





# Vitriol on Social Media: Curation and Investigation

Xing Zhao<sup>(✉)</sup>  and James Caverlee 

Texas A&M University, College Station, TX 77840, USA  
{xingzhao,caverlee}@tamu.edu

**Abstract.** Our online discourse is too often characterized by vitriol. Distinct from hate speech and bullying, vitriol corresponds to a persistent coarsening of the discourse that leads to a cumulative corrosive effect. And yet, vitriol itself is challenging to formally define and study in a rigorous way. Toward bridging this gap, we present in this paper the design of a vitriol curation framework that serves as an initial step toward extracting vitriolic posts from social media with high confidence. We investigate a large collection of vitriolic posts sampled from Twitter, where we examine both user-level and post-level characteristics of vitriol. We find key characteristics of vitriol that can distinguish it from non-vitriol, including aspects of popularity, network, sentiment, language structure, and content.

**Keywords:** Vitriol · Social media · Personal attacks  
Abusive language · Text classification

## 1 Introduction

The widespread adoption of social media has led to positive developments like community formation, information discovery, image and video sharing, and access to allies and audiences for traditionally disenfranchised groups. Alas, social media platforms have also become rife with undesired effects, including bullying [18, 20], personal attacks [22, 27], hate speech [2, 4, 5, 11, 21, 25, 26], arguments [3, 13], and trolling [7, 8, 15].

Indeed, we can view many of these examples as parts of a broad class of online discourse that is *vitriolic*. Vitriol corresponds to a persistent coarsening of the discourse that leads to a caustic, corrosive, and negative experience in our online interactions. For example, consider the following two tweets:

- *So my damn property and school taxes go up to pay for the damn illegals. Your doing crap for middle class*
- *Then if you want to switch back to produce the soil ruined. Nice going, moron*

We argue that these tweets are vitriolic: they are caustic and corrosive. However, this vitriol does not meet the requirements of hate speech, bullying, trolling,

or other anti-social activities. For example, hate speech typically is an attack on a target’s race, religion, ethnic origin, sexual orientation, and so on. Angry and resentful posts such as these two examples need not contain such attacks to be vitriolic. Similarly, bullying corresponds to a person using strength or influence to harm or intimidate those who are weaker, often including persistent and targeted behaviors to induce harm in another. Vitriol need not rise to the level of bullying, and vitriol is often initiated by ordinary people (weaker) and targeted at well-known users (stronger). While considerable previous work has focused on uncovering evidence of bullying, hate speech, and trolling, there is a research gap in curating and investigating such vitriol that creates an unwelcoming, corrosive online experience.

Hence in this paper, our goal is to begin an investigation into vitriol on social media, including: How can we define vitriol? How can we operationalize such a definition for extracting evidence of vitriol? Can we detect vitriol at scale? And how does vitriol differ from posts that just happen to include profanity? Many previous methods for extracting abusive language have focused on content-based features, and yet, some profanity can be well-meant or just joking. For example, Table 1 shows examples of what we consider to be vitriol versus profanity-laden non-vitriol posts sampled from Twitter. We find that these false positive samples are often meant as banter between friends. This observation illustrates that the detection of vitriol is challenging if we just use profanity filters or topical analysis, which are widely used in previous works.

**Table 1.** Example vitriolic tweets vs. non-vitriolic tweets

Vitriol	Non-Vitriol
<i>@HouseGOP So my damn property and school taxes go up to pay for the damn illegals. Your doing crap for middle class</i>	<i>@josel767 @rosariolopezn And remember kids, you’ll always be shit, but you wanna be the best shit to have ever been created [emoji]</i>
<i>@WolfForPA Then if you want to switch back to produce the soil ruined. Nice going, moron</i>	<i>@essjain bitch if you wasn’t my mfn friend</i>
<i>@RCorbettMEP Your peddling fear mongering bull shit. You don’t mention Fracki that’s a serious ecological risk. You arrogantly assume the ..</i>	<i>@Applied_press Weak as hell. Can you believe I’m ready to come back to Charleston</i>
<i>@DMVBlackLives You idiots are responsible for this shit</i>	<i>@Stonekettle Is this fucking fuck fucking serious?</i>
<i>@WayneDupreeShow Can you spell traitor</i>	<i>@EthanDolan @BryantEslava Your so cute wtf</i>
	<i>@Rival_Laxno @HypeWicked @VillainGoofys Holy shit theirs been hella beef today</i>

In the rest of the paper, (i) we design a curation framework for identifying vitriol from ordinary profanity-laden language online and build a vitriolic dataset; (ii) we analyze vitriolic users and their language, comparing with other users and tweets on Twitter; and (iii) we propose a suite of features to build classifiers to distinguish vitriolic tweets from other tweets, distinguish vitriolic users from random users, and ultimately detect vitriol from the wider social media space. We find that vitriolic posts vary in both user-level and post-level features compared with other tweets, with key differences in popularity, network, sentiment, language structure, and content characteristics that could provide a basis for continued exploration of vitriol in social media.

## 2 Related Work

Many existing studies focus on hate speech. For example, Banks examined the complexities of regulating hate speech on the Internet through legal and technological frameworks [2]. Warner *et al.* further presented an approach to detecting hate speech in online text, and contributed a mechanism for detecting some commonly used methods of evading common “dirty word” filters [26]. Burnap *et al.* developed a supervised machine learning classifier for hateful and antagonistic content on social media, which can assist policy and decision makers in monitoring the public reaction to large-scale events [4]. To detect hate speech incorporating context information, Gao *et al.* presented a logistic regression model with context feature, and a neural network model with learning components for context [11]. Chandrasekharan *et al.* studied the 2015 ban of two hate communities on Reddit in terms of its effect on both participating users and affected subreddits [5]. Clarke *et al.* used a new categorical form of multidimensional register analysis to identify the main dimensions of functional linguistic variation in a corpus of abusive language, specifically consisting of racist and sexist Tweets [9]. While certainly hate speech is a kind of online vitriol, we seek to find corrosive vitriolic posts even in the absence of specific targeting of race, religion, and other features of hate speech.

Trolling is another antisocial behavior on social media. Hardaker *et al.* defined “troll” as a person that engages in negative online behavior [15]. Cheng *et al.* characterized trolling behavior in three large online discussion communities – CNN, IGN, and Breitbart – by analyzing their suspended users [8]. In their latest study, they analyzed the causes of trolling behavior on discussions, and their predictive model indicates trolling can be better explained by incorporating mood and discussion context [7]. Many of these anti-social phenomena – and specifically vitriolic posts in news comments – have been attributed to granting “*someone anonymity and he or she is apt to behave poorly, namely with malevolence in their comments*” [24].

In a related, but potentially less harmful direction, sarcasm is a form of speech act in which the speakers convey their message in an implicit way [10]. Davidov *et al.* experimented with semi-supervised sarcasm identification on Twitter and Amazon dataset [10], and Bamman improved the detection performance

by including extra-linguistic information from the context of an utterance [1]. González-Ibáñez *et al.* provided a method for constructing a corpus of sarcastic Twitter messages in which determination of the sarcasm of each message has been made by its author, and investigated the impact of lexical and pragmatic factors on machine learning effectiveness for identifying sarcastic utterances [13].

### 3 Curating Vitriol

In this section, we propose a vitriol curation framework for sampling vitriol from social media, before turning in the following section to an investigation of the factors impacting what is and is not considered vitriol. Since vitriol may come in many forms, our key intuition is to focus on posts that demonstrate three observable characteristics:

- *Personal*: the post should target another user, rather than just “shouting to the wind”;
- *Context-free*: the post should ignore the substance of what the target user cares about (the context); and
- *Unilateral*: the post should be one-way from a vitriolic user to a target user, and not a back-and-forth argument.

While not representative of all forms of online vitriol, these three characteristics do allow us to operationalize our definition of vitriol for sampling evidence at scale from social media. And while vitriol exists on every social media and content-based platform – including Facebook, Twitter, Reddit, and commenting systems on news websites – we focus on Twitter since Twitter collects many user-level features, such as the popularity and social relationships, and we can track a specific user using the *user\_id* to analyze the user’s history of posts.

#### 3.1 Raw Data Collection

First, to collect a sample of *potentially* vitriolic posts (English language only) from Twitter, we begin by sampling based on a keyword list derived from Liu *et al.* ’s Negative Opinion Word List [19], augmented with a set of frequently used abusive words on Twitter and their synonyms.<sup>1</sup> Some of these keywords are shown in Table 2.

**Table 2.** Vitriolic wordbag

bullshit	lie	fake	fuck	shit	ass	stupid	spew
idiot	liar	crap	asshole	moron	damn	hell	corrupt
fool	shutup	horseshit	bastard	bitch	traitor	fraud	...

<sup>1</sup> All data, annotated samples, code, and experiments are available at <https://github.com/xing-zhao/Vitriol-on-Social-Media>.

In total, we sampled more than 3 million *potential vitriolic tweets* (denoted  $P_{VT}$ ) sent by 1.7 million *potential vitriolic users* (denoted  $P_{VU}$ ) over the period June 30<sup>th</sup> 2017 to September 14<sup>th</sup> 2017. We additionally sampled the target of these posts (recall that our definition requires a post to be sent in response to another post). We call these original targeted posts the set  $P_{ST}$  and the users of these original targeted posts as  $P_{SU}$ . This raw dataset is summarized in Table 3.

### 3.2 Refining the Sample

Of course, using these keywords alone to select vitriolic tweets is insufficient – for example, many of these selected keywords can be used as jokes or banter between friends. For reducing such false positives, we further refine the sample down to a curated set of vitriolic tweets  $V_T$  sent by vitriolic users  $V_U$ . Our goal here is to focus on precision (identifying only real vitriol) rather than on recall (finding all possible vitriol, but at the risk of many false positives). We adopt the following curation strategies:

**Table 3.** Raw dataset statistics

Set	Size	Set	Size
$P_{VU}$	1,720,281	$P_{VT}$	3,336,477
$P_{SU}$	1,374,420	$P_{ST}$	2,883,092

**Direct Replies Only.** *The vitriolic tweets must be the first layer replies of an originally generated post.* We aim to find those vitriolic users that directly targeted the person being replied to. However, on Twitter, there are many formats of tweets, such as retweets and replies. This diversity can bring noise into our curation method. For instance, user  $A$  could reply to user  $B$ 's tweet which is retweeted from user  $C$ . In such a case, it is hard to identify if  $A$ 's target is  $B$  or  $C$ . To maximize the likelihood that the attack target from a reply tweet is the person who is replied to, we restrict the format of the replied tweet to be the original tweet, and restrict the format of reply tweets to be at the first layer, which means it directly replies to the original poster rather than other repliers or re-tweeters.

**Avoid Copy-Paste Tweets.** *The replies posted by a vitriolic user cannot be identical to each other.* Through our manual investigation, we found that some users repeatedly send reply tweets with identical content to different users, in essence spamming out the same (or similar) content to a wide audience. We assume such behaviors can be dealt with using traditional spam detection methods and do not reflect vitriol sent by real users.

**Avoid Copy-Paste Tweets.** *The replies posted by a vitriolic user cannot be identical to each other.* Through our manual investigation, we found that some users repeatedly send reply tweets with identical content to different users, in essence spamming out the same (or similar) content to a wide audience. We assume such behaviors can be dealt with using traditional spam detection methods and do not reflect vitriol sent by real users.

**Focus on “Real” Active Users.** *Users must have sent at least some minimum number of tweets, but not too many repeated tweets.* There is wide evidence of paid posters and bots that frequently post similar comments or articles on different online communities and Websites for hidden purposes, e.g., to influence the opinion of other people towards certain social events or particular markets [6]. Since our focus is on the behavior of real users and not bots or other spam-like accounts, we set an upper bound of to avoid these accounts. We additionally set a lower bound of tweeting frequency to capture users who are actually active and

not isolated users with only a few tweets. We do experiments for maximumly avoiding pre-annotated isolated or spam-like accounts using different settings of the lower and upper bound (please refer to the project website for further details: <https://github.com/xing-zhao/Vitriol-on-Social-Media>). In practice, we ultimately consider users with a statuses count  $\geq 200$ , and a total number of tweets during our collection time of between 25 and 200.

**Unilateral Relationship.** *The relationship between the original poster and vitriolic replier should be unilateral.* To avoid bullying-specific tweets (which have been studied in previous works) and to focus on vitriol originating from a power imbalance (from “weaker” to “stronger” users), we consider the relative popularity of both a vitriolic user and the targeted user. We use both *# of followers* of the user and *# of retweet times* of a tweet to represent a person’s popularity. We do experiments for maximizing the number of unilateral relationship using pre-annotated dataset (see <https://github.com/xing-zhao/Vitriol-on-Social-Media>). Ultimately, we keep only users who have *# of followers*  $< 500$  but who target users with *# of followers*  $> 5000$ .

### 3.3 The Curated Vitriol Dataset

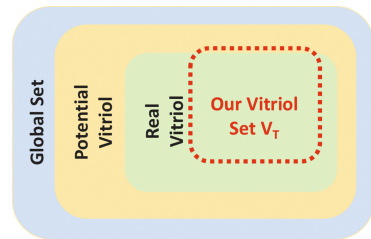
With these selection strategies, we refine our raw dataset to arrive at the curated vitriol dataset shown in Table 4. We identify 14,001 Vitriolic Tweets ( $V_T$ ) sent by 926 Vitriolic Users ( $V_U$ ). Furthermore, we collect all the users who are targeted by these vitriolic users during our observation, denoted as  $S_U$ , and their targeted tweets set  $S_T$ .

### 3.4 Validation

To validate the quality of our curation framework, we solicited three annotators to manually label a set of 500 randomly selected tweets from the sample of vitriolic tweets  $V_T$ . We took the majority vote as the ground truth for each tweet (see <https://github.com/xing-zhao/Vitriol-on-Social-Media> for details). After annotation, we find that 477 of the 500 tweets are considered vitriol, indicating a precision of 95.4%. Hence, while our curation strategies are aggressive in terms of focusing on particular kinds of vitriol (meaning that there are certainly many forms of vitriol that this initial framework misses), we see that the output is of fairly high quality. See Fig. 1 for a summary of the scope of this investigation. In our continuing work, we are interested to vary these curation strategies to better explore the trade-offs between precision and recall.

**Table 4.** Vitriol dataset statistics

	Tweets	Users
Vitriolic	14,001	926
Targeted	11,938	3,188



**Fig. 1.** The scope of our investigation.

## 4 Exploratory Analysis

In this section, we explore both tweet-centric and user-centric differences between vitriol and others. For comparison, we consider an equally-sized sample of general English tweets not contained in  $P_{VT}$ , defined as *Non-Vitriolic Tweets*; and define their posters as *Non-Vitriolic Users*. Last but not least, we present exploratory analysis of the people who were most targeted by vitriolic users.

### 4.1 Mood-Based Features

We begin by exploring the mood-based features of the tweets themselves. Since vitriol is fundamentally caustic, corrosive, and negative, we explore here the *emotional* attributes of the tweets as well as the underlying *social tendencies* of the users through an application of the IBM Watson Tone Analyzer [16] to the content of each tweet.

**Emotional Attributes.** We begin by considering five kinds of emotional attributes – anger, disgust, fear, joy, and sadness. Figure 2 shows the score for all vitriolic tweets versus a random sample of non-vitriolic tweets. The y-axis captures a likelihood score for each emotion; higher scores indicate higher degrees of each emotion. Overall, we see that vitriolic tweets score is high in anger, disgust, and sadness relative to non-vitriolic tweets, while scoring lower in joy. The original keywords that powered our curation method (see Table 2) overwhelmingly drive the *anger* score, but have little or no impact on the other scores. This suggests that even for vitriol not containing one of these original keywords, there may be clear patterns of disgust and sadness that can be used to identify additional vitriolic tweets.

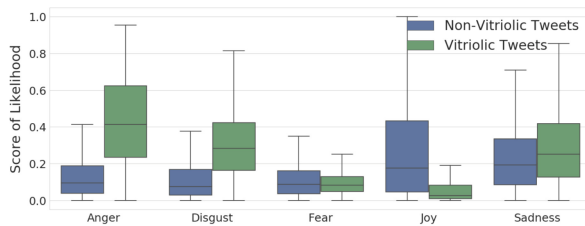


Fig. 2. Emotions of vitriolic and non-vitriolic tweets

**Social Tendencies.** We pair the emotional attributes of the tweets with five additional features that capture the social tendencies of the underlying user based on their language use – openness, conscientiousness, extroversion, agreeableness, and emotional range.

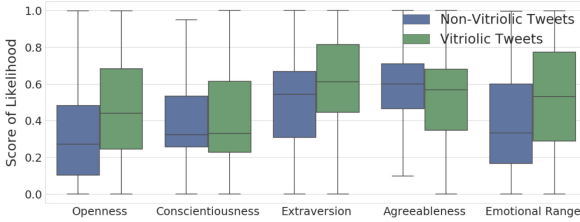


Fig. 3. Social tendencies of vitriolic and non-vitriolic tweets

Figure 3 shows the score for all vitriolic tweets versus a random sample of non-vitriolic tweets. The y-axis captures a likelihood score for each social tendency; higher scores indicate higher degrees of each tendency. Overall, we see obvious differences. Vitriolic tweets are more likely to demonstrate openness, extroversion and emotional range, and they are less likely to display agreeableness in comparison with non-vitriolic tweets.

### 4.2 User-Based Features

In addition to these content-based properties of vitriol, we also consider the popularity and activity properties of the users themselves.

**Popularity.** We measure a user’s popularity from two aspects – their followers count and friends count. Both counts indicate whether a user has a certain level of being paid attention to by other users. Figure 4 shows the comparison between vitriolic users’ and non-vitriolic users’ popularities. In summary, both counts of vitriolic users are lower than non-vitriolic users, especially in term of follower count. This shows that vitriolic users are much less popular than average; that is, vitriolic users are recognized and accepted by a group of a smaller size.

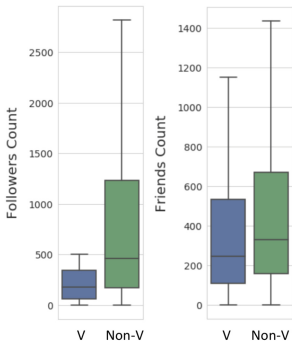


Fig. 4. Followers and friends count of each set of users.

**Activities.** To analyze a user’s degree of activity, we examine their statuses count and social age. The statuses count in Twitter is the number of tweets (including retweets) issued by a user, which can be intuitively regarded as an indicator of user’s degree of activity. Instead of using a user’s actual age, which is not publicly accessed to in Twitter, we choose another determinant, social age, for indicating a user’s activity durations on social media. This determinant is the lifespan of the account (from creation onward), calculated by subtracting the creation date of a user’s account from the creation date of

the latest tweet in our dataset. Social age could become an indicator of activity



degree since on average, longer social age means the one have had more experiences on a certain social media platform, furthermore, the user was more active on this platform. Both activity features, statuses count and social age, are used in the comparison of user’s degree of activity.

Through comparing the activity of vitriolic users and non-vitriolic users, we find that although the average of statuses count of vitriolic users (average about 7464) is slightly lower than non-vitriolic users (average about 9231), however, it is obvious that the social age of vitriolic users (average about 497 days) is much lower than non-vitriolic users (average about 1472 days). These observations indicate that vitriolic users are less active in social media than others.

### 4.3 Who Is Most Targeted by Vitriol?

But who are these users that are most targeted with vitriol? Recall that our operational definition of vitriol focuses on users who ignore the substance of a target user’s post (that is, they do not engage on the merits, but rather rely on caustic or corrosive language). Here, we consider the users who have been targeted in dataset  $S_U$  where we see that vitriolic users, while often ignoring the substance of a post, do care about the social identity of target users.

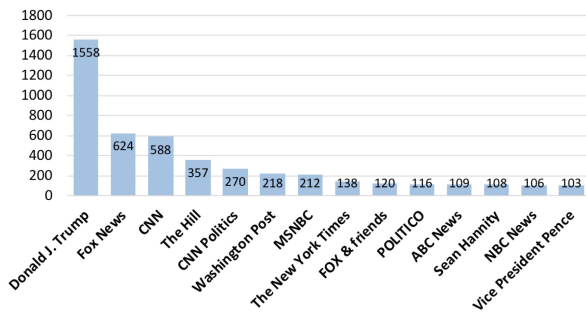
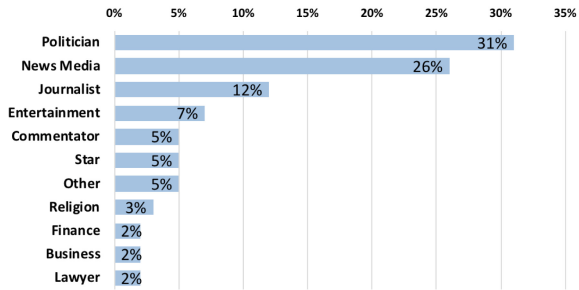


Fig. 5. The most targeted users

Figure 5 shows the top users who have been targeted by vitriolic users more than 100 times in our dataset. We find that (perhaps, unsurprisingly) most of these users are composed of politicians and news media accounts. To further study the categories of the users who were targeted the most, we manually labeled the categories of the top-100 most targeted users. The percentage of each category is presented in Fig. 6, which shows the largest slice is “politician” (31%), followed by “news media” (26%) and “journalist” (12%). This makes sense considering the divisiveness of politics, news and opinions among many people.

## 4.4 Summary

In this section, we present several data-driven analysis from user-centric and tweet-centric perspectives. We found that there are obvious differences between vitriolic users and non-vitriolic users in terms of user popularity and the degree of activity, as well as the obvious differences between their tweets in terms of emotions and social tendencies. Finally, our exploratory analysis of people who were targeted by vitriolic users shows clear patterns in their composition.



**Fig. 6.** Categories of the Top-100 most targeted users.

## 5 Vitriol Detection

In this section, we explore the potential of using features of vitriol – including from the perspectives of language patterns, communication sentiment, content relevance and latent topics – to distinguish vitriolic tweets from non-vitriolic tweets, and further distinguish vitriolic users from other users. Such models could power vitriol detection beyond our curated collection.

### 5.1 Features

To build our classifier to distinguish vitriolic tweets from others, we adopt four categories of features which can help us to characterize vitriol:

**Language Patterns (LP).** Through our manual annotation, we observed that vitriolic users use fewer at-mention markers (@), hashtags (#) and emoticons (i.e. “:-)”, “:b”), but more adjectives and strong punctuations (i.e. “?!”) than other tweets. Thus, we hypothesize that vitriolic users have certain patterns in their writing habits. To rigorously verify this hypothesis, we use Part-of-Speech Tagging [12] to analyze the language patterns of vitriolic tweets.

**Communication Sentiment (CS).** Unlike normal tweets, we have seen that vitriolic tweets include a certain set of emotions to fully express and vent writers’ feelings. To fully analyze the sentiment of language style, we apply the IBM Watson Tone Analyzer [16] and Google Sentiment Analyzer [14]. *Emotion*, a subset of these features, shows the likelihood of a writer being perceived as angry, disgust, fear, joy and sadness. Another subset of features, *Language Style*, shows the writer’s reasoning and analytical attitude about things, degree of certainty and inhibition. And the feature set *Social tendency* will help us to prove our

hypotheses that this kind of people have specific social properties in terms of openness, conscientiousness, and do on. The Google Sentiment Analysis inspects the given text and identifies the prevailing emotional opinion within the text, especially to determine a writer’s attitude as positive, negative, or neutral.

**Content Relevance (CR).** Our earlier feature category for language patterns focused on the part-of-speech patterns used on vitriol, such as nouns, adjectives, determiners, and so on. Here we consider the actual content of the tweets themselves; perhaps vitriolic tweets re-use certain phrases. Specifically, we adopt Doc2Vec [23] for learning a distributed representation [17] using hierarchical softmax. We consider each tweet as a document; Doc2Vec outputs a vector (of size 100) for each tweet such that “similar” tweets should be nearby in the dense Doc2Vec vector, where similarity here captures word order and deeper semantic similarity than in traditional bag-of-words models.

**Latent Topics (LT).** As we observed in Sect. 4, most of the people who are targeted by vitriolic users belong to categories such as famous politicians and news media accounts. We hypothesize that these vitriolic tweets are also topic-related. To fully analyze the latent topic of vitriolic tweets, we apply the LDA model [23], which allows both LDA model estimation from a training corpus and inference of topic distribution on new, unseen documents. We set the hyperparameter  $\#topics = 10$  so that the model can return a vector of likelihoods of each topic a tweet belongs to.

Table 5 shows the details of top visible features listed above, and the Fisher score of every specific feature used in different classifiers. Since features on Content Relevance and Latent Topics sets are not directly interpretable, they are not shown on Table 5. In term of language patterns, the results fit our expectation and verify our hypothesis that common nouns, adjectives, and punctuations are used in vitriolic tweets more than other tweets. This result suggests that vitriolic users do have certain patterns compared with other tweets. On the other hand, in terms of communication sentiment, anger gets the highest fisher score, which is unsurprising since our selection strategy focuses on anger words. However, we also see that disgust and joy play an important role to classify vitriol and non-vitriol.

## 5.2 Classification: Vitriol vs. Non-Vitriol

To train the classification model for vitriol vs. non-vitriol, we use all tweets in our vitriolic tweets set  $V_T$  ( $size = 14001$ ) as the positive samples, and equal-size of non-vitriolic tweets in  $R_T$  as the negative samples. We build the classifier with four different categories of features: Language Patterns (LP), Communication Sentiment (CS), Content Relevance (CR), and Latent Topics (LT), to test which features work better. We create four more feature sets by combining these four basic categories in different ways: Language Patterns + Communication Sentiment (LP-CS), Language Patterns + Content Relevance + Latent Topics (LP-CR-LT), and all features together (ALL). Note that we exclude Communication Sentiment in LP-CR-LT since our strategy of selecting the potential

**Table 5.** Part of Selected features for classification

Feature Set	Source	Top Features (Fisher Score on V vs Non-V)
Language Patterns (# = 25)	Part-of-Speech Tagging from CMU	N-common noun (0.0863)
		A-adjective (0.0682)
		P-pre- or postposition (0.0501)
		,-punctuation (0.0465)
		D-determiner (0.0351)
		V-verb (0.0204)
		U-URL or email address (0.0169)
		\$-numeral (0.0152)
		O-pronoun (0.0111)
...		
Communication Sentiment (# = 14)	IBM Watson & Google Tone Analyzers	Anger (0.4027)
		Google Sentiment (0.3721)
		Disgust (0.2928)
		Joy (0.1696)
		Openness (0.0756)
		Emotional_Range (0.0577)
		Extroversion (0.0366)
		Sadness (0.0180)
		Agreeableness (0.0142)
...		

vitriolic tweets relies on some profanity words as seed keywords. Hence, we want to evaluate the classifier when we leave out the influence of these profanity words.

We experiment with four classification algorithms: Logistic Regression, Support Vector Machine, Random Forest, and Multi-Layer Perceptron, and consider various settings of each classification algorithm. To evaluate, we perform five-fold cross validation and measure both the F1 score and AUC scores. We report the best result among all tested settings in Table 6 (F1 Score) and Table 7 (AUC Scores) for each feature set and classification algorithm.

There are many observations from Tables 6 and 7. Horizontally, Multi-Layer Perceptron outperforms the other three algorithms, and reaches the best  $F1 = 0.9200$  and  $AUC = 0.9749$ , when we use all features at the same time. Vertically, the performance tends to increase as more features are combined together. These results show the great potential of our classifier serving as a preliminary vitriol auto-filter on social media.

It is important to emphasize that since we used the Vitriolic Wordbag (See Table 2) as the keywords for crawling the potential vitriolic tweets, and most of the words in this bag have strongly emotional factors, the sentiment features of such tweets would be affected by our sampling method. Thus, we also highlighted

**Table 6.** F1 score for vitriol vs non-vitriol

	Log-Reg	SVM	RF	MLP
CS	0.8290	0.8323	0.8424	0.8384
LP	0.7636	0.8053	0.8028	0.8061
CR	0.8486	0.8386	0.8492	0.8597
LT	0.5761	0.5861	0.7265	0.6642
LP-CS	0.8564	0.8832	0.8886	0.8865
LP-CR-LT	0.8780	0.8810	0.8832	<b>0.8978</b>
all	0.8983	0.9007	0.9050	<b>0.9200</b>

the performance of our classifier only using LP+CR+LT (without sentiment) features in Tables 6 and 7. In this way, we can evaluate the performance of our classifier when we avoid bias introduced by the curation strategies we used. As a result, excluding the sentiment features, the performance of the classifier is still reasonably good ( $F1 = 0.8978$  and  $AUC = 0.9650$ ) when we use the Multi-Layer Perceptron algorithm. These results indicate that the models built over our curation strategy may be able to generalize to other domains (as we will test in more detail in a following experiment).

**Table 7.** AUC score for vitriol vs non-vitriol

	Log-Reg	SVM	RF	MLP
CS	0.8899	0.8992	0.8424	0.8384
LP	0.8289	0.8837	0.8912	0.8960
CR	0.9241	0.9154	0.9243	0.9342
LT	0.6095	0.6237	0.8101	0.7293
LP-CS	0.9263	0.9506	0.9541	0.9567
LP-CR-LT	0.9460	0.9504	0.9496	<b>0.9650</b>
all	0.9614	0.9628	0.9636	<b>0.9749</b>

### 5.3 Uncovering Vitriol In-the-wild

Since our curation framework is designed with many constraints to identify vitriol with high confidence, an open question is how well the trained models can perform over a collection of social media posts in-the-wild. That is, can we uncover evidence of vitriol even in cases where our original requirements are not observed (e.g., such as the relationship between replier and poster)? Toward answering this question, we evaluate the quality of distinguishing vitriol from all tweets containing profanity in the original set of potential vitriolic tweets  $P_{VT}$ . That

is, can our models trained over a set of highly-curated vitriolic tweets still apply to the wider space of tweets?

Concretely, we randomly select 100,000 tweets from  $P_{VT}$  and apply the Multi-Layer Perceptron classifier – which has the best performance in our previous experiments – to predict whether a tweet is vitriolic or not. In total, we uncover 55,650 tweets that are predicted to be vitriolic out of 100,000 (55.65%). We further manually annotated a random 100 of these positive tweets and find that 77 meet our definition of vitriol<sup>1</sup>. This suggests that: (i) the phenomenon of vitriol on social media seriously exceeds our expectation, and our vitriolic tweets set  $V_T$  is an accurate but small dataset out of all vitriol online; (ii) we still need to make greater efforts to design more sophisticated methods or features to capture the subtleties of vitriol, for distinguishing vitriolic tweets, because of the similarities between vitriol and posts that include profanity; and (iii) we may be able to relax our strategy of vitriol curation to identify even more vitriol for building more robust models.

#### 5.4 Vitriol in Other Domains

Complementing this tweet-based validation, we further evaluate the design of our vitriol classifier over Wikipedia comments [27], where the format and intent of the comments is quite different from our original tweet scenario. We adopt the annotated Personal Attacking comments on Wikipedia dataset [27] as an alternative dataset which has similar characteristics as vitriol. This dataset collects over 100k annotated discussion comments from Wikipedia in English. These comments from ordinary readers are similar to the replies on Twitter domain in terms of their unilateralism (there is no back-and-forth). In this data, every comment has been labeled by around 10 annotators on whether it is a personal attack or not. In summary, there are around 13,590 comments annotated as personal attacks out of 115,864 comments in total [27]. Note that this comment-based dataset does not contain any features of the repliers, so that we only use the linguistic features for analyzing and classifying. Also, the comment history of a single user in this dataset is not trackable, therefore, we can only recognize the attacking languages, not users.

First, we tested different classifiers and different feature sets over the Wikipedia personal attacks data as shown in Table 8. In this case, we train and test over the Wikipedia data, but use the features we identified from our tweet-based classifier presented earlier. We see that the MLP algorithm performs the best when we use all linguistic features, achieving an  $AUC = 0.9356$ . Not surprisingly, this result is below the result ( $AUC = 0.9719$ ) published in [27] over a classifier designed specially for Wikipedia comments. However, the good performance of our tweets-informed approach suggest that vitriol has common properties across domains that could be leveraged for high-quality vitriol detection.

We further construct a Wikipedia-specific classifier following the approach presented in Wulczyn *et al.* in [27]. Using this approach, we apply it to our set of vitriolic tweets where we see in Table 9 that the Wikipedia-based model

**Table 8.** AUC scores of attacks vs non-attacks on wikipedia dataset

	Log-Reg	SVM	RF	MLP
CS	0.8859	0.8752	0.8851	0.8953
LP	0.7120	0.6022	0.7178	0.7514
CR	0.9052	0.8982	0.8895	0.8875
LT	0.6277	0.6031	0.7997	0.6494
LP-CS	0.8389	0.8465	0.8943	0.8989
LP-CR-LT	0.8034	0.8207	0.8799	0.8942
all	0.8646	0.9063	0.9244	<b>0.9356</b>

results in an  $AUC = 0.8830$ , which is around 9.2% lower than our classifier’s performance on the same dataset.

**Table 9.** Comparing AUC scores for models across domains

	Tweets	Wiki comments
Our approach	0.9749	0.9356
Wulczyn approach	0.8830	0.9719

This suggests that while there are some commonalities across domains, that care should be taken in transferring models from domain to the other. In particular, since vitriol is an accumulated behavior that relies on a user’s history, approaches that consider user history may be more appropriate than those that rely on only a single post (be it tweet or comment).

## 6 Conclusion and Future Work

Vitriol has become a prominent societal issue, especially in social media. Distinct from hate speech and bullying, vitriol corresponds to a persistent coarsening of the discourse that leads to a more cumulative corrosive effect. Vitriol is challenging to define and study. Hence, in this paper, we have designed a vitriol curation framework as an initial step in our ongoing effort to extract vitriolic posts from social media with high confidence. We investigated a large collection of vitriolic posts sampled from Twitter, and examined both user-level and post-level characteristics of vitriol. We found key characteristics of vitriol that can distinguish it from non-vitriol, including popularity, network, sentiment, language structure, and content characteristics.

This is an initial attempt at formally studying vitriol on social media. While we have focused on one type of vitriol, our framework leaves open many questions about the size and composition of the larger space of all vitriol. What if we relax our (admittedly) strict requirements in Sect. 3? What if we change our initial tweet sampling strategy? What changes do we observe over time as vitriol evolves and transforms? While we have seen good success in distinguishing vitriol from non-vitriol posts over our sample, do these results hold over other varieties of vitriol?

In our ongoing work, we aim to continue this line of research through several avenues. First, since our curation framework of extracting vitriol on social media is primarily precision-focused, we will aim to expand our vitriol dataset by using well-designed statistical methods to relax our requirements. Second, we will relax our requirement that vitriol be unilateral to consider back and forth vitriol discourse as well. Third, we will temporally track the behaviors of our vitriolic users, e.g., to explore how many of them have been ultimately suspended. Fourth, we will incorporate more indicators to help us better characterize vitriol, such as the social networks around each user. Finally, we are interested in studying vitriol from the perspective of the users who are the targets of vitriol; are there strategies to incite or minimize the number of vitriolic attacks?

## References

1. Bamman, D., Smith, N.A.: Contextualized sarcasm detection on twitter. In: ICWSM, pp. 574–577 (2015)
2. Banks, J.: Regulating hate speech online. *Int. Rev. Law Comput. Technol.* **24**(3), 233–239 (2010)
3. Bosc, T., Cabrio, E., Villata, S.: Dart: a dataset of arguments and their relations on twitter. In: Proceedings of the 10th edition of the Language Resources and Evaluation Conference (2016)
4. Burnap, P., Williams, M.L.: Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. *Policy Internet* **7**(2), 223–242 (2015)
5. Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., Gilbert, E.: You can't stay here: the efficacy of reddit's 2015 ban examined through hate speech (2017)
6. Chen, C., Wu, K., Srinivasan, V., Zhang, X.: Battling the internet water army: Detection of hidden paid posters. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 116–120. IEEE (2013)
7. Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., Leskovec, J.: Anyone can become a troll: causes of trolling behavior in online discussions. arXiv preprint [arXiv:1702.01119](https://arxiv.org/abs/1702.01119) (2017)



8. Cheng, J., Danescu-Niculescu-Mizil, C., Leskovec, J.: Antisocial behavior in online discussion communities. In: ICWSM, pp. 61–70 (2015)
9. Clarke, I., Grieve, J.: Dimensions of abusive language on twitter. In: Proceedings of the First Workshop on Abusive Language Online, pp. 1–10 (2017)
10. Davidov, D., Tsur, O., Rappoport, A.: Semi-supervised recognition of sarcastic sentences in twitter and amazon. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pp. 107–116. Association for Computational Linguistics (2010)
11. Gao, L., Huang, R.: Detecting online hate speech using context aware models. arXiv preprint [arXiv:1710.07395](https://arxiv.org/abs/1710.07395) (2017)
12. Gimpel, K., et al.: Part-of-speech tagging for twitter: Annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, vol. 2, pp. 42–47. Association for Computational Linguistics (2011)
13. González-Ibáñez, R., Muresan, S., Wacholder, N.: Identifying sarcasm in twitter: a closer look. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, vol. 2, pp. 581–586. Association for Computational Linguistics (2011)
14. Google: Cloud natural language API (2017). <https://cloud.google.com/natural-language>. Accessed 12 Oct 2017
15. Hardaker, C.: Trolling in asynchronous computer-mediated communication: from user discussions to academic definitions (2010). <https://www.degruyter.com/view/j/jplr.2010.6.issue-2/jplr.2010.011/jplr.2010.011.xml>
16. IBM: Watson tone analyzer (2016). <https://www.ibm.com/watson/services/tone-analyzer>. Accessed 10 Oct 2017
17. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning (ICML-2014), pp. 1188–1196 (2014)
18. Lieberman, H., Dinakar, K., Jones, B.: Let’s gang up on cyberbullying. *Computer* **44**(9), 93–96 (2011)
19. Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: Proceedings of the 14th International Conference on World Wide Web, pp. 342–351. ACM (2005)
20. Macbeth, J., Adeyema, H., Lieberman, H., Fry, C.: Script-based story matching for cyberbullying prevention. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp. 901–906. ACM (2013)
21. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th International Conference on World Wide Web, pp. 145–153. International World Wide Web Conferences Steering Committee (2016)
22. Pavlopoulos, J., Malakasiotis, P., Androutsopoulos, I.: Deep learning for user comment moderation. arXiv preprint [arXiv:1705.09993](https://arxiv.org/abs/1705.09993) (2017)
23. Rehurek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Citeseer (2010)
24. Santana, A.D.: Virtuous or vitriolic: the effect of anonymity on civility in online newspaper reader comment boards. *Journalism Pract.* **8**(1), 18–33 (2014)
25. Serra, J., Leontiadis, I., Spathis, D., Stringhini, G., Blackburn, J., Vakali, A.: Class-based prediction errors to detect hate speech with out-of-vocabulary words. In: Proceedings of the First Workshop on Abusive Language Online, pp. 36–40 (2017)

26. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of the Second Workshop on Language in Social Media, pp. 19–26. Association for Computational Linguistics (2012)
27. Wulczyn, E., Thain, N., Dixon, L.: Ex machina: personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1391–1399. International World Wide Web Conferences Steering Committee (2017)