A longitudinal study of the top 1% toxic Twitter profiles

Hina Qayyum
Macquarie University
Sydney, Australia
hina.qayyum@mq.edu.au

Benjamin Zi Hao Zhao
Macquarie University
Sydney, Australia
ben_zi.zhao@mq.edu.au

Ian D. Wood
Macquarie University
Sydney, Australia
ian.wood@mq.edu.au

Muhammad Ikram
Macquarie University
Sydney, Australia
muhammad.ikram@mq.edu.au

Mohamed Ali Kaafar
Macquarie University
Sydney, Australia
dali.kaafar@mq.edu.au

Nicolas Kourtellis
Telefonica Research
Barcelona, Spain
nicolas.kourtellis@telefonica.com

ABSTRACT

Toxicity is endemic to online social networks (OSNs) including Twitter. It follows a Pareto-like distribution where most of the toxicity is generated by a very small number of profiles and as such, analyzing and characterizing these "toxic profiles" is critical. Prior research has largely focused on sporadic, event-centric toxic content (i.e., tweets) to characterize toxicity on the platform. Instead, we approach the problem of characterizing toxic content from a profile-centric point of view. We study 143K Twitter profiles and focus on the behavior of the top 1% producers of toxic content on Twitter, based on toxicity scores of their tweets availed by Perspective API. With a total of 293M tweets, spanning 16 years of activity, the longitudinal data allows us to reconstruct the timelines of all profiles involved. We use these timelines to gauge the behavior of the most toxic Twitter profiles compared to the rest of the Twitter population. We study the pattern of tweet posting from highly toxic accounts, based on the frequency and how prolific they are, the nature of hashtags and URLs, profile metadata, and Botometer scores. We find that the highly toxic profiles post coherent and well-articulated content, their tweets keep to a narrow theme with lower diversity in hashtags, URLs, and domains, they are thematically similar to each other, and have a high likelihood of bot-like behavior, likely to have progenitors with intentions to influence, based on high fake followers score. Our work contributes insight into the top 1% toxic profiles on Twitter and establishes the profile-centric approach to investigate toxicity on Twitter to be beneficial. The identification of the most toxic profiles can aid in the reporting and suspension of such profiles, making Twitter a better place for discussions. Finally, we contribute to the research community with this large-scale and longitudinal dataset[1], annotated with six types of toxic scores.

CCS CONCEPTS

• **Information systems** → **Social networks**; • **Social and professional topics** → **Hate speech**; **Political speech**.

KEYWORDS

Twitter, profile, toxicity, longitudinal, measurement, Perspective score

ACM Reference Format:
Hina Qayyum, Benjamin Zi Hao Zhao, Ian D. Wood, Muhammad Ikram, Mohamed Ali Kaafar, and Nicolas Kourtellis. 2023. A longitudinal study of the top 1% toxic Twitter profiles. In *15th ACM Web Science Conference 2023 (WebSci '23), April 30-May 1, 2023, Evanston, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3578503.3583619

1 INTRODUCTION

Verbal misbehavior and toxicity on Online Social Networks (OSNs) are receiving a huge amount of attention in the research community, with efforts to identify [5, 7, 19, 59, 63], characterize [6, 8–10, 14], and automatically detect [3, 11, 17] online misbehavior, especially on Twitter. Despite all these ongoing efforts, toxicity has increased over time. We note that almost all efforts to study toxicity on Twitter come from the content study of tweets posted around sporadic high-profile campaigns and events such as elections [23], important world events of COVID19 and the MeTooMovement [2, 24], or controversies about topics like Bitcoin [38]. However, these studies do not explore the prolonged involvement of a profile in spreading toxic content, so its utility in the characterization of overall toxicity was hindered. Works like [45, 60] investigate toxic profiles responsible for disseminating toxic content on small manually annotated datasets. In essence, efforts toward the automatic detection and characterization of toxicity on Twitter are mostly event-centric or small-scale, user-centric. This scenario leaves a gap in understanding the entire landscape of misbehavior on Twitter.

Toxic content follows a Pareto-like distribution on Twitter [43], hence we focus on the most toxic profiles in our dataset based on the median Perspective "Toxicity" score of the profile's tweets. We contrast these profiles with the remainder of our dataset to find out how much their behavior is different from base Twitter profiles. We focus on research questions that will assist us in better understanding toxic profiles:

- Are toxic profiles prolific content generators, with a specific tweeting pattern?

[1] https://github.com/hqayyum/twitter_top_toxic_1percent

WebSci '23, April 30-May 1, 2023, Evanston, TX, USA

ACM ISBN 979-8-4007-0089-7/23/04…$15.00
https://doi.org/10.1145/3578503.3583619

- Do toxic profiles tweet in a legible way to effectively convey their message?
- What type of misconduct is expected of a toxic profile?
- Do toxic profiles leverage auxiliary content, such as URLs and hashtags?
- Do toxic profiles demonstrate special trends with respect to name, location, counts of friends or followers, and such?
- Can we expect very toxic profiles to be bad bots?

Our dataset, described in §3, is seeded with seven smaller public datasets from past works studying online misbehavior on Twitter. These seed datasets cover multiple themes of online misbehavior: hostility, racism, abuse, hatefulness, homophobia, spam, and sexism, and are balanced in the number of toxic and non-toxic users. A key limitation of the seed datasets is that users are classified as toxic or not from the content of a single or a few tweets, which does not allow deeper analysis of the users' average toxic behavior. To enable such analysis, we crawl the tweet timeline of each of the users present in the seed datasets. Our resulting dataset contains 142,987 (143K) Twitter profiles and 293,401,161 (293M) individual tweets posted between 2007 and 2021. Human annotations are untenable given the size of our dataset, hence, we turn to Google's Perspective APS [22] models to assign toxicity scores to each tweet, providing estimates of the following types of misbehavior: *Toxicity, Severe Toxicity, Identity Attack, Inflammatory, Threat, and Insult.* We use only the production-ready scores from the Perspective API, which provide highly reliable estimates.

In §4, we investigate these highly toxic profiles with respect to tweeting frequency and dynamics, drawing on the distribution of inter-tweet times and a measure of burstiness. Next in §5, we look at the tweet content. We explore the Perspective scores and their consistency among each profile's tweets using the Gini Index [20]. Next, we study the number and quality of hashtags and URLs with help of Fortiguard, a service that categorizes URLs by topic. We then perform a readability analysis of tweets, using Flesch reading ease and difficulty scores, Linsear write score, and the Automatic Readability Index (ARI). In §6, we characterize toxic profiles based on the profile metadata including the number of friends, followers, statuses, favorites, membership of lists, location, creation date, and profile status. In addition, we use to apply the Botometer API [48] to our profiles, obtaining scores that quantify astroturfing, spamming, fake followers, self-declared bots, and financial bots.

This work makes the following main contributions:

- We collect and curate a longitudinal dataset of tweets, spanning 16 years, consisting of 293M tweets (§3) and augmented with six perspective scores. To our knowledge, this is the largest annotated dataset on online misbehavior. To foster further research, upon publication, we plan to share our dataset with the research community.
- We identify that the top 1% toxic profiles post fewer, shorter, but more articulate tweets than the rest. We find that the Gini Index of Perspective scores on each toxic profile's tweets is lower, indicating consistency of misbehavior among their tweets.
- We observe that the highest Perspective scores among tweets from toxic profiles are "Inflammatory" and "Insult", and that "Identity attack" is relatively low, especially when compared to where it sits among baseline tweets.

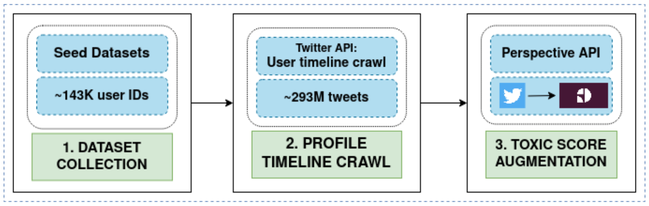

Seed Datasets
~143K user IDs
1. DATASET COLLECTION
Twitter API: User timeline crawl
~293M tweets
2. PROFILE TIMELINE CRAWL
Perspective API
3. TOXIC SCORE AUGMENTATION

Figure 1: Our dataset collection and augmentation pipeline; including dataset collection, user timeline crawl, and augmentation.

- The top 1% toxic profiles tend to use fewer but more coherent and similar URLs and hashtags.
- We find that the top 1% toxic profiles have lower friends and follower counts than baseline profiles.
- We observe a notable increase in the creation of toxic profiles between 2014-2016. Interestingly, we note that despite being the most toxic, none of the top 1% profiles has been deactivated, banned, or deleted in the 18 months that passed between timeline data crawling and profile metadata extraction.
- Notably, we identify that just under half the top 1% toxic profiles would be classified by Botometer as fake followers, which is important evidence of instrumented trolling campaigns on Twitter.

2 RELATED WORK

Online misbehavior detection on social networks and on Twitter has been extensively explored by several studies such as [18, 21, 44, 47, 58]. However, almost all approaches to indicate hate or toxicity is content-centric, the inherent shortfall of collecting and annotating toxic tweets is due to the incompleteness and insufficiency of OSN text i.e., tweets, and the sparsity of toxic hateful speech. These limitations are often amplified by oversimplifying the problem, such as considering only tweets collected around extremist events or collected with keyword searches. In this work, we partially address these limitations by accumulating a user-centric dataset, such work is done on a very limited level by [45].

Past studies [21, 44] have relied on human annotations to differentiate between toxic and non-toxic tweets. Our work leverages the ML models of the Perspective API to rate the collected tweets to explore the different dimensions of misbehavior beyond the prior works, and at the larger scale of data, i.e., 293M tweets. Similarly, previous work [25, 27] has studied Google's Perspective API [22] and its resilience against adversarial attacks. Those studies leveraged Perspective API to score and analyze the toxicity of tweets. This work takes precedence over these studies in terms of the size of the data set (293M tweets) and the number of misbehavior dimensions not studied in the past, namely Insult, Inflammatory, Threat, and Identity Attacks.

Automated accounts, paid bots, or trolls' role in toxic and false content creation and dissemination [15, 40, 61], is the base of a consistent spread of toxicity on OSNs. Content-based features best predict coordinated efforts of these malefactors [13], but unsupervised ML for detection of coordinated efforts of profiles in carrying these operations are infeasible at scale [1]. Analysis of unlabeled profiles' longitudinal and unlabeled content provides a characterization of the most toxic users and their content on Twitter.

Seed Dataset					
Dataset	**TIDs**	**UIDs**	**RUIDs**	**Labels/Keywords**	**Annotation**
[21]	149,823	-	895	sexist, racist, homophobic, religion, other hate, no hate	Amazon Mechanical Turk
[29]	817,344	-	19,859	Keyword: MeTooMovement	Twitter Streaming API
[44]	-	100,385	100,385	hateful, not hateful	CrowdFlower (appen)
[18]	-	98,378	98,378	normal, abusive, spam, hateful	CrowdFlower (appen)
[28]	10,583	-	324	benevolent, hostile, other	SVM (TF-IDF)
[58]	16,907	-	891	sexist, racist, neither	CrowdFlower (appen)
[57]	6,909	-	870	sexist, racist, both, neither	CrowdFlower (appen)

Table 1: Overview of 7 datasets used as a seed with a collection of User IDs (UIDs) or Tweet IDs (TIDs), whatever was made publicly available. Note that TIDs were used to recover User IDs (RUIDs).

3 DATASET COLLECTION METHODOLOGY AND CHARACTERIZATION

In this section, we detail our methodology for data collection and augmentation. We introduce our seed datasets, detail how the timelines of Twitter profiles were crawled, and how we augmented the collected tweet data with Google's Perspective API. We finish with an overall characterization of the augmented dataset, and our definition of the top 1% toxic Twitter profiles i.e., the upper echelon of toxic profiles as determined by the "Toxicity" Perspective score of their tweets. A summary of the selected datasets, details of their size, and labels can be found in Tab. 1 and Tab. 2. Further, a flowchart of our data collection and augmentation methodology is overviewed in Fig. 1.

3.1 Crawling User Timelines

To provide broad coverage of themes, we merged 7 balanced seed datasets (in terms of toxicity) from various topic domains (Tab.1), and we estimate that this results in a dataset that closely reflects the Twitter community. Based on Twitter's terms and conditions, Twitter User IDs (UIDs) and tweet content cannot be publicly shared. Consequently, our seed datasets contain only Tweet IDs (TIDs) and their respective annotation (Tab.1). Thus, the first step was to query Twitter's API [53] to recover the UID responsible for each TID. Next, we queried the Twitter API to retrieve each UIDs' timeline. Twitter API only permits the retrieval of 3,200 most recent tweets from a profile, whilst not the entire timeline, this still allows us to study a significant length of the historical record of each profile and its evolving behavior. We were unable to retrieve tweets from banned and deleted profiles. From the retrieved tweets, we extract relevant details such as the text, timestamp of tweet creation, hashtags, and URLs, shared within tweets. For this study, we only consider English tweets.

3.2 Dataset Augmentation with Perspective API

In the aforementioned seed datasets, a profile was labeled toxic or not based on mostly one or at most as few as 3 tweets. However, it is unrealistic to assume that this label can be a representation of a profile's overall tweeting behavior. We needed a measure of misbehavior for the entire timelines of the 143K profiles we crawled. We obtain this quantitative measure of misbehavior across all tweets through Google's Perspective API [22]. The Perspective API provides multiple Convolutional Neural Network (CNN) based models trained with GloVe word embeddings [39] for the evaluation of misconduct in text. This API offers 16 models of which 10 are considered experimental. Each model, for every given input text, yields a score from 0 to 1 representing the intensity of a type of misbehavior. We retrieve scores from the 6 production-ready Perspective API models [51]:

Recovered Profiles	142,987	**Total Profiles >10 Tweets**	138,533
Total Tweets	293,401,160	**Avg Total Tweets per profile**	2,051
Unique Tweets	230,283,810	**Avg Unique Tweets per profile**	1,610

Table 2: Summary of Seed datasets (cf. Tab. 1) and recovered profiles.

- *Toxicity*: Rude, disrespectful, or unreasonable comments, likely to make people leave a discussion.
- *Severe Toxicity*: Comments are very likely to make users leave a discussion or give up sharing their opinion.
- *Identity Attack*: Negative or hateful comments targeting someone's identity, ethnicity, sexual orientation, and other characteristics.
- *Inflammatory*: Intended to provoke or inflame others.
- *Insult*: Insulting or negative comments towards a person or a group of people.
- *Threat*: Intentions to inflict pain, injury, or violence against an individual or group.

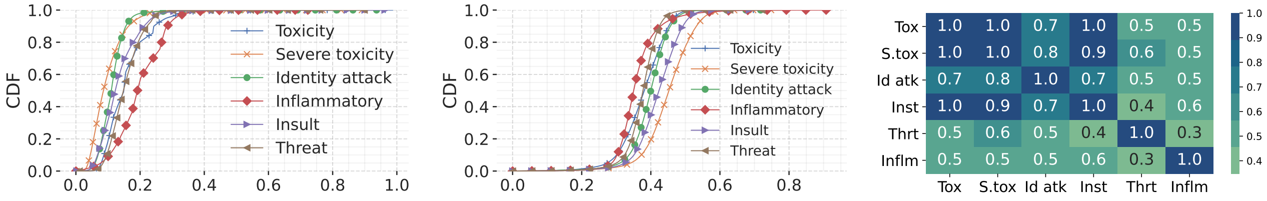
We query all 293M tweets from 143K profiles for the 6 perspective scores. The collective time for dataset collection and augmentation was about 5 months and the size of the augmented data is close to 2TB. We characterize the dataset in the following section.

3.3 Characterization of Augmented Dataset

To better understand the composition of our final dataset, we first inspect the Cumulative Distribution Function (CDF) of each Perspective score, across all tweets in all profile timelines (Fig. 2a). We can observe that the median score of a tweet's score for any of the six misbehavior dimensions lies approximately in the range of 0.1 – 0.2. A steady rise in the curve in the low ranges of scores indicates that a majority of tweets do not strongly exhibit any specific form of misbehavior.

Additionally, the strongest signal for misbehavior is in the dimension of Inflammatory content. A tail is observed with a small number of tweets acting as an exception to the rule, propagating a large amount of misbehavior (score → 1.0).

To measure a given profile's consistency in toxic behavior, we leverage the Gini Index over a profile's tweet perspective scores. The ***Gini Index*** was originally proposed to measure the concentration of wealth [20] within a population, but can equally be used to identify the extent of toxicity distribution among a profiles' tweets. A consistent set of scores (low or high) will produce a Gini Index closer to 0, whereas high variability scores produce a Gini Index approaching 1. We visualize the CDFs of the Gini index for all profiles across six dimensions of toxicity in Fig. 2b. We see that the median Gini is between 0.35 for Insult and 0.46 for Severe Toxicity. The majority of profiles have a Gini-Index in a range of 0.3-0.5



(a) Perspective API scores

(b) Gini index

(c) Pearson Correlation

Figure 2: (a) Cumulative Distribution Function (CDF) of median Perspective API scores per profile, across all tweets; (b) Gini Index calculated on all 6 Perspective scores per profile; and (c) Pearson pairwise correlation matrix across amongst all Perspective API scores.

with a median of 0.4, which indicates that these profiles are not consistently toxic; however, we observe profiles in the lowest and highest range of Gini-Index 0-0.3 and > 0.6, pointing to profiles being constantly toxic. The Gini Index of Inflammatory scores is the lowest, implying that Inflammatory behavior is exhibited most consistently. Fig. 2c illustrates the correlation among all perspective scores, we note that toxic profiles are also likely to produce tweets that show identity attack and insult, and inflammation.

3.3.1 **Takeaways:**

- With the range of median toxicity scores for all Twitter profiles between 0.14-0.16, we note that the majority of tweets on Twitter are not toxic.
- The majority of Twitter profiles have a low Gini index (0.4), thus they skew towards being consistently toxic across their tweets.

3.4 Top 1% Toxic Twitter Profiles

We identify and study the upper echelon of toxic profiles as determined by the average "Toxicity" Perspective score of their tweets. We sort all the 143K profiles based on the median toxicity scores calculated on the toxicity scores of all tweets in their respective timelines (at this point we remove all profiles with less than 10 tweets and consider the rest 138K profiles). As a final step, motivated by the fact that toxicity follows the Pareto effect on Twitter [43], we skim the top 1% of profiles as a sample of the most toxic Twitter profiles, we refer to these profiles (1,380) as ***'Top 1% toxic profiles'***. Note that 80% of the 1% contribute 1000 or more tweets. We shall compare toxic profiles with the remainder of the population (136,620, or 99%), referred to as ***'baseline profiles'*** in text. To contextualize the 1% on toxicity scores, the 1% profiles have a median toxicity of 0.40, while those in the baseline have a median of 0.15. Further, almost all tweets from the 1% have a Toxicity larger than 0.35, whereas it lies at <3% for the baseline.

4 TWEET FREQUENCY AND PATTERN ANALYSIS

Impactful Twitter profiles post a significant number of tweets over time. In this section, we shall investigate the number of total and unique (tweets with the exact same content were removed) tweets, and the percentage of unique tweets posted by toxic 1% and baseline profiles. We also consider the tweeting pattern as a measurable trait. It reveals the longitudinal nature of a profile's posting behavior.

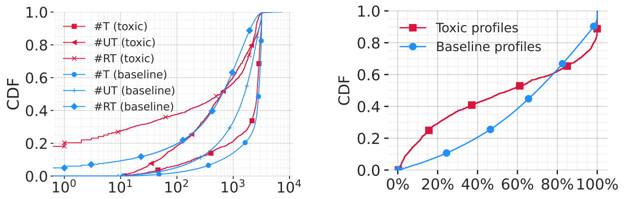
4.1 Tweet Frequency

4.1.1 Are toxic profiles prolific? In order to uncover the answer, we first note the total number of tweets of toxic 1% and baseline profiles. Unique tweets (repetitions removed) were further counted to observe whether profiles repeatedly repost the same tweet. We present a CDF with the number of total and unique tweets in Fig. 3a. From this figure, we observe that 80% of toxic profiles in our dataset post more than 1,000, as compared to 82% in baseline profiles. It is interesting to observe that half of the toxic profiles at most tweeted 500 unique tweets and half of the baseline profiles at most posted 1500 unique tweets showing that toxic profiles almost tweet equal to baseline but post fewer original tweets. We note that base 40% of our toxic and 20% of baseline profiles post retweets less than 100 times. Fig. 3b details the repetitive behavior of our profiles, it is evident that half of the toxic profiles in our dataset posted at most 55% unique tweets (no repeats) and this is true for only a quarter of baseline profiles. Interestingly, a larger proportion of profiles in the toxic set occupy lower percentages of unique tweets, before crossing over with the baseline at 77% unique tweets. At the higher percentages of unique tweets, the baseline increases gradually, whilst the bulk of the remaining toxic profiles have near 100% unique tweets, with 25% of toxic 1% profiles posting more than 95% unique tweets compared to only 15% baseline profiles and 15% of toxic 1% profiles with no repetition vs. 7%. Thus there is the occurrence of toxic profiles that repeatedly re-post the same toxic message, and profiles with individually crafted tweets containing toxicity. We report that the median toxicity of re-posted tweets is 0.47 for the toxic and 0.32 for the baseline.

4.1.2 **Takeaway:** Toxic profiles are comparable to general Twitter profiles in the total number of posted tweets but they retweet less than the baseline profiles. Notably, about 20% of toxic profiles have a much higher proportion of unique tweets than the baseline.

4.2 Tweet Pattern

4.2.1 Do toxic profiles follow any particular tweeting pattern? To capture a profile's tweeting manner, we consider the ***Time delta*** i.e., the time between sequential or consecutive tweets (time noted from tweet timestamp) for each profile in our dataset. We consider histograms of time deltas in seconds and days in Fig. 4a and 4b respectively. For both toxic and baseline groups, the most populated period of time between tweets is within a few seconds. It is interesting to observe the occurrence of periodic behavior within the baseline profiles, indicating the presence of automation, despite



(a) Tweets count. **(b) Percentage of unique tweets.**

Figure 3: Tweeting frequency of toxic and baseline profiles (§4.1). In a, T, RT, and UT, respectively, represent the number of Tweets, retweets (RT), and unique tweets per profile.

these profiles not being overly toxic. On the other hand, toxic profiles do not appear to post regularly at fixed intervals. There is a clear skew to the shorter time deltas in toxic profiles compared to the baseline.

4.2.2 Is there consistency in the temporal tweeting pattern of the toxic profiles? We now explore if the time differences between the tweets that we have observed are distributed consistently within a profile's timeline, or if profiles 'activate' briefly for bursts of activity and then go dormant.

To this end, we employ a normalized measure of burstiness to compare the tweeting behavior of toxic and baseline profiles. ***Burstiness*** [31] is a quantification of inter-event time distribution from a given event sequence, that is, the distribution of time deltas between consecutive events. The *Burstiness Score* $B = \frac{\sigma-\mu}{\sigma+\mu} = \frac{r-1}{r+1}$, where $r = \sigma/\mu$ is the coefficient of variation and σ, μ denote the standard deviation and mean of inter-event time distribution respectively. B ranges continuously between -1 and 1; regular time series (near constant inter-event times) would have scores closer to $B = -1$, $B = 0$ is a random sequence, and $B = 1$ is an extremely bursty time series (as $\sigma \longrightarrow$ inf for finite μ). It is known that burstiness is dependent on the length of the time series, and since the number of events (tweets) differs among profiles, we adopt ***Normalized Burstiness*** [31], which removes this dependency, to facilitate direct comparison among profiles. $B(n,r) = \frac{\sqrt{n+1}r-\sqrt{n-1}}{(\sqrt{n+1}-2)r+\sqrt{n-1}}$, note that $B(n,r)$ can take values greater than one and less than -1. Fig. 4c provides a probability density function or PDF of burstiness per profile. We observe that toxic profiles skew towards being more bursty with a curve peak at 0.6 as compared to 0.5 for the baseline Twitter profile. A two-sample t-test yields a p-value of 4.26×10^{-26}, considerably less than ($\alpha = 0.05$) to indicate that the distribution of Burstiness is significantly different between the toxic and baseline profile groups.

From this, we conclude that the toxic profiles are more irregular (and more bursty) in their tweeting behavior than the baseline profiles in general.

4.2.3 **Takeaways:**

- Toxic profiles can be long-lived accounts with activity gaps of 8-9 years, and they are more likely to tweet in quick succession with minimal activity intervals.
- Toxic profiles do not appear to tweet at regular fixed intervals (a sign of automation), a behavior observed in the baseline profiles.
- Generally parallels can be drawn between the temporal behavior of baseline and toxic profiles, however toxic profiles skew to favor shorter intervals between tweets and are more bursty.

5 CONTENT ANALYSIS

The nature of a profile's tweets is determined by the actual text and, also by attached auxiliary content such as URLs and hashtags. We perform a content analysis of timeline tweets with respect to each toxic or baseline profile. Specifically, we analyze the quality of the tweet's text (§5.1), toxicity level in the text (§5.2), and additional tweet attributes like URLs (§5.3), and hashtags (§5.4).

5.1 Tweet lexicon

The way a tweet is constructed tells the degree of authority of the author and potential target audience. Thus, we now analyze the text within our profiles' tweets for length, grammatical correctness, and semantic correctness.

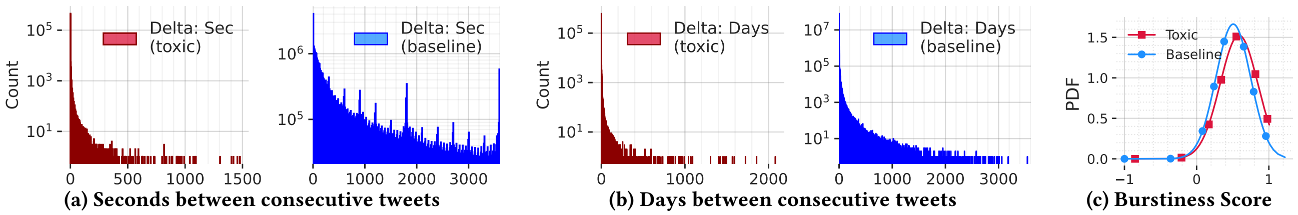
5.1.1 Do toxic profiles share verbose tweets? We question, how much of the character allowance in a tweet is utilized by our toxic profiles. To this end, we parse each tweet to extract the number of words and characters for both toxic and baseline profiles. The boxplots in Fig. 5f and 5g display the distribution of the average number of words and characters in tweets. We observe that toxic profiles post shorter tweets with fewer words than baseline Twitter profiles with an average of 11 words and 70 characters.

5.1.2 Do toxic profiles share legible and easy-to-read tweets? With the tweet text in hand, we measure the Flesch Score [16] (ease and difficulty), Linsear write scores, Automated Readability Index [49], and Lexical Diversity of toxic and baseline profiles.

The ***Flesch score*** indicates how difficult or easy it is to read the text [12], and is computed as: $206.835 - 1.015 \times (\frac{total words}{total sentences}) - 84.6 \times (\frac{total syllables}{total words})$. ***Linsear write score*** measures the length of words in the number of syllables and divides this score "r" by the number of total sentences [36]. If (r > 20, Lw = $\frac{r}{2}$) and if (r≥20, Lw = $\frac{r}{2-1}$). ***Automated Readability Index (ARI*** estimates the comprehensibility of a text corpus and is computed as (4.71×average word length)+(0.5×average sentence length)-21.43 [4]. ***Lexical diversity***, defined as the ratio of a number of unique words to the total number of words, reveals noticeable repetitions of distinct words [33]. Higher values of the ARI, Flesch scores and lexical diversity of a given text indicate increased comprehensiveness, improved readability, and range and variety of vocabulary. A given text with a high Linsear write score generally includes words with more syllables and/or is written with richer language.

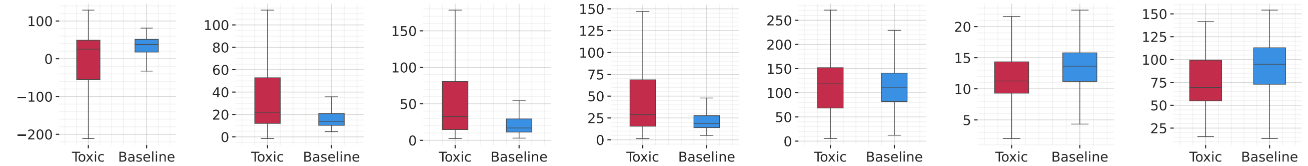
Fig. 5 provides a summary of the 3 metrics. In comparison to baseline profiles, toxic profiles share more legible and readable tweets. Toxic profiles use richer and more profound vocabulary, which as explained above is a predictable use of language and a sign of good writing style.

5.1.3 **Takeaway:** Top toxic 1% profiles share shorter tweets written in more understandable language than baseline.



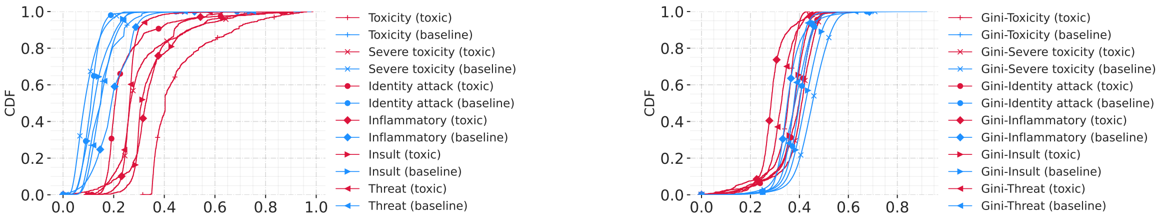
(a) Seconds between consecutive tweets (b) Days between consecutive tweets (c) Burstiness Score

Figure 4: Tweeting pattern of toxic and baseline profiles in time between sequential tweets, and burstiness in time (cf. §4.2).



(a) Flesch Kincaid reading ease score (b) Flesch Kincaid difficulty score (c) Linsear write score (d) ARI score (e) Lexical diversity score (f) # Average tweet words (g) # Average tweet char

Figure 5: Lexical analysis of toxic and baseline profiles' tweets (cf. §5.1 for details) in our dataset.



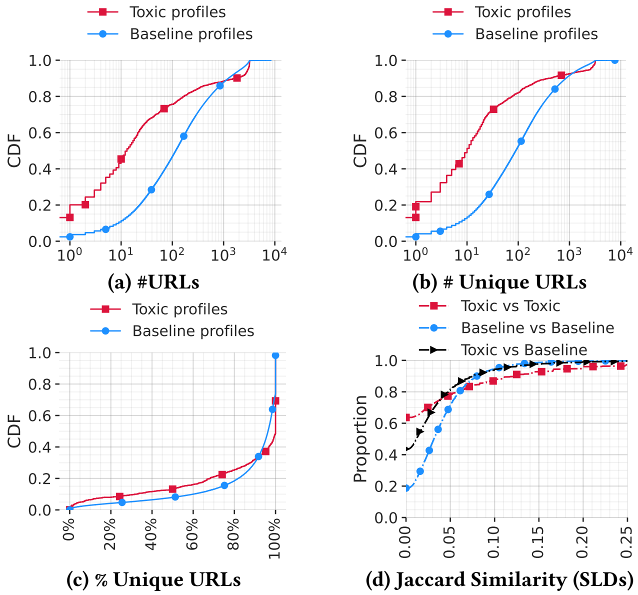
(a) Median Perspective scores (b) Gini index of Perspective scores

Figure 6: Median (a) and Gini index (b) of Perspective API scores (cf. §5.2 for details).

5.2 Tweet Toxicity

5.2.1 What type of misbehavior is common in toxic profiles? The 6 scores from the Perspective API provide granular insight into the specific types of misbehavior exhibited by a profile. We plot the median scores of Toxicity, Severe Toxicity, Identity Attack, Inflammatory, and Insult per profile as a CDF in Fig. 6a. We observe that toxic profiles on all 6 dimensions of misbehavior exceed that of general Twitter profiles. We note that beyond "Toxicity", tweets high in "Inflammatory" and "Insult" are the next most prevalent within our toxic profiles. Interesting to note that Inflammatory is comparatively less prominent (after toxicity) among toxic profiles than baseline profiles. On the other hand, the lowest score for toxic profiles is "Identity Attack", This would be consistent with Twitter policy, which states that racist tweets are not tolerated [55]. Figure 6b is a CDF of the Gini index calculated on all 6 toxicity scores to gauge the consistency of misbehavior amongst a profile's tweets. We observe that toxic profiles in comparison to baseline profiles have lower Gini scores, implying the top 1% toxic profiles exhibit misbehavior relatively consistently compared to the baseline.

5.2.2 **Takeaway:** The most common forms of toxicity among the Toxic 1% profiles are "Inflammatory" and "Insult", with "Identity Attack" the least common.



(a) #URLs (b) # Unique URLs (c) % Unique URLs (d) Jaccard Similarity (SLDs)

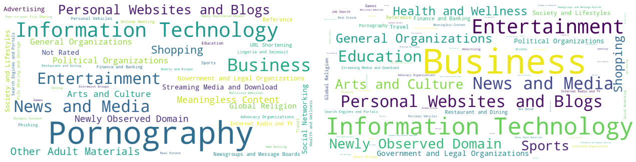
Figure 7: Profile level URL analysis (§5.3).

5.3 URLs

A shared URL is an indication that a profile seeks to point a reader to a resource external to the Twitter platform, either as a corroborative source of validation or for further reading about their tweet's subject matter. On the other hand, The repetition of a shared URL and posting URLs related to one subject (category) shows how much a profile emphasizes a topic. From our 1,380 toxic profiles and 136,620 baseline profiles, we detect and extract a total of 57,725,668 URLs and 43,916,037 unique URLs in total.

5.3.1 How frequently do toxic profiles share URLs as part of their tweet text? To answer this question, we count total number of URLs and also note the number of URLs without repetition (we inspect the full length of the URLs, extracted from the tweet's metadata and we did not rely on the shortened version used in the tweet text). Fig. 7a illustrates a CDF on the total number of URLs per profile for both groups. On the low end of the figure, it is clear that baseline profiles engage more with sharing URLs than toxic profiles. We observe that 45% of the toxic profiles shared 10 or fewer URLs in their tweets, compared to only 10% of baseline profiles. On the other extreme, approximately 10% of both profile groups are heavy URL hitters with more than 1,000 URLs in total, and 5% of toxic profiles posted 3,200 URLs compared to nearly no baseline profiles. 3,200 corresponds to the maximum number of tweets obtainable from a single profile. We note that 3.3% baseline profiles shared 6-8K URLs, these profiles shared multiple URLs per tweet in short form and on average shared 3 URLs per tweet — these profiles were predominantly news services can also comment from §4.2 that there are profiles in baseline which are persistent with regular tweeting pattern, which might indicate these are bots.

Fig. 7b presents the unique number of URLs per profile (i.e.: not counting repetitions). Half of the toxic profiles shared at most 0-10 unique URLs and half of the baseline profiles shared at most 0-95 unique URLs. It is interesting to observe that 22% of toxic profiles used a singular unique URL and 13% did not post any URLs at all. Fig. 7c shows the proportion, per profile, of URLs that are repetitions among toxic and baseline profiles. We see that around 48% of toxic profiles do not repeat URLs at all, compared to only 27% of baseline profiles (right-hand side of the plot). In contrast, looking at profiles that repeat URLs the most, the top 33% of toxic profiles (CDF values 0.0 to 0.33) have substantially more repetition than the corresponding group of baseline profiles. We note that toxic profiles in general share a lower percentage of unique URLs in their tweets. Fig. 7d shows us the Jaccard similarity of URLs amongst and between toxic and baseline groups. We observe that toxic profiles share URLs of the same nature. A large proportion of toxic profiles (63.7%) have no URL similarity with one another, in comparison to between baseline profiles (18.7%), this could be the result of the toxic profiles operating independently, or with uniquely crafted/tracking URLs. The remaining 20% of toxic profiles however do have a heightened shared URL similarity compared to the baseline, indicating the existence of coordination.



(a) Toxic profiles **(b) Baseline profiles**

Figure 8: Second level domains (SLDs) categories shown as word clouds for toxic (a) and baseline (b) profiles (cf. §5.3.2 for details).

5.3.2 What is the nature of categories in the URLs shared from tweets of toxic and baseline profiles? For the URLs that have been shared, the domain can provide an indication of the type of content linked. For example, `www.example.com`'s second level domain (SLD) is `example.com`. Proceeding forwards, "Domain" and "SLD" are interchangeable. We classify the content type of the domain with the ***FortiGuard*** classification service [26]. FortiGuard uses link crawlers, customer logs, and machine learning to categorize websites [37]. We successfully categorize 98.2% unique domains for baseline and 95.4% for toxic profiles with FortiGuard. The total number of found unique categories between the SLDs was 596 for toxic and 1,008 for the baseline. A toxic profile has on average 49 unique SLDs and a baseline profile has 487 unique SLDs in all of the tweets. We present in Figs. 8a and 8b a weighted word cloud of the SLD categories, the size of text represents the percentage of SLDs in each group assigned the category label. We observe that toxic profiles are linked to domains categorized as Pornography (examples intentionally omitted), and Information Technology (e.g. `youtube.com, SelfieSwipes.com, Kailani-Kai.com`). For the baseline profiles, the largest category is business (`huffingtonpost.com, manchester.ac.uk,gotthevote.org`)

5.3.3 Do toxic profiles share SLDs about the same subject/topic? For this, we now do a direct comparison between the nature of SLDs of toxic and baseline profiles, amongst themselves and between each other. Specifically, in Fig. 7d we compute and show the CDF of pairwise Jaccard similarity calculated on the sets of SLDs present in each toxic and baseline profile. The ***Jaccard Index*** is computed between two sets A and B as $\frac{|A \bigcup B|}{|A \bigcap B|}$, and ranges between 0 (for no common elements between the two sets) to 1 (for a perfect match or overlap). We observe that the SLDs in toxic profiles have the greatest overlap. Also, 64% of pairs of toxic 1% profiles have no similarity compared to only 18% of baseline pairs. 44% of toxic 1%-baseline pairs have disjoint sets of SLDs, indicating the presence of many SLDs present in toxic 1% tweets that are absent from baseline tweets. Overall, there is little similarity, with 95% of baseline-toxic 1% and baseline-baseline pairs, and 88% of toxic 1%-toxic 1% pairs with Jaccard similarity less than 0.1.

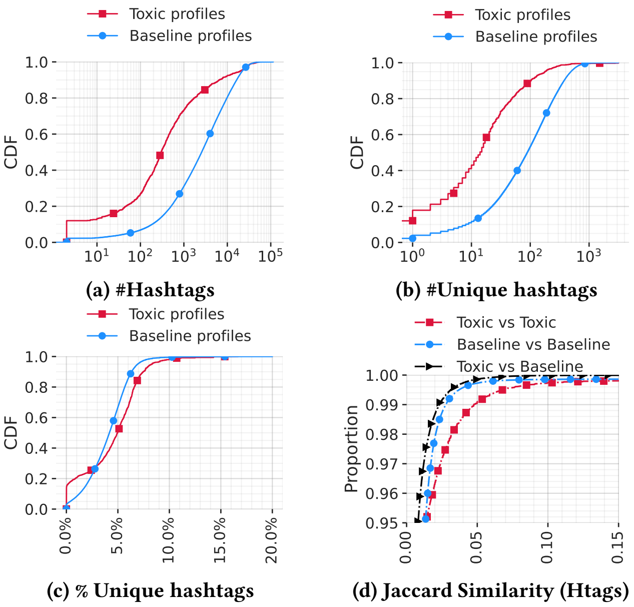
5.3.4 Takeaways:

- Toxic Profiles use fewer URLs and generally refer to unique URLs suggesting they refer less to external sources than the rest of the Twitter population.
- Of the domains linked by toxic profiles, we observed that the most popular domain categories are pornography, news, and information technology.
- Toxic 1% profiles have a larger proportion of profiles (63.7%) with no similarity between shared SLDs than baseline profiles (18.7%). This indicates high uniqueness, either from the independent operation or customized tracking domains.

5.4 Hashtags

Adding hashtags to tweets is a popular and easy way for users to convey a message to an interested audience, and to have a voice within intended communities. We compare the tendency of sharing hashtags between the toxic and baseline profiles.

5.4.1 Do toxic profiles take the help of hashtags in their tweets? Hashtags place your message within the context of a topic or community. First, with the total number of hashtags per profile (including repetitions), we discern from Fig. 9a that 50% of the toxic profiles at least shared 300 hashtags in total but 50% of baseline profiles shared many more: at least 1150 hashtags. Next, on the number of unique hashtags per profile (discounting repetitions), we see from Fig. 9b that half of the toxic profiles at most shared about 15 unique hashtags, and half of the general profiles shared at most 100 unique hashtags. As such it is evident that the toxic users are using hashtags less than the baseline. Next, Fig. 9b shows us that the toxic profiles also share fewer unique hashtags than the baseline. Fig. 9c tells us that half of the toxic profiles at most shared 50% unique hashtags as compared to 60% of baseline. Jaccard similarity Fig. 9d of the hashtags shows us the strongest similarity amongst the hashtags of toxic profiles again, pointing to their narrow focus as observed through a small number of unique categories of SLDs in §5.3.3.



(a) #Hashtags

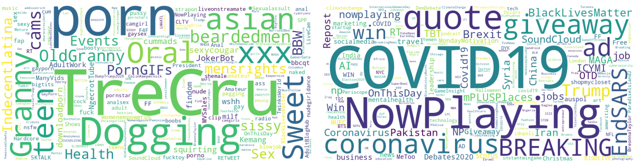
(b) #Unique hashtags

(c) % Unique hashtags

(d) Jaccard Similarity (Htags)

Figure 9: Hashtag analysis of toxic and baseline profiles (cf. §5.4).

5.4.2 Nature of hashtags shared by the top 1% toxic and baseline profiles. We now provide examples of highly occurring hashtags within the dataset. We present in Fig. 10a and 10b the weighted word clouds of all hashtags collected from tweets of toxic and baseline profiles. The largest hashtag shared by toxic profiles is 'TreCru' which is an online video game known as Treasure Cruise. The remainder of the hashtags by toxic profiles are of a very explicit nature. On the other hand, the general Twitter profiles share hashtags about diverse topics including Covid, news, and politics such as #BlackLivesMatter, #Trump, #Brexit, #EndSARS which is a protest against police brutality in Nigeria. There are also everyday benign hashtags about music #nowplaying, #SoundCloud and shopping #Giveaway, #Job. On average toxic profiles share 59 unique hashtags per profile vs 275 by a baseline profile.



(a) Toxic profiles

(b) Baseline profiles

Figure 10: Hashtag word clouds of profile groups (cf. §5.4.2).

5.4.3 Do toxic profiles share the same or similar hashtags in their tweets? By leveraging the same Jaccard similarity metric defined in §5.3.3, we inspect the overlapping hashtags used within and between the toxic and baseline profiles. While not visible in Fig. 9d, it is noted that there is zero hashtag similarity in 90.2% of toxic–toxic profiles, 85.5% between toxic-baseline profiles, and 62.9% of baseline–baseline profiles. What is visible in Fig. 9d illustrates that a small proportion of toxic-toxic profiles have a much higher overlap of hashtags and thus an aligned area of discussion.

5.4.4 **Takeaway:** Toxic profiles share fewer total and unique hashtags than baseline profiles. A larger majority of toxic profiles do not have overlapping hashtags with other toxic profiles (90.2%), compared to the baseline (62.9%). In the toxic profiles that do share hashtags with other toxic profiles, they are more aligned than the most overlapping baseline-baseline profiles.

6 PROFILE LEVEL ANALYSIS

In this section, we observe profile-level characteristics that emerge from our toxic and baseline profiles. We shall start by analyzing details directly registered with Twitter (§6.1), followed by an analysis of automation as provided by the Botometer (§6.2).

6.1 Account Metadata

'Metadata is a "data dictionary" attached to every Twitter profile providing additional insight about a profile. It is a dictionary of 17 fields including name, location, account creation date, counts of friends, followers, statuses, and favorites. It also contains information if an account that is protected and/or verified.

6.1.1 What does a toxic profile's Twitter account metadata say about them? We first inspect the proportion of profiles that are still present on Twitter. It is seen from Tab. 3 that all toxic profiles are still present on Twitter to this day, whereas 3.6% of baseline profiles no longer exist. We unfortunately cannot further determine if these accounts were deleted or banned. A majority of toxic profiles are verified; *A profile with a blue badge to show that an account is Twitter verified.* Twitter allows automation [54] and verifies the bot account [62], it, however, does not allow misconduct. Also, 95% of the Twitter profiles are protected; *A profile that does not appear in third-party search engines, i.e, Google, [53].* No toxic profiles are banned in any specific country, whereas 0.002% of baseline profiles are in multiple countries like Russia, Austria, and Belgium to name

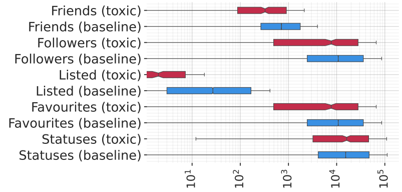


Figure 11: Toxic and baseline profiles' features (e.g. #friends, #followers, etc.) (cf. §6.1)

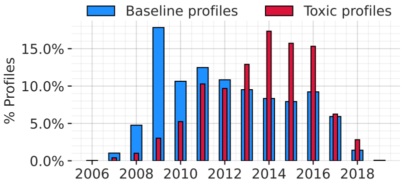


Figure 12: Creation dates of toxic and baseline profiles (cf. §6.1).

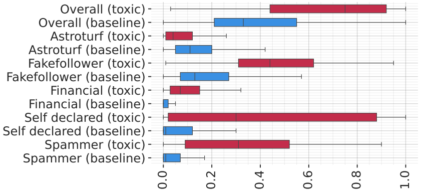


Figure 13: Botometer automation scores of toxic and baseline profiles (cf. §6.2).

a few. Additional numeric data is provided in the Twitter profile object, and the distribution of the values is represented in Fig. 11.

"Friends and Followers" Friends of a Twitter profile are other profiles followed by said profile. While followers are other profiles that follow the said profile. We observe a toxic profile has on average 500 friends, and 9,500 followers, whereas the average baseline profiles have 800 friends and 10K followers in total.

"Listed" gives the number of public lists that this profile is included within, these lists are used to collect similar accounts to strategize the timelines. As few as a couple of toxic profiles exist as part of any list, whereas a median baseline profile on average is included within 50 lists.

"Favourites" is the number of Tweets liked by a profile. We note that an average profile in both groups has liked an equally significant number of tweets (10K).

"Statuses" are the number of tweets (including retweets) created by the profile in totality, beyond the 3200 restrictions imposed during the crawling of a profile's timeline. It is interesting to note that the number of tweets posted by a median toxic and baseline profile is approximately the same if all the historical tweets are taken into account.

"Location" Another dimension is the reported location of the profile. This is a text string typically populated with a description of the profiles home town, state, and/or country. We leverage "Geopy" [42], a python library to resolve these strings into country names. The countries with the highest occurrence are presented in Tab. 4, only the top 3 countries are shown due to the quantity of found countries.

We can observe that the majority of toxic and baseline profiles are based in the US, UK, and Canada, but there exists a long tail of other countries with which profiles are associated. Interestingly, the toxic profiles are more strongly concentrated in the US (61.36%) compared to the proportion of baseline profiles in the US (29.42%). This finding is likely to differ when a different language is considered. We also note that only 0.04% of toxic and 14% of baseline profiles enabled "Geolocation" in their profile.

"Creation date" Finally, we inspect the creation date of the profiles in Fig. 12. We observe that toxic profiles skew younger than the baseline profiles. A possible explanation is an increase in the creation of toxic profiles around 2014-2016, as observed by [32]. We acknowledge that the forced and voluntary deletion of toxic profiles may also bias these numbers. Interestingly, there was a notable decline in the growth of profiles after 2016, which coincides with a plateau of active users on Twitter [50].

	Active profiles	Protected	Verified	Withheld in countries
Toxic Profiles	100%	92.74%	96.5%	None
Baseline Profiles	96.4%	95.74%	82.6%	0.002%

Table 3: Twitter profiles data (cf. §6.1).

	Top 3 locations found in profiles
Toxic Profiles	US(61.36%), Canada(9.09%), UK(9.09%), 9 others(20.04%)
Baseline Profiles	US(49.42%), UK(22.61%), Canada(7.0%), 33 others(20.09%)

Table 4: Twitter profiles location (cf. §6.1).

6.1.2 **Takeaways:**

- Toxic profiles, in general, have fewer friends, and followers and are not part of public lists in other accounts. Of the toxic profiles that have a location, 61% of them are based in the US.
- We observe an increase in the creation of toxic profiles in the US election years between 2014 and 2016, matching previous work.

6.2 Automation

6.2.1 Are toxic profiles automated bots? Automated accounts or "bots" have been observed on Twitter [61], however, Twitter permits automated accounts when they behave well according to Twitter's policy [56]. Thus, in addition to investigating the percentage of bots in our toxic and baseline groups, we also scrutinize the percentage of Twitter policy breaching "bad bots", e.g. Spammer, Fake Follower bots from *Botometer API v4* [48]. The Botometer API provides scores from five classifiers that estimate a profile's similarity to different kinds of bot behavior, including Fake Follower bots, Financial bots, self-declared bots, spammer bots, and astroturf accounts. Botometer API leverages features of a profile including the number of friends, social network structure, temporal activity (e.g. tweeting, likes, retweets), tweet language, and sentiment. Botometer provides scores in the [0,1] range, using either English or Universal (language-independent) features (we report overall universal feature scores). Botometer API defines each as:

- **Bot score**: An overall probability of profile being a bot
- **Astroturf**: A profile being one of the manually labeled political bots. These accounts systematically delete content over time.
- **Fake Follower**: An account being a bot purchased to increase follower counts.
- **Financial**: A profile used to post cashtags. Cashtags are stock symbols used with the "$" symbol. Cashtags bots promote low-value stocks by exploiting the popularity of high-value ones.
- **Self Declared**: A profile that is a bot registered with `botwiki.org`.
- **Spammer**: A profile labeled as spam bots from several datasets.

The scores for every profile are presented in Fig. 13. Each boxplot details the mean and standard deviation of all scores for both toxic and baseline profiles. The scores range from [0-1], with 0 being the most human-like and 1 as the most bot-like.

We observe that toxic profiles generally have higher overall Botometer scores with a median of 0.7, however, there still exists toxic profiles that are human-like with scores in the 1st standard deviation range <0.45 overall score. Baseline profiles skew more human-like in comparison. Astroturf (participating in politics [41]) scores are fairly low for both sets, albeit baseline profiles skewing slightly higher, this may explain that despite the most toxic these profiles are not automatically removed by Twitter, there are still long-lived spam and profane accounts on Twitter, also reported by [34]. Very few of the baseline profiles are likely to be fake followers, with a median probability of 0.14. On the other hand, just under half of the toxic profiles have a probability above 0.5 and are likely to be purchased, followers. This indicates the presence of maliciously toxic actors amplifying their toxic message through these profiles. Neither set of profiles are likely to engage in financial market updates, though there are notably more among toxic profiles. We observe that toxic profiles are widely spread on the spectrum of "self-declared" in stark contrast to baseline profiles. Spamming is not a trait of baseline profiles whereas toxic profiles have notably higher scores, and there are toxic profiles with spammer scores as high as 0.85.

6.2.2 Takeaways:

- We confirm the findings reported in [30] that the distribution of toxic profiles is less likely to be associated with politics, despite their toxic nature.
- Toxic and baseline profiles are unlikely political or financial bots. Our study validates prior work [34, 52] that the toxic profiles have a high likelihood to be spam bots and have behavior consistent with self-declared bots. 86.5% of toxic profiles are verified (§6.1)– as also reported in [52].
- Many validated toxic profiles are verified which makes their content more viral as also found by [35].

7 ETHICAL CONSIDERATIONS

Our research is non-commercial, and in line with Twitter's Terms and Conditions for research purposes, our data will not be shared with any third party for commercial purposes. We used the standard Twitter API to collect tweets from public user profiles. In all of our experiments, any result produced and shown cannot be used to re-identify, or track said users, as no user profiles are specifically named. During our experiments, we follow ethical guidelines outlined in [46]. Given our experimentation on human-produced data, we obtained IRB approval[2] from our institution.

8 DISCUSSION

Understanding the prevalence of hateful information on social media platforms is the primary driver behind this investigation of the most toxic profiles. In order to characterize levels of consistency of such behavior, to be able to make early predictions of the spread of such content, and in essence to prevent the proliferation of the most hateful content providers, it would be beneficial to study the population of profiles who produce the most toxic content.

Being able to adequately understand the behavior of the most toxic Twitter profiles is valuable in and of itself. It provides more well-informed choices about how and what to research in subsequent research investigations. It makes it possible for toxicity-reduction strategies to be designed more intelligently in many ways.

We concede that we had to work within certain constraints, such as the Perspective API limiting our work to only English tweets. Also, the Twitter API only allowed us to scrape the most recent 3200 tweets and not the entire timeline per profile. We will also like to acknowledge that because our seed data were balanced, with equal amounts of toxic and non-toxic profiles, the 1% is not a complete portrayal of the entire Twitter-sphere.

In the future, we plan to use our findings of highly toxic Twitter profiles to identify toxic Twitter profiles that are responsible for the highest toxicity in important Twitter discussions about politics, sports, and religion, among others.

9 CONCLUSION

In the past, much research has been devoted to locating toxicity spreader accounts and bots based on a few tweets; however, our work examines the timeline of Twitter accounts and takes into account the consecutive tweets posted by Twitter accounts, as well as investigates the consistent toxicity exhibited by certain profiles.

We present a profile-centric approach to survey toxicity on Twitter and characterize the most toxic profiles. Our methodology is distinct from prior works that focus on particular events or hashtags and phrases over short time windows. We focus instead on the entire profile timeline, obtaining longitudinal data that reveals the bigger picture of a profile's toxic behavior. We annotate entire timelines with the Google Perspective API. Based on the toxicity of profile tweets, we isolate the 1% most toxic profiles in our dataset and contrast their behavior with the remainder. This approach provides extra context to a profile's toxic behavior, providing new insights into toxicity on Twitter.

We find that the most toxic 1% of profiles are likely to be fake followers, indicating a level of coordinated and targeted toxic activity. They are likely to post inflammatory and profane content. Their tweets are typically textually eloquent and tend to repeat their posted content less. They are less likely to leverage auxiliary content such as URLs and hashtags in their posts.

Inspecting toxic profiles on their longitudinal data provides additional contextual insights that are otherwise missing when scrutinizing a profile on a single post. However, our approach still has limitations as obtaining this data is a challenging task. Specifically, Twitter limits the availability of the timeline to a profile's 3,200 most recent tweets. Certainly, with the full timeline, further insights could be obtained.

Findings regarding characteristics of the most toxic profiles such as inflammatory and insulting behavior, repetitive and explicit hashtags, bursty tweeting patterns, and short and well-written tweets with supporting URLs to websites and blogs in their tweets can be used to identify the most toxic Twitter accounts in a specific scenario, such as profiles discussing politics or following a specific

[2] Reference no: 520211000835379

motivational movement. Furthermore, identifying and deleting such accounts will aid in the removal of toxicity from important Twitter discussions. Our approach is not limited to Twitter and can be applied to any social media platform discussion.

In the future, we plan to further study the details of the topics discussed by toxic profiles and investigate and characterize the coordinated toxic activity as evidence for toxic influence operations present in the data.

ACKNOWLEDGMENTS

This work was partially supported by the Macquarie University Cybersecurity Hub (MQCHUB) and the EU H2020 Research and Innovation program. Hina Qayyum was supported by Macquarie University Domestic High Degree Research Scholarship Program. Nicolas Kourtellis was partially supported during this project by the EU H2020 Research and Innovation program under grant agreements No 830927 (Concordia), 871793 (Accordion), and 101021808 (Spatial). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors or originators and do not necessarily reflect the views of the MQCHUB or of the EU H2020 Research and Innovation program.

REFERENCES

[1] Meysam Alizadeh, Jacob N Shapiro, Cody Buntain, and Joshua A Tucker. 2020. Content-based features predict social media influence operations. *Science advances* 6, 30 (2020).

[2] Raghad Alshalan, Hend Al-Khalifa, Duaa Alsaeed, Heyam Al-Baity, Shahad Alshalan, et al. 2020. Detection of Hate Speech in COVID-19–Related Tweets in the Arab Region: Deep Learning and Topic Modeling Approach. *Journal Medical Internet Research* 22, 12 (8 Dec 2020).

[3] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *ICANLIS*.

[4] ari. 2019. The Automated Readability Index.

[5] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *WoSE*.

[6] Matthew C Benigni, Kenneth Joseph, and Kathleen M Carley. 2017. Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PloS one* 12, 12 (2017).

[7] Carlos Arcila Calderón, Gonzalo de la Vega, and David Blanco Herrero. 2020. Topic modeling and characterization of hate speech against immigrants on Twitter around the emergence of a far-right party in Spain. *Social Sciences* 9, 11 (2020), 188.

[8] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *ACM Web Science*.

[9] Isobelle Clarke and Jack Grieve. 2017. Dimensions of abusive language on Twitter. In *ALO*.

[10] Shaniece Criss, Eli Michaels, Kamra Solomon, Amani Allen, and Thu Nguyen. 2020. Twitter Fingers and Echo Chambers: Exploring Expressions and Experiences of Online Racism Using Twitter. *Journal of Racial and Ethnic Health Disparities* 8 (10 2020).

[11] Ashwin Geet d'Sa, Irina Illina, and Dominique Fohr. 2020. Bert and fasttext embeddings for automatic detection of toxic speech. In *OCTA*.

[12] Derar Eleyan, Abed Othman, and Amna Eleyan. 2020. Enhancing Software Comments Readability Using Flesch Reading Ease Score. *Information* 11, 9 (Sep 2020), 430. https://doi.org/10.3390/info11090430

[13] Norwegian Defence Research Establishment. 2019. *Social network centric warfare - understanding influence operations in social media*. Technical Report. FFI - Forsvarets forskningsinstitutt - Norwegian Defence Research Establishment.

[14] Miriam Fernandez, Moizzah Asif, and Harith Alani. 2018. Understanding the roots of radicalisation on twitter. In *ACM Web Science*.

[15] Emilio Ferrara. 2020. Bots, elections, and social media: a brief overview. *Disinformation, Misinformation, and Fake News in Social Media* (2020), 95–114.

[16] Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221.

[17] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM CSUR* 51, 4 (2018).

[18] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

[19] Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. 2018. Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach. *arXiv preprint arXiv:1809.08651* (2018).

[20] Corrado Gini. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]*. Tipogr. di P. Cuppini.

[21] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2019. Exploring Hate Speech Detection in Multimodal Publications.

[22] Google. 2021. Perspective API - Using machine learning to reduce toxicity online. https://www.perspectiveapi.com/.

[23] Lara Grimminger and Roman Klinger. 2021. Hate Towards the Political Opponent: A Twitter Corpus Study of the 2020 US Elections on the Basis of Offensive Speech and Stance Detection.

[24] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2020. Racism is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis.

[25] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. arXiv:1702.08138

[26] Fortinet Inc. 2021. Web Filter Categories. https://fortiguard.com/webfilter/categories.

[27] Edwin Jain, Stephan Brown, Jeffery Chen, Erin Neaton, Mohammad Baidas, Ziqian Dong, Huanying Gu, and Nabi Sertac Artan. 2018. Adversarial Text Generation for Google's Perspective API. In *CSCI*.

[28] Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*. 7–16.

[29] Kaggle. 2020. Hatred on Twitter During MeToo Movement - Kaggle. https://www.kaggle.com/rahulgoel1106/hatred-on-twitter-during-metoo-movement.

[30] Tobias R. Keller and Ulrike Klinger. 2019. Social Bots in Election Campaigns: Theoretical, Empirical, and Methodological Implications. *Political Communication* 36, 1 (2019), 171–189.

[31] Eun-Kyeong Kim and Hang-Hyun Jo. 2016. Measuring burstiness for finite event sequences. *Physical Review E* 94, 3 (2016), 032311.

[32] Bence Kollanyi, Philip N Howard, and Samuel C Woolley. 2016. Bots and automation over Twitter during the first US presidential debate. *Comprop data memo* 1 (2016), 1–4.

[33] Kristopher Kyle, Scott A. Crossley, and Scott Jarvis. 2021. Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. *Language Assessment Quarterly* 18, 2 (2021), 154–170. https://doi.org/10.1080/15434303.2020.1844205 arXiv:https://doi.org/10.1080/15434303.2020.1844205

[34] Po-Ching Lin and Po-Min Huang. 2013. A study of effective features for detecting long-surviving Twitter spam accounts. In *2013 15th International Conference on Advanced Communications Technology (ICACT)*. 841–846.

[35] Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712* (2018).

[36] Bryan C McCannon. 2019. Readability and research impact. *Economics Letters* 180 (2019), 76–79.

[37] Arian Akhavan Niaki, Nguyen Phong Hoang, Phillipa Gill, Amir Houmansadr, et al. 2020. Triplet Censors: Demystifying Great {Firewall's}{DNS} Censorship Behavior. In *FOCI*.

[38] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2021. Uncovering Coordinated Networks on Social Media: Methods and Case Studies.

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

[40] Pew Research Center. 2018. Bots in the Twittersphere. https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/.

[41] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2011. Detecting and tracking political abuse in social media. In *Proceedings of the International AAAI Conference on Web and social media*, Vol. 5. 297–304.

[42] GeoPy Repository. 2022. GeoPy Documentation. https://geopy.readthedocs.io/en/stable/.

[43] Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth international AAAI conference on web and social media*.

[44] Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, and Wagner Meira Jr au2. 2018. "Like Sheep Among Wolves": Characterizing Hateful Users on Twitter. In *MWSDM*.

[45] Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgílio A. F. Almeida, and Wagner Meira. 2018. Characterizing and Detecting Hateful Users on Twitter.

[46] Caitlin M Rivers and Bryan L Lewis. 2014. Ethical research standards in a world of big data. *F1000Research* 3 (2014).

[47] Karla Dhungana Sainju, Niti Mishra, Akosua Kuffour, and Lisa Young. 2021. Bullying discourse on Twitter: An examination of bully-related tweets using supervised machine learning. *Computers in human behavior* 120 (2021), 106735.

[48] Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2020. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 2725–2732.

[49] RJ Senter and Edgar A Smith. 1967. *Automated readability index*. Technical Report. AMRL.

[50] Statista. 2019. Number of monthly active Twitter users worldwide.

[51] Hannah Stevens, Muhammad Ehab Rasul, and Yoo Jung Oh. 2022. Emotions and Incivility in Vaccine Mandate Discourse: Natural Language Processing Insights. *JMIR Infodemiology* 2, 2 (13 Sep 2022), e37635. https://doi.org/10.2196/37635

[52] Mikael Thalen. 2022. Twitter verified a number of bot accounts. https://www.dailydot.com/debug/twitter-verified-bot-accounts/.

[53] Twitter. 2021. Twitter API Documentation. https://developer.twitter.com/en/docs/twitter-api.

[54] Twitter. 2023. Automation Rules.

[55] Twitter. 2023. hateful conduct policy.

[56] Twitter.com. 2022. Twitter automation rules. https://help.twitter.com/en/rules-and-policies/twitter-automation.

[57] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*. 138–142.

[58] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.

[59] 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 (2018), 13825–13835.

[60] Yifang Wei and Lisa Singh. 2018. *Detecting Users Who Share Extremist Content on Twitter*. Springer International Publishing, Cham, 351–368.

[61] Jason wise. 2022. Twitter Bot Accounts: How Many Bots Are On Twitter in 2022? https://earthweb.com/how-many-bots-are-on-twitter/.

[62] Dale John Wong. 2023. Twitter now has verified bot accounts to help you follow the good ones.

[63] Ziqi Zhang and Lei Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web* 10, 5 (2019), 925–945.