An Examination of Regret in Bullying Tweets

Jun-Ming Xu, Benjamin Burchfiel, Xiaojin Zhu
Department of Computer Sciences
University of Wisconsin-Madison
Madison, WI 53706, USA
{xujm,burchfie,jerryzhu}@cs.wisc.edu

Amy Bellmore
Department of Educational Psychology
University of Wisconsin-Madison
Madison, WI 53706, USA
abellmore@wisc.edu

Abstract

Social media users who post bullying related tweets may later experience regret, potentially causing them to delete their posts. In this paper, we construct a corpus of bullying tweets and periodically check the existence of each tweet in order to infer if and when it becomes deleted. We then conduct exploratory analysis in order to isolate factors associated with deleted posts. Finally, we propose the construction of a regrettable posts predictor to warn users if a tweet might cause regret.

1 Introduction

A large body of literature suggests that participants in bullying events, including victims, bullies, and witnesses, are likely to report psychological adjustment problems (Jimerson, Swearer, and Espelage, 2010). One potential source of therapy for these issues can be self-disclosure of the experience to an adult or friend (Mishna and Alaggia, 2005); existing research suggests that victims who seek advice and help from others report less maladjustment than victims who do not (Shelley and Craig, 2010).

Disclosure of bullying experiences through social media may be a particularly effective mechanism for participants seeking support because social media has the potential to reach large audiences and because participants may feel less inhibition when sharing private information in an online setting (Walther, 1996). Furthermore, there is evidence that online communication stimulates self-disclosure, which leads to higher quality social relationships and increased well-being (Valkenburg and Peter, 2009).

Online disclosure may also present risks for those involved in bullying however, such as revictimization, embarrassment, and social ostracization. Evidence exists that some individuals may react to these risks retroactively, by deleting their social media posts (Child et al., 2011; Christofides, Muise, and Desmarais, 2009). Several relevant motives have been found to be associated with deleting posted information, including conflict management, safety, fear of retribution, impression management, and emotional regulation (Child, Haridakis, and Petronio, 2012).

Our previous work (Xu et al., 2012) demonstrates that social media can be a valuable data source when studying bullying, and proposes a text categorization method to recognize social media posts describing bullying episodes, *bullying traces*. To better understand, and possibly prevent, user regret after posting bullying related tweets, we collect bullying traces using the same method and perform regular status checks to determine if and when tweets become inaccessible. While a tweet becoming inaccessible does not guarantee it has been deleted, we attempt to leverage http response codes to rule out other common causes of inaccessibility. Speculating that regret may be a major cause of deletion, we first conduct exploratory analysis on this corpus and then report the results of an off-the-shelf regret predictor.

2 Data Collection

We adopt the procedure used in (Xu et al., 2012) to obtain bullying traces; each identified trace contains

at least one bullying related keyword and passes a bullying-or-not text classifier.

Our data was collected in realtime using the Twitter streaming API; once a tweet is collected, we query its url (`https://twitter.com/USERID/status/TWEETID`) at regular intervals and infer its status from the resulting http response code. We interpret an HTTP 200 response as an indication a tweet still exists and an HTTP 404 response, which indicates the tweet is unavailable, as indicating deletion. A user changing their privacy settings can also result in an HTTP 403 response; we do not consider this to be a deletion. Other response codes, which appear quite rarely, are treated as anomalies and ignored. All non HTTP 200 responses are retried twice to ensure they are not transient oddities.

To determine when a tweet is deleted, we attempted to access each tweet at time points $T_i = 5 \times 4^i$ minutes for $i = 0, 1 \ldots 7$ after the creation time. These roughly correspond to periods of 5 minutes, 20 minutes, 1.5 hours, 6 hours, 1 day, 4 days, 2 weeks, and 2 months. While we assume that user deletion is the main cause of a tweet becoming unavailable, other causes are possible such as the censorship of illegal contents by Twitter (Twitter, 2012).

Our sample data was collected from July 31 through October 31, 2012 and contains 522,984 bullying traces. Because of intermittent network and computer issues, several multiple day data gaps exist in the data. To combat this, we filter our data to include only tweets of unambiguous status. If any check within the 20480 minutes (about two weeks) interval returns an HTTP 404 code, the tweet is no longer accessible and we consider it *deleted*. If the 20480 minute or 81920 minute check returns an HTTP 200 response, that tweet is still accessible and we consider it *surviving*. The union of the surviving and deleted groups formed our cleaned dataset, containing 311,237 tweets in total.

3 Exploratory Data Analysis

A user's decision to delete a bullying trace may be the result of many factors which we would like to isolate and understand. In this section we will examine several such possible factors.

3.1 Word Usage

Our dataset contains 331,070 distinct words and we are interested in isolating those with a significantly higher presence among either deleted or surviving tweets. We define the odds ratio of a word w

$$r(w) = \frac{P(w \mid \text{deleted})}{P(w \mid \text{surviving})},$$

where $P(w \mid \text{deleted})$ is the probability of word w occurring in a deleted tweet, and $P(w \mid \text{surviving})$ is the probability of w appearing in a surviving tweet. In order to ensure stability in the probability estimation, we only considered words appearing at least 50 times in either the surviving or deleted corpora.

Following (Bamman, OConnor, and Smith, 2012), we qualitatively analyzed words with extreme values of $r(w)$, and found some interesting trends. There was a significant tendency for "joking" words to occur with $r(w) < 0.5$; examples include "xd," "haha," and "hahaha." Joking words occur less frequently in deleted tweets than surviving ones. On the other end of the spectrum, there were no joking words with $r(w) > 2$. What we found instead were words such as "rip," "fat," "kill," and "suicide." While it is relatively clear that joking is less likely to occur in deleted tweets, there was less of a trend among words appearing more frequently in deleted tweets.

3.2 Surviving Time

Let N be the total number of tweets in our corpus, and $D(T_i)$ be the number of tweets that were first detected as deleted at minute T_i after creation. Note that $D(T_i)$ is not cumulative over time: it includes only deletions that occurred in the time interval $(T_{i-1}, T_i]$. Then we may define the deletion rate at time T_i as

$$R_T(T_i) = \frac{D(T_i)}{N(T_i - T_{i-1})}.$$

In other words, $R_T(t)$ is the fraction of tweets that are deleted during the one minute period $(t, t+1)$.

We plot R_T vs. t using logarithmic scales on both axes in Figure 1 and the result is a quite strong linear trend. Fitting the plot with a linear regression, we derive an inverse relationship between R_T and t of the form

$$R_T(t) \propto 1/t.$$

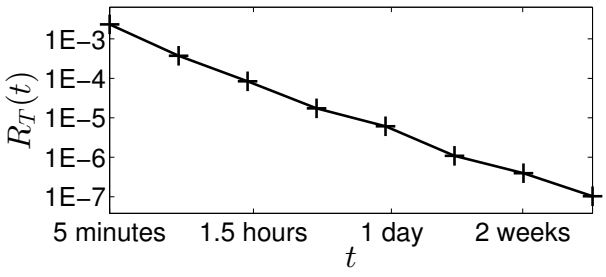

Figure 1: Deletion rate decays over time.

This result makes sense; the social effects of a particular bullying tweet may decay over time, making regret less of a factor. Furthermore, the author may assume an older tweet has already been seen, rendering deletion ineffective. Additionally, because the drop off in deletion rate is so extreme, we are able to safely exclude deletions occurring after two weeks from our filtered dataset without introducing a significant amount of noise. Finally, $\sum_{t=0}^{\infty} R_T(t)$ gives the overall fraction of deletion, which in our case is around 4%.

3.3 Location and Hour of Creations

Some bullying traces contain location meta-data in the form of GPS coordinates or a user-created profile string. We employed a reverse geocoding database (`http://www.datasciencetoolkit.org`) and a rule-based string matching method to map these tweets to their origins (at the state level; only for tweets within the USA). This also allowed us to convert creation timestamps from UTC to local time by mapping user location to timezone. Because many users don't share their location, we were only able to successfully map 85,465 bullying traces to a US state s, and local hour of day h. Among these traces, 3,484 were deleted which translates to an overall deletion rate of about 4%.

Let $N(s, h)$ be the count of bullying traces created in state s and hour h. Aggregating these counts temporally yields $N_S(s) = \sum_h N(s, h)$, while aggregating spatially produces $N_H(h) = \sum_s N(s, h)$. Similarly, we can define $D(s, h)$, $D_S(s)$ and $D_H(h)$ as the corresponding counts of deleted traces. We can now compute the deletion rate

$$R_H(h) = \frac{D_H(h)}{N_H(h)}, \text{ and } R_S(s) = \frac{D_S(s)}{N_S(s)}.$$

The top row of Figure 2 shows $N_H(h)$, $D_H(h)$, and $R_H(h)$. We find that $N_H(h)$ and $D_H(h)$ peak in the evening, indicating social media users are generally more active at that time. The peak of $R_H(h)$ appears at late night and, while there are multiple potential causes for this, we hypothesize that users may fail to fully evaluate the consequences of their posts when tired. The bottom row of Figure 2 shows $N_S(s)$, $D_S(S)$, and $R_S(s)$. The plot of $N_S(s)$ shows that bullying traces are more likely to originate in California, Texas or New York which is the result of a population effect. Importantly however, the deletion rate $R_S(s)$ is not affected by population bias and we see, as expected, that spatial differences in $R_S(s)$ are small. We performed χ^2-test to see if a state's deletion rate is significantly different from the national average. We chose the significance level at 0.05 and used Bonferroni correction for multiple testing. Only four states have significantly different deletion rates from the average: Arizona (6.3%, $p = 5.9 \times 10^{-5}$), California (5.2%, $p = 2.7 \times 10^{-7}$), Maryland (1.9%, $p = 2.3 \times 10^{-5}$), and Oklahoma (7.1%, $p = 3.5 \times 10^{-5}$).

3.4 Author's Role

Participants in a bullying episode assume well-defined roles which dramatically affect the viewpoint of the author describing the event. We trained a text classifier to determine author role (Xu et al., 2012), and used it to label each bullying trace in the cleaned corpus by author role: *Accuser*, *Bully*, *Reporter*, *Victim* or *Other*.

Table 1 shows that compared to tweets produced by bullies, victims create more bullying traces, possibly due to an increased need for social support on the part of the victim. More importantly, $P(\text{deleted} \mid \text{victim})$ is higher than $P(\text{deleted} \mid \text{bully})$, a statistically significant difference in a two-proportion z-test. Possibly, victims are more sensitive to their audience's reaction than bullies.

3.5 Teasing

Many bullying traces are written jokingly. We built a text classifier to identify teasing bullying traces (Xu et al., 2012) and applied it to the cleaned corpus.

Table 2 shows that $P(\text{deletion} \mid \text{Teasing})$ is much lower than $P(\text{deletion} \mid \text{Not Teasing})$ and the difference is statistically significant in a two-proportion z-

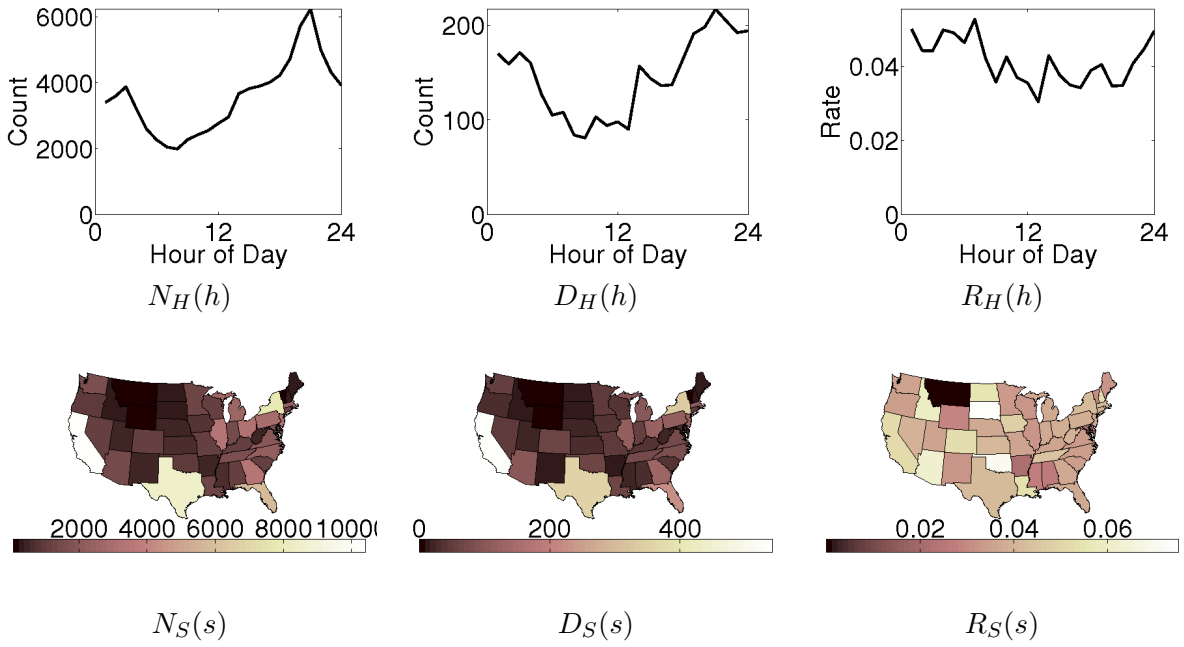

Figure 2: Counts and deletion rates of geo-tagged bullying traces.

	Deleted	Total	$P(\text{deleted} \mid \text{Role})$
Accuser	2541	50088	5.07%
Bully	1792	30123	5.95%
Reporter	11370	147164	7.73%
Victim	6497	83412	7.79%
Other	41	450	9.11%

Table 1: Counts and deletion rate for different roles.

	Deleted	Total	$P(\text{deleted} \mid \text{Teasing?})$
Yes	858	22876	3.75%
Not	21383	288361	7.42%

Table 2: Counts and deletion rate for teasing or not.

test. It seems plausible that authors are less likely to regret teasing posts because they are less controversial and have less potential to generate negative audience reactions. This also corroborates our findings in word usage that joking words are less frequent in deleted tweets.

4 Predicting Regrettable Tweets

Once a bullying tweet is published and seen by others, the ensuing effects are often impossible to undo. Since ill-thought-out posts may cause unexpectedly negative consequences to an author's reputation, relationship, and career (Wang et al., 2011), it would be helpful if a system could warn users before a potentially regrettable tweet is posted. One straightforward approach is to formulate the task as a binary text categorization problem.

We use the cleaned dataset, in which each tweet is known to be surviving or deleted after 20480 minutes (about two weeks). Since this dataset contains 22,241 deleted tweets, we randomly sub-sampled the surviving tweets down to 22,241 to force our deleted and surviving datasets to be of equal size. Consequentially, the baseline accuracy of the classifier is 0.5. While this does make the problem artificially easier, our initial goal was to test for the presence of a signal in the data.

We then followed the preprocessing procedure in (Xu et al., 2012), performing case-folding, anonymization, and tokenization, treating URLs, emoticons and hashtags specially. We also chose the unigrams+bigrams feature representation, only keeping tokens appearing at least 15 times in the corpus.

We chose to employ a linear SVM implemented in LIBLINEAR (Fan et al., 2008) due to its efficiency on this large sparse text categorization task and a 10-fold cross validation was conducted to eval-

uate its performance. Within the first fold, we use an inner 5-fold cross validation on the training portion to tune the regularization parameter on the grid $\{2^{-10}, 2^{-9}, \ldots, 1\}$; the selected parameter is then fixed for all the remaining folds.

The resulting cross validation accuracy was 0.607 with a standard deviation of 0.012. While it is statistically significantly better than the random-guessing baseline accuracy of 0.5 with a p-value of 5.15×10^{-10}, this accuracy is nevertheless too low to be useful in a practical system. One possibility is that the tweet text contains very limited information for predicting inaccessibility; a user's decision to delete a tweet potentially depends on many other factors, such as the conversation context and the characteristics of the author and audience.

In the spirit of exploring additional informative features for deletion prediction, we also used the teasing and author role classifiers in (Xu et al., 2012), and appended the predicted teasing, and author role labels to our feature vector. This augmented feature representation achieved a cross validation accuracy of 0.606, with standard deviation 0.007; not statistically significantly different from the text-only feature representation. While it seems that a signal does exist, leveraging it usefully in real world scenarios may prove challenging due to the highly-skewed nature of the data.

5 Discussion

There have been several recent works examining causes of deletion in social media. Wang *et al.* (2011) qualitatively investigated regret associated with users' posts on social networking sites and identified several possible causes of regret. Bamman *et al.* (2012) focused on censorship-related deletion of social media posts, identifying a set of sensitive terms related to message deletion through a statistical analysis and spatial variation of deletion rate.

Assuming that deletion in social media is indicative of regret, we studied regret in a bullying context by analyzing deletion trends in bullying related tweets. Through our analysis, we were able to isolate several factors related to deletion, including word usage, surviving time, and author role. We used these factors to build a regret predictor which achieved statistically significant results on this very noisy data. In the future, we plan to explore more factors to better understand deletion behavior and regret, including users' recent posts, historical behavior, and other statistics related to their specific social network.

Acknowledgments

We thank Kwang-Sung Jun, Angie Calvin, and Charles Dyer for helpful discussions. This work is supported by National Science Foundation grants IIS-1216758 and IIS-1148012.

References

Bamman, David, Brendan OConnor, and Noah Smith. 2012. Censorship and deletion practices in chinese social media. *First Monday*, 17(3-5).

Child, Jeffrey T., Paul M. Haridakis, and Sandra Petronio. 2012. Blogging privacy rule orientations, privacy management, and content deletion practices: The variability of online privacy management activity at different stages of social media use. *Computers in Human Behavior*, 28(5):1859 – 1872.

Child, Jeffrey T, Sandra Petronio, Esther A Agyeman-Budu, and David A Westermann. 2011. Blog scrubbing: Exploring triggers that change privacy rules. *Computers in Human Behavior*, 27(5):2017–2027.

Christofides, Emily, Amy Muise, and Serge Desmarais. 2009. Information disclosure and control on facebook: are they two sides of the same coin or two different processes? *CyberPsychology & Behavior*, 12(3):341–345.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Jimerson, Shane R., Susan M. Swearer, and Dorothy L. Espelage. 2010. *Handbook of Bullying in Schools: An international perspective*. Routledge/Taylor & Francis Group, New York, NY.

Mishna, Faye and Ramona Alaggia. 2005. Weighing the risks: A child's decision to disclose peer victimization. *Children & Schools*, 27(4):217–226.

Shelley, Danielle and Wendy M Craig. 2010. Attributions and coping styles in reducing victimization. *Canadian Journal of School Psychology*, 25(1):84–100.

Twitter. 2012. The twitter rules. `http://support.twitter.com/articles/18311-the-twitter-rules`.

Valkenburg, Patti M and Jochen Peter. 2009. Social consequences of the internet for adolescents a decade of research. *Current Directions in Psychological Science*, 18(1):1–5.

Walther, Joseph B. 1996. Computer-mediated communication impersonal, interpersonal, and hyperpersonal interaction. *Communication research*, 23(1):3–43.

Wang, Yang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. "I regretted the minute I pressed share": a qualitative study of regrets on facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, pages 10:1–10:16. ACM.

Xu, Jun-Ming, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 656–666, Montréal, Canada, June. Association for Computational Linguistics.