Are These Comments Triggering? Predicting Triggers of Toxicity 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

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

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

ABSTRACT

Understanding the causes or triggers of toxicity adds a new dimension to the prevention of toxic behavior in online discussions. In this research, we define *toxicity triggers* in online discussions as a non-toxic comment that lead to toxic replies. Then, we build a neural network-based prediction model for toxicity trigger. The prediction model incorporates text-based features and derived features from previous studies that pertain to shifts in sentiment, topic flow, and discussion context. Our findings show that triggers of toxicity contain identifiable features and that incorporating shift features with the discussion context can be detected with a ROC-AUC score of 0.87. We discuss implications for online communities and also possible further analysis of online toxicity and its root causes.

CCS CONCEPTS

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

KEYWORDS

Reddit, toxicity, trigger detection, neural networks, online discussion

ACM Reference Format:
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 Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3366423.3380074

1 INTRODUCTION

Social media has revolutionized how users communicate with each other, as it enables information exchange with great ease. Online platforms, such as Reddit, YouTube, and Facebook, allow users to discuss various topics by building communities around shared interests [29]. With the increase of user engagement in online communities, moderating discussions for incivility becomes more challenging. These challenges stem from the difficulty in controlling toxic posts that target other users with various types of harassment and rudeness [29, 30]. The manifestation of toxic posts in online discussions discourages users from having healthy interactions, leading to conflicts and unpleasant user experiences [22].

There are multiple reasons why toxicity occurs, relating to motivations linked to cyberbullying, trolling, and socio-dynamic effects, such as group polarization [48, 50]. Various approaches have been suggested to combat online harassment, including counter-speech and creation of community guidelines [47]. One common approach is automatic detection and moderation of toxic comments [31, 44, 45]. Automation can be seen especially useful for active communities where users send thousands of messages daily, making manual moderation extremely challenging.

Despite the scholarly interest in detecting toxicity in online posts, studies that investigate the causes of such toxic posts within online discussions seem scarce [5]. To stop malicious posting from reoccurring or even prevent them from the beginning, detection of toxicity triggers is a novel and impactful research goal. To achieve this goal, understanding of the *initiation* of toxic discussions through the detection of toxicity triggers is required [5].

In this work, we identify toxicity triggers (i.e., the initiating actions), leading to toxicity in discussion threads. We define toxicity triggers as *non-toxic initiators* sparking following toxic interaction. We formulate our research questions for detecting toxicity triggers as follows: *Can we predict toxicity triggers that are likely to lead to toxic replies?* To address this research question, we analyze an extensive collection of more than 104 million Reddit comments during a period spanning nearly two years. We start with detecting toxic comments in our dataset and reconstructing discussion threads from the detected toxic comments. From the reconstructed discussion threads, we find toxicity triggers based on their definition that they are non-toxic, but their child (i.e., reply) comments are toxic.

To predict whether a given comment is a toxicity trigger, we incorporate its textual features and context feature. The context refers to the discussion that happened before that comment. Additionally, from a comprehensive review of previous literature, we extract two specific features from the context: topic shift [40] and emotional shift [33]. We then build a prediction model and find the toxicity triggers across multiple subreddits.

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

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380074

The contribution of this work is as follows: We formulate a novel problem of predicting toxicity triggers and build a neural network-based prediction model based on previous literature of online communications. Our model achieves a high ROC-AUC of 0.87 across multiple subreddits, showing its generalizability.

2 RELATED WORK

2.1 Toxicity Detection

There is a high level of interest in investigating online hate and toxicity, as shown by the number of peer-reviewed publications on the subject (see [38] for a review). The problem of toxicity is prevalent in the wider society. For example, a 2016 report [6] states that 66% of the harassment that users report takes place on social media and that 21% of the harassment victims refrain from using all forms of social media. This report shows that toxicity can deter users from participating in online discussions and can negatively affect users' online social experience. Other studies focused on cyberbullying show similar findings [11, 18, 23], and individuals' freedom to participate in online discussions become *de facto* reduced by the presence of toxicity.

For toxic comment detection, the goal is generally to classify comments as either toxic or non-toxic for further actions such as deletion or analysis. Researchers have used machine learning models [31], such as deep neural networks [7] in a variety of ways to detect toxic comments. For instance, Nobata et al. (2016) used syntactic 2012, linguistic [44], distributional semantics (embedding) [50], and n-gram [45] features to detect abusive and non-abusive comments. Their research showed that with variations of NLP features, the regression model outperforms a state-of-the-art deep learning model by not maintaining word embeddings in each iteration.

Similarly, Wulczyn et al. (2017) deploy machine learning models to detect toxicity at different levels, defined as targeted personal attacks. The features in this study were extracted from the comments text and did not exploit information associated with users or the communication network. Comparable to [31], 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 [48]. Salminen et al. (2018) found that, in 137,098 comments from YouTube and Facebook videos in an online news media, most targets of hate were politicians and the media. Additionally, Spiros et al. (2018) used the same dataset as [48] to experiment with Convolutional Neural Networks (CNNs) for toxic comment classification. The outcomes of the study show that CNNs can outperform classic bag-of-word text classification techniques.

Finally, a prominent effort for toxic comment detection is the Perspective API [43] aimed at providing platform owners and researchers with a tool that scores comments based on their level of toxicity. The models used in the PerspectiveAPI were built using machine learning techniques from an extensive collection of comments labeled by respondents of online surveys. Each model returns a score between 0 and 1, indicating the toxicity of the comment. Some of the experimental models offered by the Perspective API classify severe toxicity, obscenity, spam, and attacks on people. The experimental models of the Perspective API receive frequent revisions and updates that have improved its performance over time [43].

Overall, prior research focused on several aspects of toxic conversation using different angles and techniques to approach the problem and often focusing on the detection and classification of toxic comments. However, no prior work that we could locate has specifically focused on the detection of triggers of toxic online discussions. Our approach, therefore, addresses an open and vital research question with a diverse set of computational techniques and implications in a variety of domains.

2.2 Contagion of Toxicity

Apart from experimenting with online hate detection, prior research has established that contagion of toxicity is a considerable contamination risk for the health of online discussions [29]. Del Vicario et al. (2016) referred to this effect as emotional contagion. Kwon and Gruzd (2017) investigated how hatefulness formed a cascading effect and discovered that when parent comments in a discussion thread include swearing, the following child comments exhibit a higher degree of swearing as well.

The concept of topic shift is relevant to the problem of toxicity trigger analysis because the emotions of commentators [40] profoundly influence the tone of social media discussions. For example, Topal et al. (2016) provide indications that a topic shift could be an indicator of emotional shifts in online discussions. However, this premise was not explored in the context of toxicity. Moreover, the impact of tone shifts has not been studied using a large corpus of online discussions, such as the one in this research.

Therefore, our premise is that for toxicity to spread in conversations, there might be causes - *triggers* - that turn healthy conversations toxic. These triggers of toxicity may differ by the community and by topic due to different norms and uses of language [35, 46]. We could locate no prior research specifically on toxicity triggers in online discussion threads. However, in information behavior research, the concept of 'trigger' has been used to describe what causes users to initiate information-seeking activity [32]. Reddy and Jansen (2008) considered 'trigger' as a point for transitioning from individual to collaborative information-seeking behavior. This triggering behavior requires interaction, defined as a series of events that need at least two actions and two actors [42].

3 REDDIT DATA COLLECTION AND PREPROCESSING

Reddit is an online community with more than one million [37] communities, which are commonly known as "subreddits", denoted as "r/" in the platform. Overall, Reddit is considered a 'community of communities' where each subreddit is moderated independently; hence, it incorporates a mix of cultures [28].

We focused on the ten subreddits with the highest number of subscribers as of 2017 [4] , as shown in stage (a) of Figure 1 and Table 1. Then, for each subreddit, we retrieved all the comments posted between January 2016 and August 2017 (stage (b) in Figure 1). Using Pushshift's public Reddit collection [3] , we retrieved the comments and ensured that the comment objects included a)

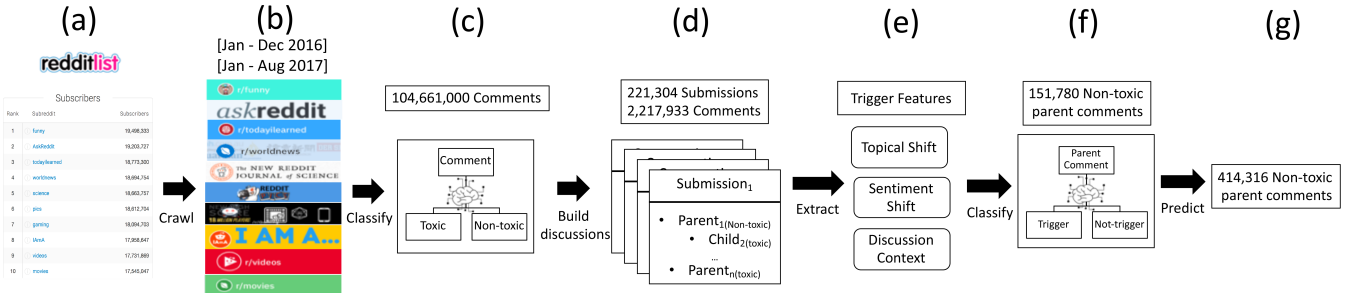

Figure 1: Pipeline for investigating toxicity triggers in Reddit discussions.

Table 1: Summary of our data collection. The Subreddits are sorted by the number of subscribers.

Subreddit	Jan - Dec 2016	Jan - Aug 2017
r/funny	6,471,278	3,598,968
r/AskReddit	52,231,835	33,652,845
r/todayilearned	214,206	221,734
r/science	474,845	161,174
r/worldnews	1,986,401	876,046
r/pics	69,713	134,872
r/IAmA	140,637	57,468
r/gaming	65,496	68,037
r/videos	527,500	222,529
r/movies	3,292,659	192,757
Total	**65,474,570**	**39,186,430**

the time stamp, b) ID of the comment, c) ID of the comment's parent, and d) the comment text. After collecting the comments, we removed the comments shown as deleted as well as comments that were removed by moderators. Then, we constructed discussion threads from the comments using the ID and parent ID of the corresponding comment. Following this procedure, we removed any comments that did not belong to a thread (i.e., parent comments with no children). We ended up with an extensive collection (i.e., corpus) of online discussions from the top 10 subreddits on Reddit, as shown by the statistics in Table 1.

4 TOXICITY DETECTION

To fully understand toxicity triggers in online discussions and build a statistical model to detect them, we propose an analysis framework as in Figure 1. In this section, we explain how to detect toxicity.

4.1 Building the Ground-Truth

To perform toxic comment detection, the first approach that we tried was scoring comments through Google's Perspective API [2]. We scored 1,000 comments that we randomly sampled from our collection and examined the retrieved scores manually for their accuracy. Unfortunately, as the API was not stable at the time of our experiments in 2018, we could not rely on the scores obtained from the API. Also, as the Perspective API has a rate-limit of sending requests, it is infeasible to label each of 104 million comments through the API. To overcome this limitation, we built our own toxic comment detection model based on comments sampled from our collection.

To build our model for toxic comment detection, we required a labeled collection of toxic and non-toxic comments to train the model. Since there is no Reddit-specific test collection for the task of toxic comment detection, we opted to create our own labeled collection. For this task, we used Figure Eight [1] to collect judgment labels for the 10,100 randomly sampled comments from r/AskReddit. The reason we choose r/AskReddit is that it is one of the largest and most popular question-answering communities on Reddit that deals with a wide range of types of topics [25]. Therefore, it is reasonable to assume that r/AskReddit captures various behaviors of Reddit users and minimizes the risk of dealing with subreddit-specific keywords. We verify this assumption by evaluating our model with sampled comments from multiple subreddits §4.3.

The designed labeling job required workers to label a given comment as either toxic or non-toxic. The designed labeling job required workers to label a given comment as either toxic or non-toxic. To help workers understand what toxicity means, we used the definition of the Perspective API, which states that a toxic comment is "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

The results from the labeling task showed that 81.57% of the 10,100 comments are non-toxic, while the remaining 18.43% are toxic. As this is the dataset of highly imbalanced, we measure the agreement between annotators by using Gwet's gamma, which takes the average-distribution approach that outperforms the Kappa statistic measurements in handling class imbalance [15]. Gwet's gamma score was 0.70, which is considered high.

4.2 Training the Model

Then, we build models that can classify comments into toxic or non-toxic, as shown in stage (c) of Figure 1. For this task, we employed a variety of features that capture the semantic and syntactic properties of the comments. First, we looked at n-gram features at different configurations [19]; then, we incorporated an advanced set of features based on word embeddings such as word2vec [21] and sentence2vec [27]. We also extracted 13 content-based features used in [38], such as the number of emoticons, capital letters, and misspellings.

The toxic comment classification experiment takes into account several issues that persisted in the collection, such as the skewness

of the classes. To solve the problem of class imbalance, we used a technique that combines Synthetic Minority Over-sampling TEchnique (SMOTE) and Tomek Links (an under-sampling method) [8]. This technique performs over-sampling using SMOTE and cleaning with Tomek links; then, we performed feature transformation by scaling each feature to a range between 0 and 1.

The next step in the classification pipeline was to perform feature selection. For this purpose, we used the Random Forest algorithm to perform the classification, ranking, and selection of features. Feature selection is essential in reducing the dimensionality of the large feature vectors, such as the highly dimensional N-gram features, and it can be useful in many algorithmic approaches. The last step in the classification experiment was to perform the classification using 10-fold cross-validation and grid search for parameter tuning. As a result, we extracted 8,623 features.

We tested a variety of classifiers to choose the best performing model, including Decision Tree, Random Forest, Adaboost, and Logistic Regression. Moreover, we experimented with two neural network models: Bidirectional Encoder Representations from Transformers (BERT) in its vanilla form with fine tuning [13], and Long Short Term Memory (LSTM) pre-trained word embeddings from GloVe (840B tokens, 2.2M vocabulary terms, cased, 300d) [34]. For BERT, we used the uncased version of the vocabulary model to train the neural network on the relationship between sentences. For the fine-tuning step, we adjusted the maximum sequence length to 128 characters, used a batch size of 32, a learning rate of 0.00003, and a total of 4 epochs. For LSTM, we used 'Nadam' as our optimizer. The training task ran in batches of size 64 on a total of 3 epochs. The maximum features for the GloVe embeddings were 30,000, and the maximum length of each comment was 400 words. Table 2 shows the max performance of the models we tested. As the LSTM-based model outperforms other models, we chose the LSTM-based model to performs the toxicity prediction for the following experiments.

Table 2: Max Performance of each model

Model	F_1
Decision Tree	0.85
Random Forest	0.86
Logistic Regression	0.74
Adaboost	0.91
BERT-based	0.81
LSTM-based	**0.94**

4.3 Additional Model Evaluation with Other Subreddits

While we assume that comments from r/AskReddit are general enough to study toxicity in Reddit, that assumption should be tested with the actual data. Therefore, we design an additional evaluation with comments that are randomly sampled from other subreddits. To evaluate the predictive accuracy of the model, we computed a simple agreement score on 1,000 random comments from the other nine subreddits (excluding r/Askreddit). We used crowdsourcing (Figure Eight) to label the comments, then we computed the agreement between the crowd worker labels the labels of the prediction model. The achieved agreement score is 0.95 , which further supports our choice r/AskReddit to obtain labeled comments for training the model.

Table 3: Percentage of toxic and non-toxic comments in each subreddit.

	2016		*2017*	
Subreddit	**%Toxic**	**%Non-toxic**	**%Toxic**	**%Non-toxic**
r/funny	13.32%	86.69%	12.62%	87.38%
r/AskReddit	13.33%	86.67%	12.86%	87.14%
r/todayilearned	12.41%	87.59%	11.81%	88.19
r/science	2.84%	**97.16%**	2.74%	**97.26%**
r/worldnews	13.45%	86.55%	**13.01%**	86.99%
r/pics	11.23%	88.77%	12.27%	87.73%
r/IAmA	8.73%	91.27%	8.71%	91.29%
r/gaming	10.28%	89.72%	10.82%	89.18%
r/videos	**15.44%**	84.56%	12.87%	87.13%
r/movies	11.18%	88.82%	9.82%	90.18%

With the built prediction model, we performed the prediction on all the comments in the years 2016 and 2017. We then computed the percentage of toxic and non-toxic comments in each subreddit, as shown in Table 3. The outcomes of this experiment show that the subreddit r/videos was the most toxic in the year 2016 (and still quite high in 2017), while the subreddit r/worldnews was the most toxic in the year 2017 (and was quite high in 2016). Furthermore, the subreddit r/science exhibits the lowest percentage of toxicity across both years 2016 and 2017. Our findings indicate that toxicity is prevalent in Reddit communities; therefore, the problem of detecting toxicity triggers is important in online communities. Furthermore, our results show that subreddits vary in their toxic nature.

5 TOXICITY TRIGGER DETECTION

5.1 Feature Modeling for Trigger Prediction

Changes in the topic of a discussion, the context of the discussion, or the overall emotion could tell us something about the comments that might trigger toxicity [40]. This insight motivated us to introduce topical and emotional shifts along with the discussion context as features to predict toxicity triggers.

The features were extracted from discussion threads that we show in stage (d) from Figure 1 to capture the occurrence of changes that might lead to toxicity triggers. Hence, in the following, we explain the proposed features that include topical shift, sentiment shift, and the discussion context, as in stage (e) in Figure 1.

5.1.1 Topical Shift. By using pre-trained word embeddings to represent the comments in the discussion, we can measure the topical similarity of the comments using the cosine similarity measure [26], as in Equation (1). This measure captures the topical similarity between the non-toxic parent comment and all preceding comments that came before it within the discussion thread. For any given parent comment p and a preceding comment c that came before p within a discussion, we calculate the similarity between p and c as:

$$cos(\vec{v_p}, \vec{v_c}) = \frac{\vec{v_p} \cdot \vec{v_c}}{|\vec{v_p}| \cdot |\vec{v_c}|} = \frac{\sum_{t=1}^{n} v_p(t) v_c(t)}{\sqrt{\sum_{t=1}^{n} v_p(t)^2} \sqrt{\sum_{t=1}^{n} v_c(t)^2}} \quad (1)$$

where v_p and v_c correspond to the vector space representations of both p and c, while n corresponds to the total number of terms.

To build the word embedding representations in comments, we used the GloVe pre-trained model on 5 billion tokens to generate vectors of 100 dimensions for each comment. Next, we computed the cosine similarity between the parent comment and all preceding comments. Then, we used a threshold to determine if the comments were on-topic or off-topic. To compute this threshold, we relied on the k-means clustering algorithm [17] to construct two clusters that denote on-topic and off-topic comments. Then, we considered the smallest cluster to be an indicator of comments that exhibit topical shift [40]. As a threshold, we used the centroid, which is the arithmetic mean of all the data points that belong to a cluster. So, if the cosine similarity of a comment <= the threshold, then a topic shift had occurred in the discussion thread.

5.1.2 Sentiment Shift. To study the sentiment of each comment within the discussion thread [49], we used AFINN's lexicon [16] to score each comment based on the sentiment's polarity. The lexicon contains 2,477 words and phrases labeled for their valence in the range [-5,5], where minus represents negative polarity, and the plus represents positive polarity. To compute the lexicon score, we mapped terms in comments with words or phrases in the lexicon. Then, we aggregated the score of all the terms in the comment to end up with a single score that represents the sentiment of each comment. The final score of each comment is used to cast judgment on the overall sentiment present in the comment. To capture the sentiment shift, we followed the same approach that we used in the topic similarity shift analysis, where we clustered the obtained scores into two classes that correspond to comments with sentiment change and comments without sentiment change. The smallest centroid value was also used as a threshold to determine if comments had any sentiment shift [40].

5.1.3 Discussion Context. When users initiate online discussions, the topics and events discussed earlier usually dictate the flow of the upcoming responses in the discussion thread [10]. This premise motivated us to introduce the discussion context as a feature to detect toxicity triggers. To compute the discussion context, we collected all the comments that came before the toxicity trigger, granted that these comments fall between the post or comment that the trigger replies to and the trigger itself (i.e., creating a sense of a context). Then, we generated the word embedding representation of these comments, as we mentioned previously.

For each toxicity trigger, we averaged the word embeddings of all the comments that came before it to create the discussion context. The averaging of word embeddings was necessary to generate a single context for each toxicity trigger in the collection.

5.2 Training a Statistical Model

For the prediction experiment that we show in stage (f) of Figure 1, we used the LSTM-based neural network. As for the training data, we randomly sampled 151,780 non-toxic parent comments (50% for training and 50% for testing). We checked the toxicity of their child comments and marked whether they are τ-toxicity triggers (i.e., it belonged to the class Trigger) or not. In our analysis, we ensured that the training data from the class Trigger and Non-trigger were evenly split (balanced). Then, we evaluated the performance of the prediction model and used it to predict the occurrence of toxicity triggers in the remaining 414,316 non-toxic parent comments, as in stage (g) of Figure 1. We formally define a τ-toxicity trigger as non-toxic comment having equal or more than τ toxic child comments. Intuitively, higher τ leads to clear but lower coverage of toxicity triggers. To determine the appropriate threshold for further analysis, we tested varying values of τ, the threshold of the number of toxic child comments, from one toxic child (τ=1) to five toxic children (τ=5). We experimented with the value of τ by running small prediction experiments at each threshold and manually evaluated the prediction performance. Starting with τ=1, when we set the training set size to 57,760, the prediction accuracy on the 6,418 test set was 0.63. Then, we increased the value of τ to 2, where the training set size was 6,476, and the prediction accuracy on 720 comments was 0.46. Increasing τ to 3 reduced the training set size to 1,288, and the prediction accuracy on 144 comments also dropped to 0.40. As for τ=4, the training set size became 394, and the prediction accuracy of 44 comments dropped to 0.10. Lastly, by setting τ to 5, the training set size became 182, and the prediction accuracy on 20 comments was 0.10. Our findings show that having two or more toxic children is enough to detect toxicity triggers as long as the training set size is large.

5.3 Performance Evaluation

The results of the classification experiment are presented in Table 4. As a baseline model, we trained the LSTM model using GloVe word embeddings only. Then, we added sentiment shift features to the model by concatenating the shift scores with the embedding feature vector. The results of including topical shift features did not show a significant improvement over the baseline. Similarly, we added topical shift features to the word embedding feature vector. The improvement over the baseline was slightly better than sentiment shift features. Then, we added the discussion context feature, which provided another slight improvement over the topical shift features. Lastly, we combined topical shift, sentiment shift, and discussion context features with GloVe embeddings. The achieved average accuracy of the model was 78.2%, which shows a 6% improvement over the baseline model. This result indicates that topical and sentiment shift features, along with the discussion context, improve the detection of toxicity triggers. As for the predictive model's agreement score, we manually labeled 100 comments from our collection and computed the agreement between the labels and the prediction of the model. The agreement that we got was 84%, which is good enough for us to consider the model for performing toxicity trigger detection.

6 ERROR ANALYSIS

Lastly, we tie-in our findings from the prediction experiments by examining some of the online discussions in which toxicity triggers might have occurred. The discussion depicted in Figure 2 shows an example of a correctly and incorrectly classified toxicity trigger.

Table 4: Performance of the neural network models in terms of accuracy, macro F_1, and ROC-AUC score. BL:Baseline, sent:sentiment, cont:context. (The results are obtained by taking the average of 5 random runs.)

Features	ROC-AUC	Accuracy	Macro F_1
GloVe(BL)	0.81	72.8%	0.73
GloVe+sent.	0.81	72.9%	0.72
GloVe+topic	0.82	73.8%	0.73
GloVe+cont.	0.83	74%	0.78
All features	**0.87**	**78.2%**	**0.78**

The correctly identified toxicity trigger in Figure 2 shows some form of disagreement in opinion, which leads to toxicity.

The topic of the discussion (i.e., the submission post) aims at provoking reactions from readers. As the context of the parent-comment was in agreement with the submission post, the first child-comment was aggravated by the parent-comment, which led it to be toxic. It was triggered by the parent comment, which itself was not toxic.

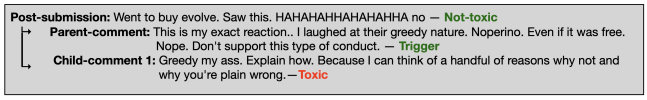

Figure 2: An example discussion with trigger parent comment that was correctly predicted as "trigger" by the model.

As for incorrectly-classified toxicity triggers, Figure 3 shows an example of a non-toxicity trigger classified as a trigger by the prediction model. The submission post talks about a racial issue that happened in Britain. While the parent comment made an assuring claim without any proof, the subsequent child comments did not respond to the claim in an aggressive manner then proceeded to provide evidence to their counter-argument. Cases where child comments behave in an unaggressive manner (like producing high-quality arguments [9]) can lead to incorrect classification.

On the other hand, one might argue that the parent comment triggers toxicity if a subsequent child comment was toxic. In this case, the classification mistake was due to the lack of information about the end of the discussion thread [20]. With this example, we demonstrate some of the challenges associated with the problem of detecting toxicity triggers in discussion threads, which mainly stem from unpredictable child comment toxicity and lack of knowledge about the discussion termination.

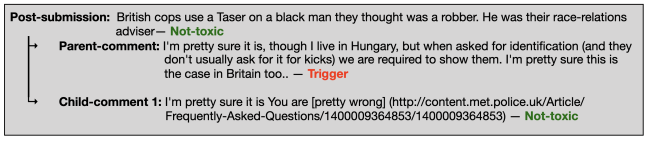

Figure 3: An example discussion with not-trigger parent comment that was incorrectly classified as "trigger" by the model.

Also, the outcomes of the examination showed that when the model correctly identifies toxicity triggers, it usually relies on specific trigger-terms that contain provocative words like *argument* and shift ques from the conversation thread. However, when the model incorrectly classifies a toxicity trigger, this is usually attributed to: a) the lack of features that detect the end of the discussion thread, and b) incorrect toxic-comment classification results, making it difficult to detect if the parent comment triggers toxicity or not.

Other incorrect classification examples shed light on some of the challenges associated with toxicity trigger detection and open areas for future work. Such as incorporating additional features into the toxicity trigger detection model that detect sarcasm or trolling [41], discussion end-points [20], and introducing semantic shift features [40] that track the stylistic writing style of comments.

7 CONCLUSION

By utilizing a large-scale dataset and multiple techniques to detect toxic comments and toxicity triggers in online discussions, this research contributes to the stream of ongoing literature on online toxicity. Our approach shows novelty by being, to our knowledge, the first study to examine online toxicity triggers by using an extensive collection of online discussions and incorporating two shift features and a discussion context feature to detect toxicity triggers.

In this study, we detect toxicity in online communities and build discussion threads to use along with toxicity predictions to detect toxicity triggers. In the context of a discussion thread, we defined toxicity trigger as a non-toxic parent comment that has at least one toxic comment as its child. Based on this definition, we built an LSTM neural network model that combines topical and sentiment shift features along with discussion context features to detect toxicity triggers.

The findings from this research pave the way toward many interesting directions. For example, we show preliminary evidence that some communities are more robust against toxicity than others - it then becomes a crucial question to isolate the "resistance tactics" that more healthy communities employ to tackle the toxicity problem. Moreover, with our large-scale dataset, we can perform a multitude of analyses, including temporal studies on the emergence of toxicity within online communities and thus create research that can impact the health of those communities positively by identifying toxicity triggers.

Regarding the practical implications of our work, we note that by studying toxicity triggers, social media platforms, community moderators, and decision-makers can deploy automatic or semi-automatic techniques to prevent toxicity from spreading in online discussions. Such attempts could shield users from the adverse effects of toxicity and encourage them to participate in meaningful discussions without disseminating toxic content in online communities.

REFERENCES

[1] [n.d.]. *The Essential High-Quality Data Annotation Platform.* Retrieved October 10, 2019 from https://www.figure-eight.com/
[2] [n.d.]. *Perspective.* Retrieved October 10, 2019 from http://www.perspectiveapi.com/
[3] [n.d.]. *Reddit Statistics.* Retrieved September 30, 2017 from https://files.pushshift.io/reddit/comments/

[4] [n.d.]. *Tracking the top 5000 subreddits.* Retrieved August 21, 2017 from http://redditlist.com

[5] 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)*. ACM, New York, NY, USA, 291–292. https://doi.org/10.1145/3342220.3344933

[6] Kathryn Zickuhr Myeshia Price-Feeney (CiPHR) Amanda Lenhart, Michele Ybarra (CiPHR). [n.d.]. *Online Harassment, Digital Abuse, and Cyberstalking in America.* Retrieved October 10, 2019 from https://datasociety.net/output/online-harassment-digital-abuse-cyberstalking/

[7] 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. https://doi.org/10.1145/3041021.3054223

[8] Gustavo Batista, Ronaldo Prati, and Maria-Carolina Monard. 2004. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explorations* 6, 1 (2004), 20–29. https://doi.org/10.1145/1007730.1007735

[9] George Berry and Sean J. Taylor. 2017. Discussion Quality Diffuses in the Digital Public Square. 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, 1371–1380. https://doi.org/10.1145/3038912.3052666

[10] Tiago Cunha, David Jurgens, Chenhao Tan, and Daniel Romero. 2019. Are All Successful Communities Alike? Characterizing and Predicting the Success of Online Communities. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3308558.3313689

[11] Laura Del Bosque and Sara Garza. 2016. Prediction of Aggressive Comments in Social Media: an Exploratory Study. *IEEE Latin America Transactions* 14, 7 (2016), 3474–3480. https://doi.org/10.1109/TLA.2016.7587657

[12] Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2016. Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Scientific reports* 6 (2016), 37825. https://doi.org/10.1038/srep37825

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] 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. https://doi.org/10.1145/3200947.3208069

[15] Kilem Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* Advanced Analytics, LLC.

[16] 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.

[17] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108. http://www.jstor.org/stable/2346830

[18] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. In *Social Informatics*, Tie-Yan Liu, Christie Napa Scollon, and Wenwu Zhu (Eds.). Springer International Publishing, Cham, 49–66.

[19] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. 2009. Patterns of query reformulation during Web searching. *Journal of the American Society for Information Science and Technology* 60, 7 (2009), 1358–1371. https://doi.org/10.1002/asi.21071

[20] Yunhao Jiao, Cheng Li, Fei Wu, and Qiaozhu Mei. 2018. Find the Conversation Killers: A Predictive Study of Thread-Ending Posts. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1145–1154. https://doi.org/10.1145/3178876.3186013

[21] Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2016. Freshman or Fresher? Quantifying the Geographic Variation of Language in Online Social Media. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13121

[22] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 933–943. https://doi.org/10.1145/3178876.3186141

[23] 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* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3739–3748. https://doi.org/10.1145/2702123.2702529

[24] Kyounghee Kwon and Anatoliy Gruzd. 2017. Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research* 27, 4 (2017), 991–1010. https://doi.org/10.1108/IntR-02-2017-0072

[25] Candice Lanius. 2019. Torment Porn or Feminist Witch Hunt: Apprehensions About the #MeToo Movement on /r/AskReddit. *Journal of Communication Inquiry* 43, 4 (2019), 415–436. https://doi.org/10.1177/0196859919865250

[26] Ray R Larson. 2010. Introduction to information retrieval. *Journal of the American Society for Information Science and Technology* 61, 4 (2010), 852–853.

[27] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning - Volume 32* (Beijing, China) *(ICML'14)*. JMLR.org, II–1188–II–1196. https://dl.acm.org/doi/10.5555/3044805.3045025

[28] Adrienne Massanari. 2017. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19, 3 (2017), 329–346. https://doi.org/10.1177/1461444815608807

[29] Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. 2017. The Impact of Toxic Language on the Health of Reddit Communities. In *Advances in Artificial Intelligence*, Malek Mouhoub and Philippe Langlais (Eds.). Springer International Publishing, Cham, 51–56.

[30] Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management* 57, 3 (2020), 102087. https://doi.org/10.1016/j.ipm.2019.102087

[31] 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. https://doi.org/10.1145/2872427.2883062

[32] Robert Orton, Rita Marcella, and Graeme Baxter. 2000. An observational study of the information seeking behaviour of Members of Parliament in the United Kingdom. *Aslib Proceedings* 52, 6 (2000), 207–217. https://doi.org/10.1108/EUM0000000007015

[33] Mourad Oussalah, F. Faroughian, and Panos Kostakos. 2018. On Detecting Online Radicalization Using Natural Language Processing. In *Intelligent Data Engineering and Automated Learning – IDEAL 2018*, Hujun Yin, David Camacho, Paulo Novais, and Antonio J. Tallón-Ballesteros (Eds.). Springer International Publishing, Cham, 21–27.

[34] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[35] Gabriele Pergola, Lin Gui, and Yulan He. 2019. TDAM: A topic-dependent attention model for sentiment analysis. *Information Processing & Management* 56, 6 (2019), 102084. https://doi.org/10.1016/j.ipm.2019.102084

[36] Madhu C. Reddy and Bernard J. Jansen. 2008. A Model for Understanding Collaborative Information Behavior in Context: A Study of Two Healthcare Teams. *Information Processing & Management* 44, 1 (2008), 256–273. https://doi.org/10.1016/j.ipm.2006.12.010

[37] Felix Richter. 2017. *Infographic: The Explosive Growth of Reddit's Community.* Retrieved May 27, 2017 from https://www.statista.com/chart/11882/number-of-subreddits-on-reddit/

[38] Joni Salminen, Hind Almerekhi, Milica Milenkovic, Soon-Gyo Jung, Jisun An, Haewoon Kwak, and Jim Jansen. 2018. Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17885

[39] Sara Sood, Judd Antin, and Elizabeth Churchill. 2012. Profanity Use in Online Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. ACM, New York, NY, USA, 1481–1490. https://doi.org/10.1145/2207676.2208610

[40] Kamil Topal, Mehmet Koyuturk, and Gultekin Ozsoyoglu. 2016. Emotion -and Area- Driven Topic Shift Analysis in Social Media Discussions. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Davis, California) *(ASONAM '16)*. IEEE Press, 510–518.

[41] Paraskevas Tsantarliotis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2017. Defining and predicting troll vulnerability in online social media. *Social Network Analysis and Mining* 7, 1 (Jun. 2017). https://doi.org/10.1007/s13278-017-0445-2

[42] Ellen D. Wagner. 1994. In Support of a Functional Definition of Interaction. *American Journal of Distance Education* 8, 2 (1994), 6–29. https://doi.org/10.1080/08923649409526852

[43] Daisuke Wakabayashi. 2017. Google Cousin Develops Technology to Flag Toxic Online Comments. *The New York Times* (2017). https://www.nytimes.com/2017/

02/23/technology/google-jigsaw-monitor-toxic-online-comments.html

[44] William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media* (Montreal, Canada) *(LSM '12)*. Association for Computational Linguistics, USA, 19–26.

[45] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. *IEEE Access* 6 (Feb. 2018), 13825–13835. https://doi.org/10.1109/ACCESS.2018.2806394

[46] Tim Weninger, Xihao Avi Zhu, and Jiawei Han. 2013. An Exploration of Discussion Threads in Social News Sites: A Case Study of the Reddit Community. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Niagara, Ontario, Canada) *(ASONAM '13)*. Association for Computing Machinery, New York, NY, USA, 579–583. https://doi.org/10.1145/2492517.2492646

[47] Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for Counterspeech on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada, 57–62. https://doi.org/10.18653/v1/W17-3009

[48] 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. https://doi.org/10.1145/3038912.3052591

[49] Frank Z. Xing, Filippo Pallucchini, and Erik Cambria. 2019. Cognitive-inspired domain adaptation of sentiment lexicons. *Information Processing & Management* 56, 3 (2019), 554 – 564. https://doi.org/10.1016/j.ipm.2018.11.002

[50] Rui Zhao, Anna Zhou, and Kezhi Mao. 2016. Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features. In *Proceedings of the 17th International Conference on Distributed Computing and Networking* (Singapore, Singapore) *(ICDCN '16)*. ACM, New York, NY, USA, Article 43, 6 pages. https://doi.org/10.1145/2833312.2849567