating additional features [6] into the toxicity trigger detection model – e.g., semantic shift [4] and the discussion context [7].

In summary, we defined toxicity triggers and detected them from diverse Reddit communities. For that, we built an LSTM model that achieved an average accuracy of 82.5% by combining shift and embedding features. Our approach shows novelty by being, to our knowledge, the first study that examines online toxicity triggers by using an extensive collection of online discussions. For future work, we will conduct more comprehensive studies on toxicity triggers to cover extended periods and include more subreddits.

REFERENCES

[1] 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*. Springer, Berlin, Heidelberg, 34–43.

[2] 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.

[3] Joni Salminen, Hind Almerekhi, Milica Milenkovic, Soon-gyo Jung, Jisun An, Haewoon Kwak, and B. J. Jansen. 2018. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. In *Proceeding of The International AAAI Conference on Web and Social Media (ICWSM 2018)*.

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

[5] T. Weninger, X. A. Zhu, and J. Han. 2013. An exploration of discussion threads in social news sites: A case study of the Reddit community. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

[6] Frank Z. Xing, Filippo Pallucchini, and Erik Cambria. 2019. Cognitive-inspired domain adaptation of sentiment lexicons. *Information Processing & Management* 56, 3 (2019), 554 – 564.

[7] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. 1350–1361.