Four Types of Toxic People: Characterizing Online Users' Toxicity over Time

Raghvendra Mall
rmall@hbku.edu.qa
Qatar Computing Research Institute, Hamad Bin Khalifa University

Mridul Nagpal
mridulnagpal07@gmail.com
International Institute of Information Technology Hyderabad

Joni Salminen
jsalminen@hbku.edu.qa
Qatar Computing Research Institute, Hamad Bin Khalifa University
Turku School of Economics at the University of Turku

Hind Almerekhi
hialmerekhi@hbku.edu.qa
Hamad Bin Khalifa University

Soon-gyo Jung
sjung@hbku.edu.qa
Qatar Computing Research Institute, Hamad Bin Khalifa University

Bernard J. Jansen
bjansen@hbku.edu.qa
Qatar Computing Research Institute, Hamad Bin Khalifa University

ABSTRACT

Identifying types of online users' toxic behavior reveals important insights from social media interactions, including whether a user becomes "radicalized" (more toxic) or "pacified" (less toxic) over time. In this research, we design two metrics to identify toxic user types: F score that captures the changes in a user's toxicity, and G score that captures the direction of the shift taking place in the user's toxicity pattern. We apply these metrics to a dataset of 4M user comments from Reddit by defining four toxic user types based on the toxicity scores of a user's comments: (a) Steady Users whose toxicity scores are steady over time, (b) Fickle-Minded Users that switch between toxic and non-toxic commenting, (c) Pacified Users whose commenting becomes less toxic in time, and (d) Radicalized Users that become gradually toxic. Findings from the Reddit dataset indicate that fickle-minded users form the largest group (31.2%), followed by pacified (25.8%), radicalized (25.4%), and steadily toxic users (17.6%). The results suggest that the most typical behavior type of toxicity is switching between toxic and non-toxic commenting. This research has implications for preserving the user-friendliness of online communities by identifying continuously toxic users and users in danger of becoming radicalized (in terms of their toxic behavior), and designing interventions to mitigate these behavior types. Using the metrics we have defined, identifying these user types becomes possible. More research is needed to understand why these patterns take place and how they could be mitigated.

CCS CONCEPTS

• **Human-centered computing** → **Social network analysis**; *Empirical studies in collaborative and social computing*; • **Social and professional topics** → *User characteristics*;

KEYWORDS

online toxicity, user analysis, social media behavior, Reddit

ACM Reference Format:
Raghvendra Mall, Mridul Nagpal, Joni Salminen, Hind Almerekhi, Soon-gyo Jung, and Bernard J. Jansen. 2020. Four Types of Toxic People: Characterizing Online Users' Toxicity over Time. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society (NordiCHI '20), October 25–29, 2020, Tallinn, Estonia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3419249.3420142

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
NordiCHI '20, October 25–29, 2020, Tallinn, Estonia
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-7579-5/20/10…$15.00
https://doi.org/10.1145/3419249.3420142

INTRODUCTION

Research on online toxicity provides a crucial inquiry into online users' experiences, as hate, profanity, and toxicity are prevalent in many social networks, communities, and websites [14–16, 18, 26, 33]. Because of its high prevalence and multiple forms, content providers, moderators, and the society at large are struggling to deal with online toxicity [24]. Toxicity has a detrimental effect on the health of online communities [11], degrading the user experience in online social networks, especially for vulnerable groups such as teens facing cyberbullying [10], women and minorities [9], and other targets of online hate [26]. Online toxicity can be defined as actions of a user or users that create a negative atmosphere for the other users of social media [13].

To mitigate the many negative effects of online toxicity, researchers are striving to better understand its antecedents. However, often the work is focused on development of machine learning classifiers [28], instead of understanding the users' shifts in toxicity [3]. In addition, most prior work on analyzing online toxicity focuses at either the community level [13, 19, 20] or comment level [24, 26], rather than analyzing toxicity *at a user level*. Yet, toxicity originates from individual users and, therefore, research focusing on individual users should be conducted to better understand the emergence of online hate. Finally, there is a general lack of approaches for longitudinal analysis of users' toxicity. In contrast, previous research tends to assume that users' toxicity would be constant or uniform over time, so it would be possible to identify (permanently) "toxic users" and "non-toxic users" [38]. Although very few studies have looked into the patterns and variation of users' toxicity over time, it has been shown that users' online reactions vary over

time [31], for example, such that radicalization takes place over time.

Therefore, a longitudinal investigation of toxicity patterns has the potential to reveal meaningful insights on the toxicity of individuals to a greater degree than mere labeling of users to toxic or non-toxic. In addition, doing so can provide means of early detection of a user's comments transforming from neutral to toxic. Following the above rationale, the goal of this research is to *analyze the change of toxicity of online users over time*. For this purpose, we obtain a dataset from Reddit, a social network that has an active user base of 1.6 billion monthly visitors[1] and a known problem of toxicity. As several researchers have found toxicity to be a major problem on Reddit [5, 13, 19, 20, 31], a longitudinal dataset collected from various subreddits is likely to provide good grounds for validation of our approach.

In the following section, we review the related literature. After this, we explain the methodology for data collection, toxicity scoring, and pattern detection. We then present our findings and discuss their implications for typologies of toxicity among online users.

LITERATURE REVIEW

Approach

To identify related research, we search for online toxicity research using relevant keywords (e.g., 'online toxicity') on two academic databases – ACM Digital Library and Google Scholar – as done in previous HCI literature reviews [27]. In these searches, we regard 'toxicity', 'profanity', and 'hatefulness' as synonyms, as done in previous studies [26]. From the search results, we select the articles that, based on their titles or abstract, focus on (a) *variation of hate over time*, or (b) *individual users* (rather than communities or groups). As we find little research on the toxicity of users over time, we expand the literature search by considering toxicity as one form of *sentiment*, which opens a connection to thematically related studies in sentiment analysis (SA). These studies suggest various ways of investigating sentiment patterns of online users over time and are thus relevant for our research focus. General literature reviews on online toxicity are given in [6, 37].

Impact of Community Norms

The previous research on online hate and toxicity has shown that contextual and user-specific conditions affect online toxicity. For example, the interpretation of a comment's toxicity may depend on the community norms and its linguistic patterns [24], so that the same language is perceived toxic in one community and non-toxic in another community. There is also topical variation, e.g. political topics being more prone to hate [14, 26]. Moreover, the interpretation of toxicity seems to vary among users, and people from different countries have different perceptions of toxicity of the same online comments [30].

Sood et al [33] confirm the importance of community context in detecting online profanity. They note that the definition of toxicity and tolerance for profanity varies across different communities. Additionally, users may circumvent profanities by purposefully misspelling or altering profane words. Traditional list-based approaches to profanity detection do not take these important factors into account. A similar conclusion about the importance of context is drawn by Aisopos et al. [1] in their comparative analysis of content and context-based approaches to sentiment analysis. They argue for two innovations in sentiment analysis: a focus on character instead of word n-gram graphs, and the increased efficacy of context-based approaches. Because sentiment polarity is reliant on the social contexts of tweets (i.e. more positive tweets engenders further positive tweets and vice versa), Aisopos et al. [1] propose a metric – the Polarity Ratio – for determining the aggregate polarity of a collection of comments thus extending the reach of traditional SA beyond the user-level. The Polarity Ratio may be applied to all contextual aspects of a comment, including all messages by the same author, all messages relating to a specific topic and all comments by followers [1].

Shifts in Online Users' Sentiments

The evolution of Internet discussions' sentiment over time has been explored by several scholars. In Reddit, Singer et al. [31] provide evidence that user performance and comment quality deteriorate throughout users' sessions. Giachanou et al. [7] suggest that one approach to studying sentiment evolution is by identifying sentiment spikes in tweets of a certain entity. By using Latent Dirichlet allocation, they extract topics contributing to the sentiment spike and rank them using Kullback-Leibler divergence [12]. The ranking of sentiment strength emerges is another approach for tracking the long-term evolution of user comments. Giachanou et al. [8] employ human Mechanical Turk workers to rank sentiment strength in a sample of tweets.

The evolution of sentiment within individual online discussions threads in relation to topic shifts is addressed by Topal et al. [36]. Their findings indicate the following: (1) the root comment appears to set the tone for the comments to follow; (2) topic shifts occur in 67-72% of all comments; (3) topic shifts increase when users use highly emotional language on any topic; (4) the first-shifted comment contain low levels of emotion; (5) after an initial first-shifted comment, all subsequent comments in different trees display an increase in topic shift; (6) the emotion of comments can predict with 75% accuracy whether a topic shift has occurred; (7) accuracy can be increased by 1-2% if tree structure is taken into account; and (8) by another 1-2% if the number of words in the comment is known.

In their studies on sentiment spike, Giachanou et al. [7, 8] employ SentiStrength (see [34, 35] to categorize tweets as positive, negative or neutral. Bamman & Smith [4, p. 575] distinguish between "tweet whole sentiment" and "tweet word sentiment". The former is determined by the Stanford Sentiment Analyzer [32] and the latter by modeling the minimum and maximum word sentiment and the distance between the two in a single comment. As Bamman & Smith [4] show, however, an analysis of comment-level features is inadequate for determining ambiguous sentiments, like sarcasm, that are embedded in cultural-specific social pragmatics, i.e., contexts. Aisopos et al. [1, p. 196] refer to the analysis of such sentiments as "fine-grained sentiment analysis". In

[1] https://www.statista.com/statistics/443332/reddit-monthly-visitors

addition to sarcasm, this rubric also includes sentiments such as tension, depression, anger and fatigue. Statistically significant improvements in sarcasm detection are obtained when context and author features are taken into account [4]. Therefore, the inference may be drawn that similar improvements will occur when context-level features are considered in the study of other fine-grained sentiments.

Park et al. [23] aim to replicate human perceptions of the quality of online comments by developing a model that defines quality through a combination of comment and user history criteria. Comment criteria include conversational relevance, length, personal experience, readability, and recommendation, while user history criteria include user comment rate, user comment length, user personal experience, user picks, user readability and user recommendations [23]. The evolution of sentiment across different online communities on Reddit is tracked by Kumar et al. [13]. This large-scale study, involving 40 months of Reddit data, including 1.8 billion comments made by over 100 million users across 36,000 communities, effectively combines different strands of SA into a single methodology to study inter-community conflict and communication. The researchers define inter-community "mobilizations" as "cases where a cross-link leads to an increase in the number of comments by current source members on the discussion thread of the target post" [13, p. 935]. As the researchers point out, such mobilizations are a major source of online conflict. They find that 1% of communities are responsible for 74% of negative inter-community mobilizations and, therefore, that online anti-social behavior may be reduced by banning these small groups of miscreants.

RESEARCH GAP

In previous research, time has been considered as an important aspect for tracking social media sentiments. However, previous research rarely considers defining user types. Defining types of toxic users can help many stakeholder groups: moderators to identify and understand particular types of toxic behavior, policy makers and researchers to detect those online users at risk to become increasingly toxic, and even the online users themselves by providing information on their own toxicity patterns over time. Thus, the research at hand has potential for positive impact towards the goal of healthy and productive online encounters between users.

Moreover, previous studies on online toxicity tend to focus on targets, types, and language of toxic behavior, not users *per se*. While some authors have incorporated user-level *features* in their models [23], users have not been the unit of analysis. Even more importantly, prior research has not focused on analyzing users' temporal toxicity trends. Some studies use features that may not be available when querying anonymous users, especially given the privacy concerns of accessing social media profiles in environments where the users are anonymous. In these cases, the toxicity patterns of the users may need to be constructed only using the historical toxicity of their comments.

Overall, there is a gap of studying the toxicity patterns of individual users over time. We address this gap by devising an algorithmic approach that utilizes sound theoretical concepts and empirical distributions to analyze social media users' toxicity over time to identify specific types of toxic users.

METHODOLOGY

Data Collection

To investigate the toxicity of users over time, we opt for collecting data from Reddit. For data collection, we focus on the 10 largest subreddits with the highest number of subscribers [2]. For each subreddit, we retrieve the 10 discussions with the highest number of comments starting from 2009 until August of 2017. So, for each subreddit, we had a total of 90 discussions per year, and a total of 900 submissions over the period of 9 years. We focused on submissions with the highest number of comments because the comment number usually implies the existence of more discussions. On average, a discussion has 4,202 comments.

Using Python's Reddit API Wrapper (PRAW), we collected all the comments associated with each discussion and ensured that the comment objects retrieved from the API included (a) the time stamp, (b) ID of the comment, (c) Username of the person that posted the comment, and (d) the comment text. After collecting the comments, we removed the long comments exceeding 3,000 characters since these pose issues for the toxicity scoring method that we used. The potential impact of removing these comments in the results is likely very small. We ended up with a large text collection; i.e., a corpus of comments and discussions from the Top 10 subreddits on Reddit, as shown by statistics about the total number of comments and discussions in Table 1.

Toxicity Scoring

Alphabet has an initiative called Perspective API, aimed at developing computational methods for providing safer environments for social media discussions. Perspective API has been trained on millions of online comments that have been manually labeled for toxicity [38]. Following other published research in HCI and related fields [17, 21, 29], we utilize the Perspective API to score the comments collected for this research. Upon obtaining access to the API from Alphabet, we tested it by inputting requests to score comments. The version of the API at the time of the study had two main types of models: (a) alpha models and (b) experimental models. In this research, we use the alpha category's toxicity model that returns a toxicity score between 0 and 1, where 1 indicates maximum toxicity. According to the API documentation, the returned scores are toxicity probability, i.e., how likely a comment is perceived to be toxic.

To retrieve the toxicity scores, we sent 4,028,324 million comments to the Perspective API. Overall, we were able to successfully score 3,727,889 comments, representing 92.54% of the comments in the dataset. According to the Perspective API documentation, failure to provide scores can be due to non-English content, and lengthy comments. To further establish the validity of the toxicity scores given by the Perspective API, we performed a manual rating of a random sample of 150 comments. An independent human rater determined if a comment is toxic or not toxic, and we compared this rating to the score given by Perspective API.

We used the threshold of 0.50, so that comments scored below that threshold by Perspective API are considered non-toxic and comments above 0.50 are considered toxic, the human rater also

[2] http://redditlist.com

classifying on this binary range. We obtained a simple percentage agreement of 86.7%. We also computed Cohen's Kappa (k) that considers the the probability of agreement by chance in the ratings. Here, we observed 135 agreements (90.0% of the observations), whereas the number of agreements expected by chance would have been 118.5 (79.0% of the observations). The Kappa metric was k = 0.524, indicating a 'moderate' strength of agreement. While the agreement score would ideally be higher, we consider it acceptable for the purposes of this study, especially given that there is evidence of toxicity ratings being subjective and thus hard to agree upon [25, 30].

Table 1: Summary of the analyzed Reddit collection

Subreddit	No of Discussions	No of Comments	No of Unique Users
Funny	90	206, 444	86, 374
AskReddit	90	1, 264, 723	357, 732
Today I Learned	90	238, 623	82, 761
Science	90	122, 841	44, 550
World News	90	365, 057	106, 191
Pics	90	337, 624	118, 918
IAmA	90	371, 616	154, 899
Gaming	90	219, 202	86, 378
Videos	90	241, 447	85, 840
Movies	90	360, 312	117, 718
Total	900	3, 727, 889	1, 241, 361

Exploratory Analysis

We perform an initial exploratory analysis on the dataset. We identified a total of 741, 994 unique users contributing in total 3, 727, 889 comments in 900 long discussions spanning the 10 subreddits (see Table 1). Here each user made at least one comment in at least one discussion out of the 90 discussions per subreddit. We then analyzed the user comment history over time i.e. observe how many users are highly active on Reddit and are making several comments over these 10 subreddits versus those users who are infrequent and sporadically comment only once or twice. The majority of the unique users comment only once or twice in all the subreddits. The maximum number of comments from one user in all discussions is 3, 413 while the average number of comments by a unique user in the entire dataset is 4.41.

We further analyzed the user activity (comment frequency) for each of the 10 subreddits as depicted in Figure 2a. Figure 2a indicates majority of the users make one or two comments and are thus infrequent users in the discussions taking place in each subreddit. We additionally highlight the average toxicity of such users (who make one or two comments) per subreddit in Figure 2b and the average toxicity of all unique users who make more than 2 comments per subreddit (see Figure 2c). Since we are interested in understanding the users' online toxicity behavior over time in long discussions, we analyze only those users who make at least 15 comments in the entire dataset. We choose the threshold to be 15 as there are a significant number of unique users (43, 031) who have commented over these 900 long discussions and enable us to capture the inherent temporal toxicity trends of such users in comparison to users with fewer comments. Moreover, the number of unique users with over 30 comments in the dataset reduces drastically (14, 520), thereby, preventing us from capturing certain user types estimated based on the toxicity scores of their comments over time. This is further validated in Figure 1.

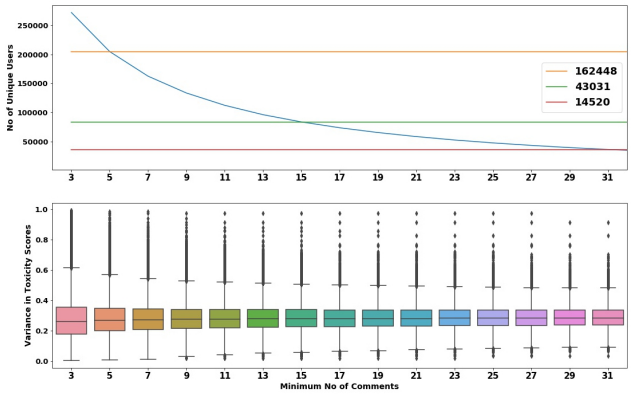


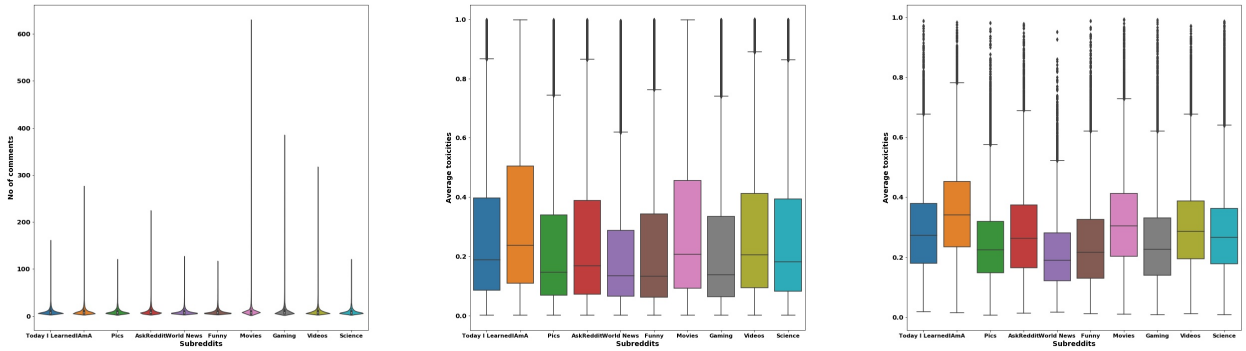
Figure 1: The first subplot showcases how the number of unique users decreases monotonically as the minimum number of comments they make in our dataset increases. We show via horizontal lines that there are 162448, 43031 **and** 14520 **unique users who made at least** 5, 15 **and** 30 **comments respectively. The second subplot showcases that how the variance in toxicity scores of these unique users reduces and stabilizes (around** 15**) as the minimum number of comments they make in our dataset increases. We observe that the variability in the toxicity score of these users plateaus near a minimum of** 15 **comments per user and we have a significant number of unique users with at least** 15 **comments in the entire dataset to apprehend the temporal toxicity trends of different types of user behaviors in Reddit.**

Metrics to Capture Temporal Toxicity Patterns

We make certain notational conventions, such as the i^{th} user in the dataset is represented by u_i, the time when the k^{th} comment was made by the user is referred as t_k, and the toxicity score of u_i at t_k is defined as $s(u_i, t_k)$. We define the metrics utilized to quantify the different temporal toxicity trends for users based on their comments as follows:

F ***score***: The F score captures the absolute changes in a user's toxicity scores. However, to account for the delay in the time spent between two consecutive comments and, hence, the change in the state of mind of the user, we utilize an additional decay term in the formulation of the F score. Mathematically, the F score can be written as:

$$F(u_i) = \sum_{k=1}^{N_{u_i}-1} \frac{\exp^{(-\frac{t_{k+1}-t_k}{\alpha})}(\mid s(u_i, t_{k+1}) - s(u_i, t_k) \mid)}{N_{u_i} - 1} \quad (1)$$



(a) (b) (c)

Figure 2: 2a illustrates that for each subreddit, a major density belongs to the users who comment rarely. 2b shows that there is larger variance in the average toxicity scores in the subreddits when considering the isolated user comments. However, users who make more than 2 comments per subreddit usually have a smaller variance in the average toxicity score distribution suggesting stability in their temporal toxicity behavior as depicted in 2c.

Here, N_{u_i} represents the number of comments made by user u_i, and α represents a constant time factor. From Equation 1, we can observe that F score primarily captures the absolute change in the toxicity scores between two consecutive comments. By seizing absolute changes in toxicity scores, the F score does not consider the direction of shift in the temporal toxicity pattern of the user, which is appropriate for modeling the fluctuation of a user's toxicity over time. Moreover, by incorporating the decay coefficient, it can take into consideration the user's state of mind between two comments, which might influence their toxicity, given the temporal patterns of Reddit users' commenting observed, e.g., in the study by Singer et al. [31]. For example, if a user is highly active in a discussion and is responding to his own comments in a short span of time, we expect a temporal smoothness in the toxicity patterns of his comments, whereas, if a user replies to his comments after a long period of time, there might be a drastic change in the toxicity score for the user's next comment due to a shift in his state of mind. Thus, to reduce the effect of such drastic changes in toxicity scores happening between two consecutive comments over a long period of time, we utilize the exponential decay term in Equation 1.

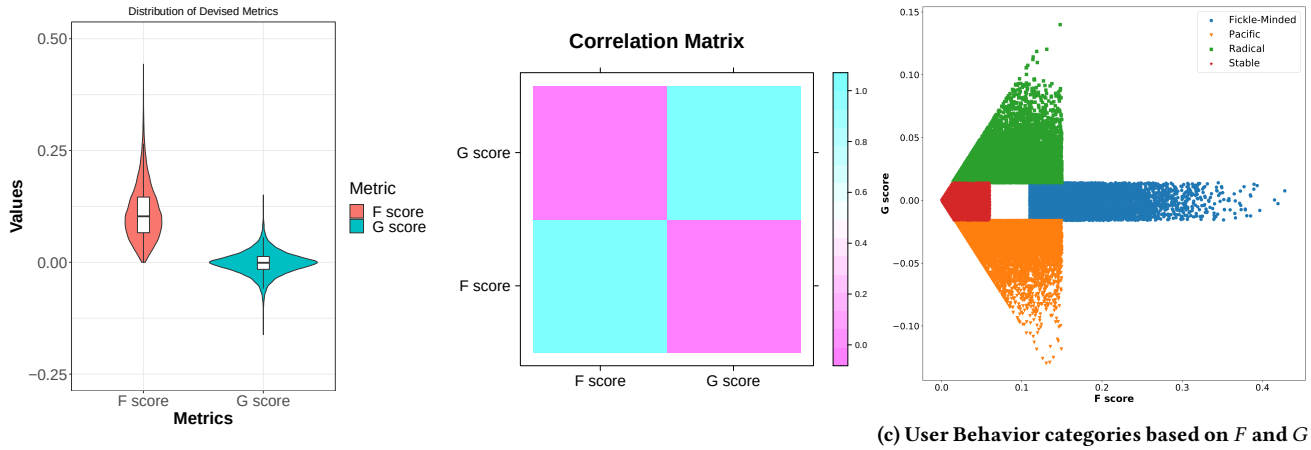
We set α as 1 hour (3,600 seconds) in all our experiments (note that α can be set to other values as well, if the commenting frequency is very sparse). Hence, for consecutive comments which are made in a short span of time say (ss), the decay factor becomes $\exp^{(-\frac{ss}{\alpha})} \approx \exp^{(-0)} = 1$, while for the comments made over a long span of time, say ll, the decay factor becomes $\exp^{(-\frac{ll}{\alpha})} \approx \exp^{(-\inf)} = 0$. This principle is inspired by the concept of random walk with restart used in the classic PageRank algorithm [22].

$F(u_i)$ can take values between $[0, 1]$, where $F(u_i)$ will be 0 only in the case when the toxicity pattern of the comments made by u_i is not changing over time. If a user u_i's comments are becoming more toxic (i.e., a user is becoming radicalized) or becoming less toxic (i.e., a user is becoming pacific), the $F(u_i)$ in both cases can be the same (since F score accounts for absolute changes in toxicity score of consecutive comments) and should usually be < 0.5. This is because the changes are over a span of $N_{u_i} \geq 15$ comments which is used in the denominator of the F score to capture average changes between toxicity scores of two consecutive comments. F score will be maximum when there are drastic changes in a short span of time between toxicity pattern of a user's comments. For example, $s(u_i, t_k) = 0$ and $s(u_i, t_{k+1}) = 1$ and this pattern appears for each pair of consecutive comments in very short span of time. In this case, $F(u_i) \approx 1$, indicating that the user u_i is extremely fickle-minded, i.e., user u_i has severe fluctuations in their toxicity scores between several consecutive comments.

G score: The G score is defined in the same way as the F score, with the only difference that it can capture the direction of the shift taking place in the temporal toxicity pattern of user u_i. Mathematically, the G score can be defined as:

$$G(u_i) = \sum_{k=1}^{N_{u_i}-1} \frac{\exp^{(-\frac{t_{k+1}-t_k}{\alpha})}(s(u_i, t_{k+1}) - s(u_i, t_k))}{N_{u_i} - 1} \tag{2}$$

The only difference between the F and G score is that the G score does not use the absolute difference in toxicity scores of two consecutive comments made by user u_i. Hence, $G(u_i)$ score can take values between $(-1, 1)$, where, if $G(u_i)$ is at the left of the first quantile (a negative number i.e < 0), the toxicity score is decreasing (as we are adding up negative numbers repeatedly). Similarly, if $G(u_i)$ is a large positive number (right of the third quantile), then the toxicity score is increasing (since we are aggregating positive changes in toxicity scores of consecutive comments). Moreover, $G(u_i)$ should be close to 0 for a fickle-minded person, where rapid shifts in direction occur when assimilating their toxicity scores. However, $G(u_i)$ can also be 0 in the case when a user's toxicity pattern remains stable over time.



(a) Distribution of F and G scores

(b) Correlation between F and G scores

(c) User Behavior categories based on F and G scores

Figure 3: We show the distribution of F and G scores in Figure 3a. Figure 3b depicts the correlation between the F and G score of all users, showing that the correlation is very low between these scores. Hence, each score is capturing distinct temporal information that is further utilized to discriminate the 4 primary types of user behaviors as shown in Figure 3c.

Categorizing User Behaviors based on Estimated Toxicity Metrics

Based on the intuitions provided in constructing the F and G scores, and their distributions and correlations depicted in Figures 3a and 3b respectively, we can define a set of rules to identify different types of temporal trends from the toxicity scores, i.e., user behaviors over time. The rules for identifying the patterns are defined as follows.

Stable Users: The rule used to identify users with nearly stable toxicity score over time is defined as:

$$G_{Q_1} < G(u_i) < G_{Q_3} \text{and} \quad F(u_i) < F_{Q_1} \tag{3}$$

Here G_{Q_1} =-0.016 and G_{Q_3} = 0.013 represents the 1^{st} and 3^{rd} quantiles of the G score distribution as also illustrated in Figure 3a. Similarly, F_{Q_1} = 0.066 corresponds to the 1^{st} quantile of the F score distribution. As mentioned earlier, small value (≈ 0) of $G(u_i)$ and small value (≈ 0) of $F(u_i)$ indicates nearly stable toxicity scores over time for u_i.

Fickle-Minded Users: The rule used to identify this class of users is:

$$G_{Q_1} < G(u_i) < G_{Q_3} \text{and} \quad F(u_i) \geq F_{Q_1} + \epsilon \tag{4}$$

Here ϵ represents a small constant value used to prevent overlapping user classes. Fickle-minded users show changes in their toxicity patterns and hence require a low G score, whereas the F score has be to $\geq F_{Q_1} + \epsilon$, as it is the sum of absolute changes in toxicity patterns over time.

Pacific Users: We define the rule to identify users who are becoming less toxic or whose comment toxicities decrease over time as follows:

$$G(u_i) < G_{Q_1} - \epsilon \quad \text{and} \quad F(u_i) \leq F_{Q_3} - \epsilon \tag{5}$$

Here F_{Q_3} = 0.15 corresponds to the 3^{rd} quantile of the F score distribution. These users have toxicity scores that has a predominant down trend ($1 \rightarrow 0$) for their comments. Hence, they should have large negative G scores and F scores lower than its third quantile, as larger F scores would rather suggest fickle-minded user behavior.

Radical Users: We define the rule to identify users whose toxicity scores predominantly increase over time as:

$$G(u_i) > G_{Q_3} + \epsilon \quad \text{and} \quad F(u_i) \leq F_{Q_3} - \epsilon \tag{6}$$

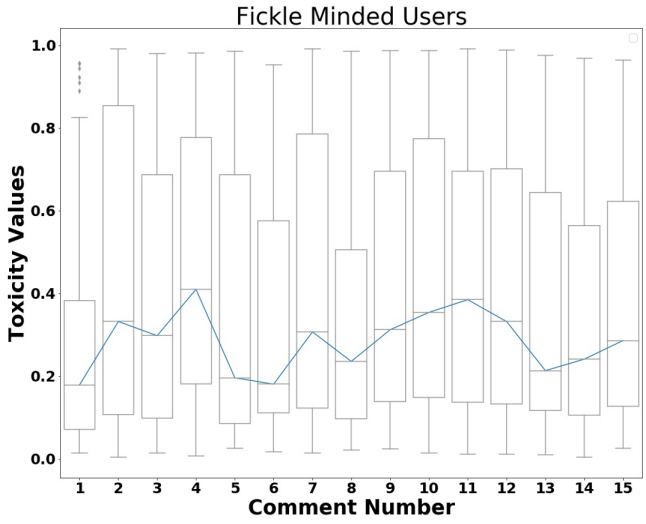
These users have toxicity scores with a primarily up trend ($0 \rightarrow 1$) for their comments. Hence, they should have large positive G scores, and F scores lower than its 3^{rd} quantile.

We used ϵ = 0.01 in all our experiments. The different user behaviors based on temporal toxicity patterns are illustrated in Figures 3c and 4. The overlaps between these clusters in Figure 3c correspond to the users whose toxicity patterns are fuzzy combinations of two or more such user behaviors. This makes it difficult to correctly assign these users to a particular behavior type based on their temporal toxicity patterns in the subreddits.

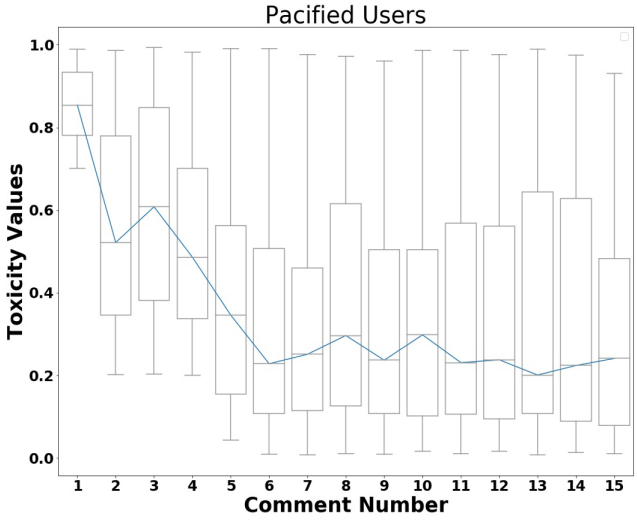
RESULTS

We first identify the total number of users belonging to each user behavior class based on their temporal toxicity patterns in Table 2. The temporal toxicity pattern for each user behavior class is highlighted in Figure 4 and an example of each behavior class is provided in Table 3. We further showcase the total number of comments made by the users belonging to each behavior class in Table 2. Moreover, we highlight the top 3 subreddits based on the ratio of the total number of comments made by the users in a given subreddit to the total number of comments made by users of a particular behavior class. These results are depicted in Table 2.

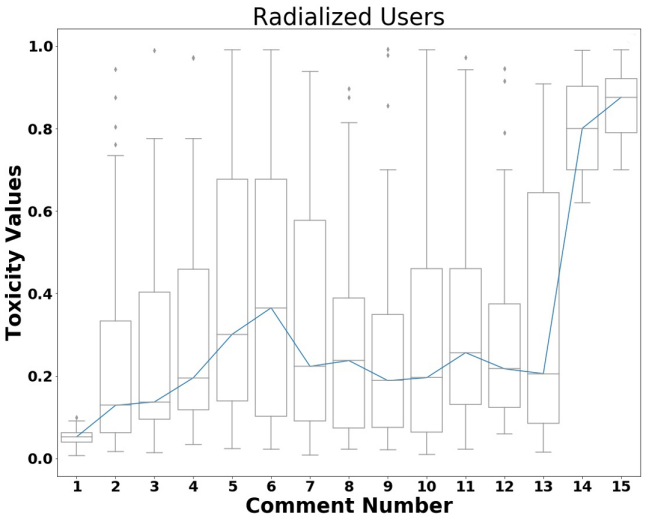
From Table 2, we find that the maximum number of users for each behavior class commented in Ask Reddit discussions. This is because Ask Reddit community had the maximum number of comments from all of these users (183, 084). Interestingly, the subreddit Movies and World News had the 2^{nd} and 3^{rd} highest number of radicalized user comments, while the Pics subreddit had the 2^{nd}



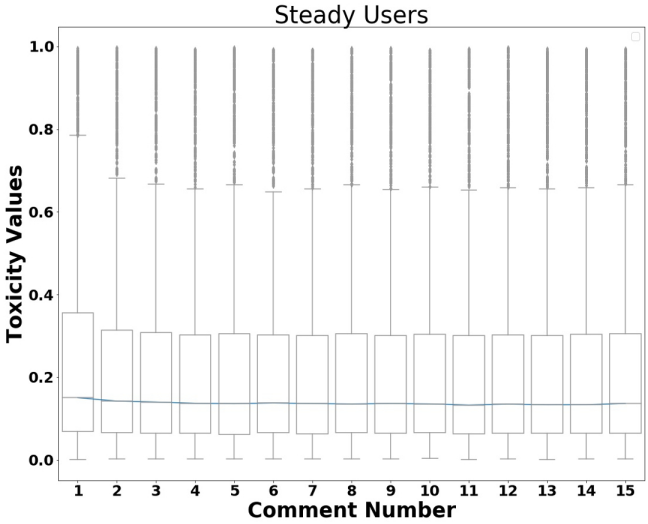
(a) Fickle-Minded Users



(b) Pacified Users



(c) Radical Users



(d) Steady Users

Figure 4: Figure illustrates the different types of toxic user behaviors. Figures 4a, 4b, 4c and 4d represents the range of toxicity scores for Fickle-Minded, Pacific, Radical and Steady users respectively in case of the 1^{st} 15 comments made by them. By connecting the medians, we capture the *different shapes* that the temporal toxicity patterns can take.

Table 2: Number (#) of users per behavior class and top 3 subreddits where these users have participated most. The number in $(\cdot)$ is the fraction of total comments made by these users in an individual subreddit.

Class	#Users (% of all	#Comments	1^{st} Subreddit	2^{nd} Subreddit	3^{rd} Subreddit
Fickle-Minded	9,664 (31.2%)	226,606	Ask Reddit (0.382)	Pics (0.111)	World News (0.108)
Pacified	8,002 (25.8%)	811,004	Ask Reddit (0.312)	World News (0.102)	Movies (0.099)
Radicalized	7,858 (25.4%)	102,416	Ask Reddit (0.297)	Movies (0.106)	World News (0.103)
Steady	5,461 (17.6%)	97,742	Ask Reddit (0.268)	IAmA (0.108)	Movies (0.106)

largest fickle-minded population. Similarly, the IAmA subreddit has 2^{nd} largest number of comments from users with steady toxicity patterns.

From this sample, we find that most users (31.2%) are fickle-minded, alternating between toxic and non-toxic comments. Pacified users are second-most common (25.8%), but the difference to radicalized users in terms of frequency (25.4%) is not large. Steadily toxic users are the least frequent (17.6%) in the sample.

Interestingly, results in Table 2 also indicate that pacified users tend to contribute more comments (M = 101.4) than all other user types combined (M = 18.6), and this difference is highly significant, $t(10,077.3) = 139.6$, $p < .0001$. The difference can be interpreted by the fact that most comments in our dataset exhibit low toxicity scores, so transitions from toxic to non-toxic commenting tends to be proportionally higher than other shifts.

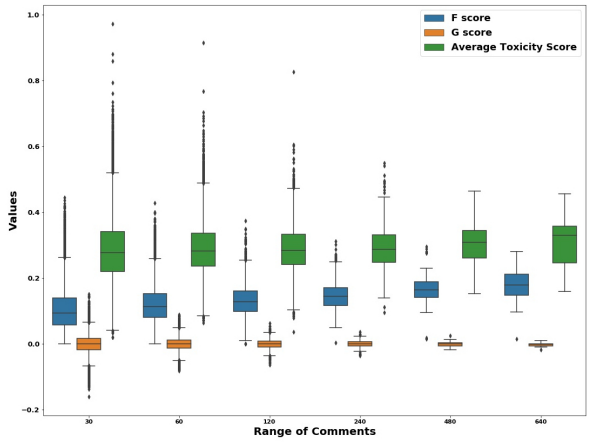


Figure 5: Here the 1^{st} value on x-axis represents the scores for all users who made a minimum of 15 and maximum of 30 comments. The 2^{nd} value represents scores for all users who made $[30, 60]$ comments and so on. The boxplots show that average toxicity, F and G score distribution *stabilizes* for users who are highly active (made large number of comments over time) in these long discussions on Reddit.

DISCUSSION

Contribution

The major contributions of this research are two-fold: first, (1) defining four types of toxic users based on user behaviors over time: (a) Stable Users (toxicity of the users' comments does not noticeably change over time), (b) Radicalized Users (increased toxicity over time), (c) Pacified Users (decreased toxicity over time), and (d) Fickle-Minded Users (fluctuating toxicity, i.e., both increasing and decreasing over a period). Secondly, (2) we provide quantitative metrics to detect these user types from online social media data.

In the context of social media users in general and Reddit in particular, our findings complement the earlier studies on toxicity in Reddit [13, 19, 20] that has largely focused on comparing the toxicity of different subreddits and thereby remained at a community level rather than investigating user level toxicity. Here, we show that toxicity patterns vary both by subreddit and by user. For instance, the Ask Reddit community has maximum number of comments from each user behavior class. The fact that fickle-minded users were the largest group in our analyzed sample is interesting. We interpret this finding such that switching between toxic and non-toxic commenting styles can be seen to describe typical social media behavior of users, whereupon users navigate many discussion threads and topics and change their attitudes based on the topic (and perhaps the participants' triggering messages [2]. For some topics and subreddits, users may exhibit less toxicity, whereas for more controversial topics (e.g., politics, religion [29]), their behavior can become more toxic. This also frames the interesting question: are some users more drawn to controversial topics than other users? While we did not explicitly investigate this question, our results indicate that this might be one potential explanatory factor behind the fluctuation of toxicity, as Reddit users typically engage in conversations in not only one, but several subreddits.

Implications for Online User Experiences

The implication for Reddit is that more moderation is needed for communities that are predisposed to radicalization. Average toxicity in a community tends to be lower than neutral (see Figure 2b). Moreover, when removing sporadic users, the variance in average toxicity scores decreases in all subreddits as showcased in 2c because participants seem to have greater stability in their behavior as they comment more (see Figure 5).

For wider purposes, we note that the approach presented in this research can be generalized to any other user data from any social media or online platform, or combination thereof, as long as User IDs, message timestamps, and message toxicity scores are available. The generalizability comes from the fact that we are using the properties of the distribution of the proposed F and G score metrics to classify the users' toxicity pattern. In particular, the rules that govern the toxicity patterns are derived from the *quantiles* of distributions of the F and G scores and are thus independent from a particular distribution of the data.

Regarding applicability, our approach could be employed in systems for automatic moderation. For example, the F and G score estimating the temporal toxicity trends of users can be incorporated in a machine learning model to identify different user behavior patterns and to moderate online communities by identifying users who are steadily toxic or are becoming radicalized. Using information on the past history of the user is likely to provide better results for automatic moderation than only looking at individual comments.

Design Implications

Because we did not carry out a research study with a direct user intervention, but rather an observational study, our findings do not provide direct implications for user interfaces (UI) of online systems.

Regarding user experience (UX) in online communities, however, our findings suggest that designing ways to monitor users' toxicity over time is worthwhile, as there the user behavior varies and can be

Table 3: Snippet of the comments made by a user belonging to Fickle-Minded, Pacified, Radicalized, and Steady user behavior classes. Here time is in hours and 0 represents the start time of a comment for a given user.

Class	Time	Comment	Toxicity Score
Fickle-Minded	0	Submitter said controversial, not controversial within reddit.	0.033
	0.109	As the first black female head of the Ku Klux Klan, I'd like to say ***America stinks!***	0.876
	0.204	Submitter didn't specify *on reddit*.	0.161
	0.383	On the other hand, men *suck* at giving birth.	0.925
	0.443	Nice try, Republicans.	0.110
	0.477	He has a dick, for god's sake! Only pre-op.	0.957
	547.7	Wiccans! Holy shit, *witchcraft bombs!*	0.957
	557.5	I hope someone is hitting "contribute a better translation".	0.275
Pacific	0	(Walking offstage)"...I don't even want the fucking thing.."	0.903
	84.77	Da-Dink Dink!!	0.578
	84.80	Haha- holy shit I can see this happening.	0.898
	85.01	"We came, we saw, we kicked it's—"	0.214
	85.19	Yeah- the "slow powerful music" thing is getting old with trailers. Every major film has it's theme slowed down.	0.120
	253.0	Elizabeth Olsen is a great actress, and she fits the role really well. But she is so damn hot, man. Whew.	0.309
	253.1	I didn't see any giant metallic looking bracelets like the 70's Spider-man.	0.158
	567.5	And make sure while you film it, chew some gum REALLY loudly like that English guy did.	0.238
	603.1	But the guns thing was a dream. It was a result of a post apocalyptic world.	0.122
Radicalized	0	But sea salt has a delicious taste that is different from just regular kosher salt.	0.027
	0.007	Just because no one in your family can cook doesn't mean red sauce is "bad."	0.249
	0.0192	The only sweetness that should be in sauce is delicious tomato sweetness. Any added sweetness can go to hell.	0.393
	0.0336	That's ridiculous, food can often enhance drinks and vice versa.	0.339
	0.0397	I'm pretty sure this is bullshit.	0.903
	0.059	This is bad for you and I imagine would also taste pretty awful.	0.616
	0.228	You sleep on your back? I've never slept with anyone who slept on their back.	0.575
	0.236	It's not weird, you're weird.	0.417
	0.261	What's wrong with your balls that you don't want them to touch your sheets?	0.789
	0.328	How is this even an issue? Put the fucking curtain INSIDE THE SHOWER, GENIUSES.	0.938
Steady	0.186	Shut your fucking mouth, nigger lover.	0.997
	0.339	Shut up, liberal faggot.	0.988
	2.366	I am not a troll. I speak the truth you fucking piece of shit. Kill yourself.	0.990
	8.506	Kill yourself now, douchebag. Hurry. This is an order from Almighty God.	0.963
	8.564	I will piss, take a shit, and spit on your ugly face. You are my fucking slave, little boy.	0.992
	9.745	Go and commit suicide you worthless, useless piece of dog shit.	0.991
	10.19	I fucked your whore of a mother. She loved it, the skank she is.	0.993

divided into different typologies. Visible metrics, gauges, penalties, or forms of positive reinforcement and feedback are among the possible UI/UX interaction techniques that could be leveraged using our statistical metrics as inputs for design implementations.

Given the complexity of issues like automated content filtering, freedom of speech, privacy, and other societal aspects, readers should be careful with drawing straightforward conclusions. Design choices in the space of online toxicity require the careful consideration of ethical and legal ramifications. For example, what would be the ethical, legal, and appropriate use of the ability to identify and perhaps modify users that are shifting to more extreme views? Could this "power" be used to shape the users' trajectory? Addressing these questions is a cross-disciplinary effort – for example, collaboration with a civil rights lawyer may greatly inform the next stages of this work.

Limitations and Future Research

An interesting aspect of the user behavior classes is that they can be sub-divided into micro classes. For example, the steady users can be sub-categorized into steady toxic, steady-pacific and steady-neutral based on their average toxicity over their lifetime. Similarly, pacific users can be divided into three categories: (a) toxic-to-neutral, (b) neutral-to-pacific, and (c) toxic-to-pacific. The same can be done for radical users. For future work, it is possible to develop rules, using F, G and average toxicity score distributions that can capture these granular classes of user behavior.

With further development, our approach could also be applied for early detection of radicalization, defined as increased toxicity of a user over time (above a specific threshold value). This would be useful for helping online communities to identify and take action (e.g., helpful interventions) to preserve the quality of discussion. Future research could, therefore, focus on analyzing users who were initially showing little or no toxic behavior but who became toxic over time. Currently, the instruments for managing and helping users with problems of toxic behavior are lacking, and the management of toxicity is often done by ignoring the users (moderation/blocking) rather than analyzing the paths of becoming toxic and devising interventions that would help users reduce their toxic behavior. To this end, our approach of modeling a user's toxicity history can, with further development, be helpful.

Our approach could also be expanded by incorporating context in which toxicity takes place. Previous research has shown that context plays a major role for users' behavior in social media [1, 4]. The context could be represented by the topic of the discussion, or the subreddit the user is writing to. For example, a user that is participating relatively more on subreddits that deal with topics he or she is hateful towards could have a higher history of toxicity than a user that is participating relatively more in discussions whose topics he or she enjoys. To properly capture the impact of the context, the model would need to normalize its impact, which is an area of future work.

Another aspect for improvement is the evaluation methods applied: in our dataset, there is no ground truth, so we cannot judge if the approach can predict a target metric such as radicalization or a user being banned. Future research can can use the F and G scores in a dataset that has information indicating that certain users

are banned due to inappropriate behavior, to investigate if such an event could be predicted.

CONCLUSION

We defined two quantitative metrics for segmenting online users to four types based on their toxic behavior: Steady, Radicalized, Pacified, and Fickle-Minded Users. Applying these metrics to a sample of social media users in Reddit, results indicate that fickle-minded users that switch between making toxic and non-toxic comments are most common (31.2%) in the sample we investigated. Pacified users are second-most common (25.8%), although their difference to radicalized users in terms of frequency (25.4%) is not large. Steadily toxic users are the least frequent (17.6%) in the sample. Further work is required to understand the reasons behind toxicity variation of users, and to design early-warning methods to mitigate toxicity outbreaks and to help users at risk of radicalization.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their helpful comments.

REFERENCES

[1] Fotis Aisopos, George Papadakis, Konstantinos Tserpes, and Theodora Varvarigou. 2012. Content vs. context for sentiment analysis: a comparative analysis over microblogs. In *Proceedings of the 23rd ACM conference on Hypertext and social media* (2012). ACM, 187–196.

[2] Hind Almerekhi, Haewoon Kwak, Joni Salminen, and Bernard J Jansen. 2019. Detecting Toxicity Triggers in Online Discussions. In *In the proceedings of the 30th ACM Conference on Hypertext and Social Media (HT'19)*. Hof, Germany, 291–292. https://doi.org/10.1145/3342220.3344933

[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 Web Conference 2020 (WWW '20)*. Association for Computing Machinery, Taipei, Taiwan, 3033–3040. https://doi.org/10.1145/3366423.3380074

[4] David Bamman and Noah A. Smith. 2015. Contextualized Sarcasm Detection on Twitter. *ICWSM* 2 (2015), 15.

[5] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can'T Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 31:1–31:22. https://doi.org/10.1145/3134666

[6] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–30. Publisher: ACM New York, NY, USA.

[7] Anastasia Giachanou, Ida Mele, and Fabio Crestani. 2016. Explaining sentiment spikes in twitter. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (2016). ACM, 2263–2268.

[8] Anastasia Giachanou, Ida Mele, and Fabio Crestani. 2017. A Collection for Detecting Triggers of Sentiment Spikes. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2017). ACM, 1249–1252.

[9] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing" trolling" in a feminist forum. *The information society* 18, 5 (2002), 371–384.

[10] 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* (2015) *(Lecture Notes in Computer Science)*. Springer, Cham, 49–66. https://doi.org/10.1007/978-3-319-27433-1_4

[11] Teo Keipi, Matti Näsi, Atte Oksanen, Pekka Räsänen, Matti Näsi, Atte Oksanen, and Pekka Räsänen. 2016. *Online Hate and Harmful Content : Cross-National Perspectives*. Routledge. https://doi.org/10.4324/9781315628370

[12] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[13] 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 on World Wide Web* (2018). International World Wide Web Conferences Steering Committee, 933–943.

[14] K. Hazel 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 (June 2017), 991–1010. https://doi.org/10.1108/IntR-02-2017-0072

[15] Jean-Charles Lamirel, Pascal Cuxac, Raghvendra Mall, and Ghada Safi. 2011. A new efficient and unbiased approach for clustering quality evaluation. In *New Frontiers in Applied Data Mining*. Springer, 209–220.

[16] Rocco Langone, Raghvendra Mall, Carlos Alzate, and Johan AK Suykens. 2016. Kernel spectral clustering and applications. In *Unsupervised Learning Algorithms*. Springer, 135–161.

[17] Kaitlin Mahar, Amy X. Zhang, and David Karger. 2018. Squadbox: A Tool To Combat Online Harassment Using Friendsourced Moderation. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–4.

[18] Raghvendra Mall, Rocco Langone, and Johan AK Suykens. 2013. Kernel spectral clustering for big data networks. *Entropy* 15, 5 (2013), 1567–1586.

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

[20] 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 *SpringerLink* (2017). Springer, Cham, 51–56. https://doi.org/10.1007/978-3-319-57351-9_6

[21] Adewale Obadimu, Esther Mead, Muhammad Nihal Hussain, and Nitin Agarwal. 2019. Identifying Toxicity Within YouTube Video Comment. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 214–223.

[22] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, et al. 1998. The pagerank citation ranking: Bringing order to the web. (1998).

[23] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016). ACM, 1114–1125.

[24] Haji Mohammad Saleem, Kelly P. Dillon, Susan Benesch, and Derek Ruths. 2017. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. *arXiv:1709.10159 [cs]* (Sept. 2017). http://arxiv.org/abs/1709.10159 arXiv: 1709.10159.

[25] Joni Salminen, Hind Almerekhi, Partha Dey, and Bernard J. Jansen. 2018. Inter-rater agreement for social computing studies. In *Proceedings of The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018)* (2018). Valencia, Spain.

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

[27] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, Shammur Absar Chowdhury, and Bernard J. Jansen. 2020. A Literature Review of Quantitative Persona Creation. In *Proceedings of the ACM Conference of Human Factors in Computing Systems (CHI'20)* (2020). ACM, Honolulu, Hawaii, USA. https://doi.org/10.1145/3313831.3376502

[28] Joni Salminen, Maximilian Hopf, Shammur A. Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J. Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences* 10, 1 (2020), 1. https://doi.org/10.1186/s13673-019-0205-6

[29] Joni Salminen, Sercan Sengün, Juan Corporan, Soon-gyo Jung, and Bernard J. Jansen. 2020. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLOS ONE* 15, 2 (Feb. 2020), e0228723. https://doi.org/10.1371/journal.pone.0228723 Publisher: Public Library of Science.

[30] Joni Salminen, Fabio Veronesi, Hind Almerekhi, Soon-gyo Jung, and Bernard J. Jansen. 2018. Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings. In *Proceedings of The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018)* (2018). Valencia, Spain.

[31] Philipp Singer, Emilio Ferrara, Farshad Kooti, Markus Strohmaier, and Kristina Lerman. 2016. Evidence of online performance deterioration in user sessions on Reddit. *PloS one* 11, 8 (2016). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0161636

[32] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (2013). 1631–1642.

[33] 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* (2012). ACM, 1481–1490.

[34] Mike Thelwall. 2017. The Heart and soul of the web? Sentiment strength detection in the social web with SentiStrength. In *Cyberemotions*. Springer, 119–134.

[35] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology* 61, 12 (2010), 2544–2558.

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

[37] Ahmed Waqas, Joni Salminen, Soon-gyo Jung, Hind Almerekhi, and Bernard J. Jansen. 2019. Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLOS ONE* 14, 9 (Sept. 2019), e0222194. https://doi.org/10.1371/journal.pone.0222194

[38] 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* (2017) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1391–1399. https://doi.org/10.1145/3038912.3052591