Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Characterizing Variation in Toxic Language by Social Context

Bahar Radfar, Karthik Shivaram, Aron Culotta
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
{bradfar, kshivara}@hawk.iit.edu, aculotta@iit.edu

Abstract

How two people speak to one another depends heavily on the nature of their relationship. For example, the same phrase said to a friend in jest may be offensive to a stranger. In this paper, we apply this simple observation to study toxic comments in online social networks. We curate a collection of 6.7K tweets containing potentially toxic terms from users with different relationship types, as determined by the nature of their follower-friend connection. We find that such tweets between users with no connection are nearly three times as likely to be toxic as those between users who are mutual friends, and that taking into account this relationship type improves toxicity detection methods by about 5% on average. Furthermore, we provide a descriptive analysis of how toxic language varies by relationship type, finding for example that mildly offensive terms are used to express hostility more commonly between users with no social connection than users who are mutual friends.

1 Introduction

Sociolinguistics posits that the social relationship between interlocutors influences linguistic choices (Hannerz 1970). As Gregory and Carroll (2018) state: "The relationship the user has with his audience, his addressee(s), is the situational factor that is involved in tenor of discourse." Many studies of online media support this notion – e.g., Pavalanathan and Eisenstein (2015) find that nonstandard language in tweets is more common when directed at small audiences rather than large audiences, and other work documents the prevalence of code switching in social media (Diab et al. 2016). In this paper, we investigate how toxic comments vary by social relationships on Twitter.[1]

To do so, we construct a dataset of tweets containing potentially toxic terms, then manually annotate them as truly toxic or not based on the context. We then categorize each tweet into one of four groups based on the social relationship of the users involved – does A follow B, B follow A, both, or neither. Doing so allows us to examine overall rates of toxicity by type of relationship, as well as to include additional features in text classification models for toxicity detection. Our main findings are that (i) of tweets containing potentially toxic terms, those sent between users with no social relationship are nearly three times as likely to be toxic as those sent between users with a mutual following relationship; (ii) including a feature indicating relationship type consistently improves classification accuracy by 4-5% across a variety of classification methods; (iii) linguistic choice in toxic tweets varies substantially across relationship type, with mutual friends using relatively more extreme language to express hostility than do users with no social connection.

Below, we first briefly summarize related work (§2), then describe the data collection (§3.1), annotation (§3.2), and classification (§3.3) methods, followed by the empirical results (§4) and concluding remarks (§5).[2]

2 Related Work

Detecting toxicity, cyberbullying, and trolling is an active area of study, with many text classification methods proposed (Yin et al. 2009; Al-garadi, Varathan, and Ravana 2016; Davidson et al. 2017; Cheng et al. 2017; Kumar, Cheng, and Leskovec 2017; Liu et al. 2018; Zhang et al. 2018). Most of the focus has been on engineering appropriate linguistic features, though some work has also considered demographics or other user attributes as possible signals.

Other work has examined linguistic variation across social communities (Jurgens 2011; Del Tredici and Fernández 2017; Danescu-Niculescu-Mizil et al. 2013). For example, Danescu-Niculescu-Mizil et al. (2013) study how user behavior changes as they learn the norms of a community, and Chandrasekharan et al. (2018) investigate how community norms affect moderation on Reddit, finding that certain types of behavior are allowed in some subreddits but banned in others. Additionally, recent work has uncovered biases introduced by race, gender, and sexual orientation (Park and Fung 2017; Dixon et al. 2018; Sap et al. 2019).

Our main contribution to this prior work is to investigate how toxicity varies by user relationship, and to improve classification using this evidence.

[1] Due to the nature of this research, this paper contains many terms that may be found offensive.

[2] All code and data are available at https://github.com/tapilab/icwsm-2020-toxic.

3 Methods

Our goal is to identify whether a tweet is *toxic* or not; for the purposes of this study, we define a tweet as toxic if it contains harassing, threatening, or offensive/harsh language directed toward a specific individual or group. Due to the directed nature of this definition, we restrict our study to tweets that are in reply to another tweet and contain a mention of another user. Thus, for each potentially toxic tweet, we identify the *author* a and *recipient* r.

To group each tweet by the relationship between the author and recipient, we consider four relationship types:

1. **no connection** $a \nleftrightarrow r$ – Neither user follows the other.
2. **symmetric connection** $a \leftrightarrow r$ – Both users follow each other.
3. **author follows recipient** $a \rightarrow r$
4. **recipient follows author** $a \leftarrow r$

Below, we describe how these data are collected and how we train classifiers to label them by toxicity.

3.1 Data Collection

We collected data using the Twitter API from January-March 2019. As our goal was to identify how social context influences the semantics of potentially toxic comments, we focused our data collection on tweets containing terms that often indicate toxicity. To identify a suitable set of keywords, we combined the top 20 most commonly used words in HateBase (HateBase 2019) with the intersection of terms on three other lists created at Shuttterstock (Emerick 2018), Google (Gabriel 2017), and CMU (von Ahn 2019). This resulted in 56 unique terms (see §6.1). We emphasize that due to this targeted sampling, we do not aim to make claims about overall rates of toxicity; instead we focus only on how the intent of potentially toxic terms varies by group.

We then used the Twitter Streaming API to track each of the 56 terms. In order to focus on messages that were directed at other users, we only retained tweets that contained a mention of another user and were in reply to another tweet. This resulted in ~178K messages collected over the three month period. For each message, we extracted the tweet's author and recipient (the user mentioned in the tweet), and queried the API again (`friendships/show`) to identify the relationship between them. Note that we executed this query immediately after collecting the tweet to best reflect the user relationship at the time the tweet was posted. Additionally, we collected the number of followers of both the author and recipient of the tweet.

Finally, for each tweet, we recollected it two weeks after it was posted to determine the number of likes and retweets it has received as well as whether or not it has been deleted. Table 1 summarizes the results of this initial data collection.

We can make a number of observations from this table. First, the $a \leftarrow r$ relationship is quite rare compared to the others (~2.7k versus ~79.9k for $a \leftrightarrow r$). This relationship is exemplified by the "celebrity strikes back" interaction — e.g., when a user with many followers responds with hostility to something said by one of their followers. This is further supported by the fact that the authors in the $a \leftarrow r$ relationship have the most number of followers on average (13.3K versus 3.3K for $a \leftrightarrow r$).

A second observation is that $a \nleftrightarrow r$ relationship is much more likely to be retweeted (.23 retweets on average versus .12 for $a \leftrightarrow r$) or deleted (15% versus 9.6% for $a \leftrightarrow r$). As we will see in a moment, this high rate of deletion suggests that many of these tweets are indeed toxic; furthermore, this indicates that there are many interactions among users who are not actually neighbors in the social graph (~33% of all tweets collected), often when users reply to more popular users whom they know about but do not follow. Insulting politicians from opposing parties is a common example in this category.

Finally, we note that the $a \leftrightarrow r$ type is the most common (~45% of all tweets), and also has the lowest deletion rate (9.6%), which matches our intuition that users who are closest to one another are more likely to use profanity, though not necessarily with hostile intent.

3.2 Data Annotation

We selected a subset of these collected tweets for human annotation, over-sampling the $a \leftarrow r$ category to have more similar sample sizes for each relationship type. We used three student annotators (only one of which is a co-author of this paper). Annotators were allowed to lookup additional context for each tweet to improve accuracy. Each tweet was given a binary toxicity label by two randomly assigned annotators; any disagreements were settled by a third annotator. 85% of annotations were in agreement by the first two annotators.

Table 2 gives statistics of the annotated data. As hinted at by the deletion rate in Table 1, the $a \leftrightarrow r$ tweets are much less likely to be toxic (28% versus 53% overall); whereas the $a \nleftrightarrow r$ tweets are the most likely to be toxic (75%). This nearly threefold increase in toxicity in $a \nleftrightarrow r$ over $a \leftrightarrow r$ indicates that relationship type is an important indicator to disambiguate truly toxic comments from those said in jest. Furthermore, the relatively high rate of toxicity for the $a \rightarrow r$ tweets (61%) was a bit surprising, indicating that a user will opt to follow another user (and thus read their tweets) while at the same time directing toxic messages towards them. This may in part be due to the phenomenon of "hate-following" (Ouwerkerk and Johnson 2016) or in part due to targeted trolling behavior.

3.3 Classification

We next train supervised text classifiers to categorize each message as toxic or not. We transform each tweet into a binary term vector, retaining emojis and hashtags but removing mentions and urls. We additionally add a feature for the length of each tweet (represented as a decile) to capture the intuition that toxic tweets are often longer than non-toxic tweets.

We consider three classification methods: logistic regression, random forests, and bidirectional LSTMs (Graves and Schmidhuber 2005). For logistic regression, we use the default setting of the scikit-learn implementation (Pedregosa et al. 2011); for random forest, we use 100 estimators with a minimum of three samples per leaf. For the LSTM,

group	tweets	tweet length	likes	retweets	author followers *mean/median*	recipient followers *mean/median*	% deleted
$a \leftrightarrow r$	79,909	15	1.5	0.12	3.3K/513	13K/0.7K	9.6%
$a \nleftrightarrow r$	59,465	22	2.6	0.23	1.3K/167	225K/1.3K	15.0%
$a \rightarrow r$	35,786	18	1.6	0.12	1.2K/129	487K/22K	11.4%
$a \leftarrow r$	2,786	20	2.9	0.16	13.3K/1.1K	5K/0.3K	12.8%

Table 1: Statistics of the tweets matching a list of toxic terms collected over a three month period.

group	tweets	% toxic
$a \leftrightarrow r$	2,179	28%
$a \nleftrightarrow r$	1,516	75%
$a \rightarrow r$	1,610	61%
$a \leftarrow r$	1,469	58%
total	**6,774**	**53%**

Table 2: Statistics of the labeled dataset.

method	auc	f1	precision	recall
rf	.772	.722	.724	.719
rf-rel	.818	**.765**	.737	**.794**
logreg	.769	.714	.733	.696
logreg-sep	.808	.748	.746	.751
logreg-rel	**.823**	.762	**.765**	.759
lstm	.751	.706	.708	.703
lstm-rel	.794	.746	.746	.747

Table 3: Classification results.

we use the TensorFlow implementation with two layers of 256 cells each; the input vector uses 300 dimensional pre-trained GloVe embeddings (Pennington, Socher, and Manning 2014). Adam optimization with default parameters is run for 40 epochs with a batch size of 64.

We considered two ways to incorporate relationship type information into classification. The first simply adds a single one-hot encoded feature indicating the relationship type (one of four possible values). The second is to train four separate classifiers, one per relationship type. This second variation was only used for logistic regression, since the other classifiers are non-linear, and thus can already represent interactions between relationship type and term features.

To summarize, the experiments compare seven classification methods:

- **rf**: Random forest without relationship type features
- **rf-rel**: Random forest with relationship type features
- **logreg**: Logistic regression without relationship type features
- **logreg-rel**: Logistic regression with relationship type features
- **logreg-sep**: Logistic regression; separate classifier trained for each relationship type.
- **lstm**: Bidirectional LSTM without relationship type features
- **lstm-rel**: Bidirectional LSTM with relationship type features

4 Results

We perform five-fold cross-validation on the 6.7K labeled tweets. Table 3 summarizes the average accuracy measures according to auc (area under the ROC curve), f1, precision, and recall, where the toxic class is considered the positive class. Comparing each baseline classifier to its variant that includes the relationship type feature (e.g., **rf** versus **rf-rel**), we see a consistent increase in f1 by 4-5% absolute. Both precision and recall improve for all methods, though recall improves a bit more on average. We find that this improvement also persists when stratifying results by relationship type. Comparing the three classifiers, both random forest and logistic regression perform similarly, and both outperform the LSTM, perhaps due to the limited number of training examples available to fit this more complex model.

Training separate logistic regression classifiers for each relationship type (**logreg-sep**) also appears to improve over the **logreg** baseline, though not as much as simply adding an additional feature. A complicating factor here is the limited number of training instances per relationship type, which most likely reduces the quality of the **logreg-sep** classifier.

To understand the types of errors that are corrected by introducing relationship type features, we identify tweets that are misclassified by **logreg** but correctly classified by **logreg-rel**. We find a number of examples of tweets that may be offensive when sent to a stranger, but could be seen as playful when written to friends. For example, tweets written between users with no relationship that are corrected to be classified as toxic include "Nigga, you disgusting" and "What the fuck is wrong with you?". Conversely, tweets between users who follow each other are corrected to be classified as non-toxic, e.g., "Kiss my ass 😂" and "bitch ur on crack." Of course, a deeper discourse analysis is often required to disambiguate these cases in general; however, here the group identity is sufficient to tip the classification decision into the correct class.

Finally, we investigated the linguistic differences in how toxicity is expressed in each relationship type. To do so, we compared the coefficients of the **logreg-rel** classifier with those of each of the four classifiers used in **logreg-sep**. That is, we aim to see how the coefficients of a classifier trained

on all relationship types differs from the coefficients of each classifier trained on only one relationship type. To visualize this, Figure 1 shows four scatter plots, one per relationship type. Each point represents a single term feature. On the x-axis is its coefficient in **logreg-rel**; on the y-axis is its coefficient in one of the **logreg-sep** classifiers. (All coefficients have been scaled with their z-score for easier comparison.)

Each plot also shows a linear fit – overall, most term coefficients are in agreement (correlations are ~0.55 for each plot). However, we can also examine some outliers to understand the differences. We plot the top ten terms based on their signed residuals (y value minus the linear fit). These are the terms that are more predictive of toxic class in one type of relationship as compared to overall. We can draw several observations from this figure. First, there are certainly some terms that are strongly indicative of toxicity regardless of relationship type (e.g., "cunt" and "twat"). On the other hand, the $a \nleftrightarrow r$ plot shows a number of terms that are perhaps only mildly toxic in general ("bitch", "dumb", "stupid"), but are toxic when mentioned between users with no preexisting relationship. Indeed, these three terms have much smaller coeffiecients in the $a \leftrightarrow r$ group than in the $a \nleftrightarrow r$ group ("bitch": 4.8 $\rightarrow$ 1.6; "dumb": 5.1 $\rightarrow$ 2.9; "stupid": 6.7 $\rightarrow$ .4). Additionally, the $a \leftarrow r$ type (the "celebrity strikes back" case) exhibits some intuitive terms like "shut" and "weird" that are often used to dismiss criticism from one's followers.

To further identify terms whose usage varies by group, we also examined the 56 original search terms and identified those whose coefficients decreased the most between $a \nleftrightarrow r$ and $a \leftrightarrow r$. The top term was "bitch", as noted above, followed by the terms "fuk" (2.4 $\rightarrow$ .6) and "bastard" (3.1 $\rightarrow$ 1.5). Thus, we find evidence confirming the intuition that some potentially toxic terms are often used in non-toxic ways amongst mutual friends.

5 Conclusion

In this paper, we collected and annotated a novel dataset of toxic messages from Twitter and investigated how toxic language varies by the nature of the relationship between two users, finding that even simple representations of this social information improves toxicity detection accuracy. On the one hand, we find that tweets containing toxic terms are more than twice as likely to have toxic intent when users have no relation than when they have a mutual following relation; on the other hand, the language used in toxic tweets is often more mild between users with no social connection, matching intuition that profanity among friends has different intent than profanity among strangers.

There are several limitations of this study. First, due to the targeted nature of sampling, we are unable to make claims about overall rates of toxicity among these groups, only about tweets containing one of the initial search keywords. Second, occasionally a directed comment is toxic towards someone other than the recipient, though it is difficult to reliably detect these. Finally, we have not attempted to infer attributes of individual users, so we have not investigated any potential sources of variation by demographic attributes. In future work, in addition to considering the above limitations, we will also investigate avenues to improve the modeling of social relationships — for example, measuring the tie strength between users rather than using four discrete groups — which may help produce more accurate content moderation algorithms.

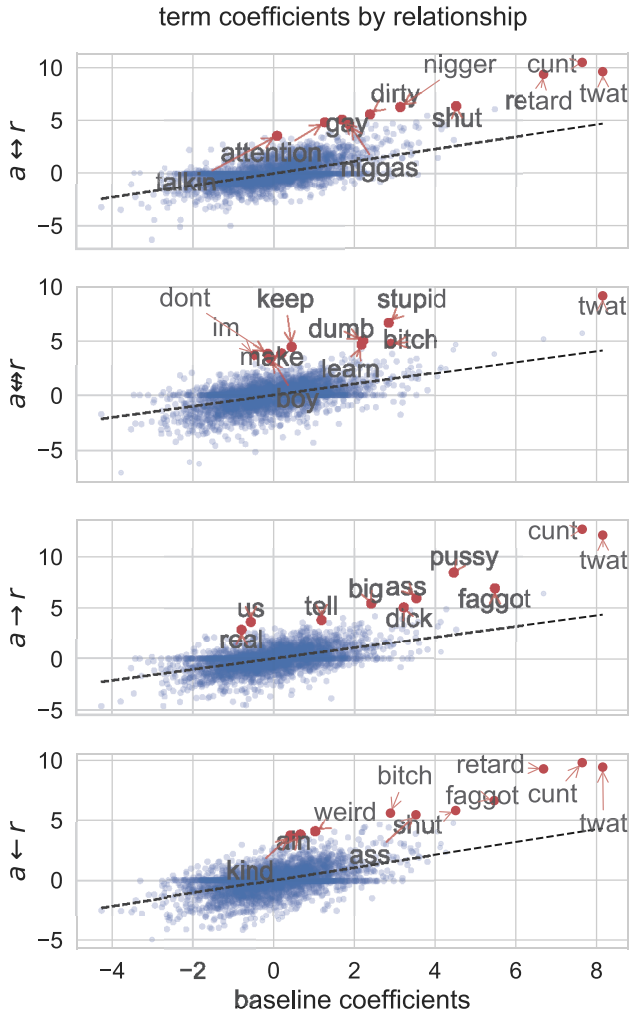

Figure 1: Coefficients from classifiers trained separately for each relationship type, compared with those of the baseline classifier trained on all types.

6 Appendix

6.1 Profane word list

We searched for tweets containing at least one of the following 56 keywords:

anal, anus, ass, bastard, bitch, boob, cock, coon, cripple, cum, cunt, dick, dildo, dyke, fag, faggot, fuck, fudgepacker, fuk, greaseball, gypo, handjob, homo, jihadi, jizz, kike, knacker, kunt, muff, muzzie, nigga, nigger, niggur, peckerwood, penis, piss, poop, porch monkey, pussy, queer, raghead, rape, retard, sand nigger, semen, sex, shit, shyster, slut, titties, twat, uncle tom, vagina, vulva, wank, yellow bone

References

Al-garadi, M. A.; Varathan, K. D.; and Ravana, S. D. 2016. Cybercrime detection in online communications. *Comput. Hum. Behav.* 63(C):433–443.

Chandrasekharan, E.; Samory, M.; Jhaver, S.; Charvat, H.; Bruckman, A.; Lampe, C.; Eisenstein, J.; and Gilbert, E. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact.* 2(CSCW):32:1–32:25.

Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, 1217–1230.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, 307–318. New York, NY, USA: ACM.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*.

Del Tredici, M., and Fernández, R. 2017. Semantic variation in online communities of practice. In *12th International Conference on Computational Semantics*.

Diab, M.; Fung, P.; Ghoneim, M.; Hirschberg, J.; and Solorio, T. 2016. Proceedings of the second workshop on computational approaches to code switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Emerick, J. 2018. List of dirty naughty obscene and otherwise bad words. https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words. [Online; accessed Feb-2019].

Gabriel, R. J. 2017. Google profanity words. "https://github.com/RobertJGabriel/Google-profanity-words. [Online; accessed Feb-2019].

Graves, A., and Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks* 18(5-6):602–610.

Gregory, M., and Carroll, S. 2018. *Language and situation: Language varieties and their social contexts*. Routledge.

Hannerz, U. 1970. Language variation and social relationships. *Studia Linguistica* 24(2):128–151.

HateBase. 2019. Hatebase. https://hatebase.org/. [Online; accessed Feb-2019].

Jurgens, D. 2011. Word sense induction by community detection. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, TextGraphs-6, 24–28. Stroudsburg, PA, USA: Association for Computational Linguistics.

Kumar, S.; Cheng, J.; and Leskovec, J. 2017. Antisocial behavior on the web: Characterization and detection. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, 947–950.

Liu, P.; Guberman, J.; Hemphill, L.; and Culotta, A. 2018. Forecasting the presence and intensity of hostility on instagram using linguistic and social features. In *Twelfth International AAAI Conference on Web and Social Media*.

Ouwerkerk, J. W., and Johnson, B. K. 2016. Motives for online friending and following: The dark side of social network site connections. *Social Media + Society* 2(3).

Park, J. H., and Fung, P. 2017. One-step and two-step classification for abusive language detection on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, 41–45. Vancouver, BC, Canada: Association for Computational Linguistics.

Pavalanathan, U., and Eisenstein, J. 2015. Audience-modulated variation in online social media. *American Speech* 90(2):187–213.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research* 12(Oct):2825–2830.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678.

von Ahn, L. 2019. List of bad words. https://www.cs.cmu.edu/~biglou/resources/bad-words.txt. [Online; accessed Feb-2019].

Yin, D.; Xue, Z.; Hong, L.; Davison, B. D.; Kontostathis, A.; and Edwards, L. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2:1–7.

Zhang, J.; Chang, J.; Danescu-Niculescu-Mizil, C.; Dixon, L.; Hua, Y.; Taraborelli, D.; and Thain, N. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1350–1361.