Tweet Properly: Analyzing Deleted Tweets to Understand and Identify Regrettable Ones

Lu Zhou
DIAC Lab, Kno.e.sis Center
Department of Computer Science and Engineering
Wright State University
Dayton, OH 45435
kbzhoulu@gmail.com

Wenbo Wang
Kno.e.sis Center
Department of Computer Science and Engineering
Wright State University
Dayton, OH 45435
wenbo@knoesis.org

Keke Chen
DIAC Lab, Kno.e.sis Center
Department of Computer Science and Engineering
Wright State University
Dayton, OH 45435
keke.chen@wright.edu

ABSTRACT

Inappropriate tweets can cause severe damages on authors' reputation or privacy. However, many users do not realize the negative consequences until they publish these tweets. Published tweets have lasting effects that may not be eliminated by simple deletion because other users may have read them or third-party tweet analysis platforms have cached them. Regrettable tweets, i.e., tweets with identifiable regrettable contents, cause the most damage on their authors because other users can easily notice them. In this paper, we study how to identify the regrettable tweets published by *normal individual users* via the contents and users' historical deletion patterns. We identify normal individual users based on their publishing, deleting, followers and friends statistics. We manually examine a set of randomly sampled deleted tweets from these users to identify regrettable tweets and understand the corresponding regrettable reasons. By applying content-based features and personalized history-based features, we develop classifiers that can effectively predict regrettable tweets.

Keywords

Twitter; Regret; Deleted Tweets; User Clustering; Regrettable Tweets

1. INTRODUCTION

Twitter is a popular online social network where users can post their thoughts, share photos, and have conversions with other users publicly in a real-time fashion. While it is convenient to communicate with others on Twitter, people sometimes mistakenly post tweets that they will feel regret later [23]. For example, people may feel inappropriate after venting out frustrations about friends or managers. Moreover, people may disclose *personally embarrassing information* [5] when interacting with their friends on Twitter, forgetting that tweets are visible to the public [2]. Table 1 lists a few sample tweets that were deleted due to regrettable reasons. These inappropriate tweets may be read and retweeted by a large number of people before authors delete them. Furthermore, the deleted tweets may still be available for people to obtain from third-party tweet miners [1]. As a result, tweet deletion does not eliminate the risk of privacy disclosure or harm to self-image. It would be ideal if a system can identify such regrettable tweets before authors publish them.

Table 1: Sample content-identifiable regrettable tweets

1	My *sister* is so *childish* oh my goodness
2	Feeling better no more *hangover* x) http://t.co/selfie_pic
3	Work work work seems like that's all I do since I started my job! Ughhh I need more time

However, developing such a system is challenging for several reasons. First, training a model to automatically identify regrettable tweets typically needs a large number of training examples, i.e., labeled regrettable tweets. However, there is no effective method to automatically collect a large number of such tweets. Survey-based methods [29] can be used to collect a small number of examples, but they are expensive and difficult to be extended to the scale of Twitter.

Second, whether an author will regret about a tweet is an entirely personal choice and probably subject to many factors. One person's regrettable content might be acceptable to another person. It is also understandable that an individual tweet only becomes regrettable in a certain context. However, this contextual information may be beyond what is available on Twitter, making it difficult to extract regrettable examples.

Inspired by the observation that users will delete tweets when they feel regret, we believe that the deleted tweets are a valuable source to study regrettable tweets. It is fortunate that we can retrieve deleted tweets via Twitter's streaming API [1], which provides abundant data for this study. However, deleted tweets are naturally noisy as deleted tweets are not necessarily attributed to regrets. Almuhimedi et al. [1] have presented a statistical study of deleted tweets. They find tweets might be deleted for many reasons such as misspelling, rephrasing, spamming, and regrets. Besides, according to recent studies, spammers may also deliberately

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883052.

delete tweets to mimic normal users [14]. Twitter also actively takes fighting tweets published by identified (or reported) spammers[1]. Thus, it is challenging to extract regrettable tweets from deleted ones.

A previous study [1] suggests that simple content analysis may not distinguish deleted tweets from normal ones. In particular, both of them contain approximately the same percentage of sensitive information, such as offensive comments, alcohol/illegal drug use, and sexual activity. This result is intuitive as it is an entirely personal choice to publish such contents. Thus, we believe that identifying regrettable contents should be personalized. How to capture the personalization signals is another challenge.

Scope and Contributions. In this paper we will address the challenges in collecting, mining, and identifying a subset of regrettable tweets: *content-identifiable regrettable tweets*. Content-identifiable regrettable tweets are those that readers can capture the sensitive meaning based on the content only. We believe identifying such a subset is a top priority because their privacy damage can be easily noticed and quickly propagated. In contrast, more subtle regrettable tweets that depend on the contextual information may escape from most readers' attention. We will develop classifiers to predict whether a user will regret about publishing a sensitive tweet based on their personal preferences.

Extracting possibly regrettable tweets from noisy deleted tweets is difficult. Our strategy is to focus on the tweets by *normal individual users*, as they form the majority of the Twitter users. In contrast, *verified users* are typically celebrities and organizations, who publish a lot of tweets and proportionally delete more tweets than normal users do. Other users may include bulk deletion users and spam users. Bulk deletion is regularly performed by some users, who may not regret about publishing the deleted tweets, and thus, their portion of deleted tweets should be excluded from our study.

The different types of users are approximately identified by *user clustering*. Note that normal individual users cannot be simply identified from their Twitter profiles as there is no effective way to identify whether a profile is fake or not. A more reliable way is to understand normal individual users' tweeting behaviors and use the behavioral features to describe the group. We design a set of features to describe the user tweeting characteristics and apply self-organized map [16] to find the group of likely normal individual users. Based on the clustering result, we can eliminate about 17% of deleted tweets that are not contributed by the normal individual users.

Then, we sample tweets deleted by these likely normal users for content analysis. They are further cleaned by removing the tweets deleted for non-regrettable reasons such as retweets and rephrasing. The remaining tweets are examined, understood, and manually labeled by annotators. We find that the deleted tweets with identifiable regrettable contents compose about 18% of the deleted tweets. After summarizing the possible regrettable reasons, we build lexicons for each specific reason with the help of WordNet, Urban Dictionary, and relevant studies [28, 24], which are used to derive features describing regrettable tweets.

We capture users' preferences of publishing regrettable contents by mining their tweeting history. With one month of historical published and deleted tweets, for each user we derive the publishing and deletion statistics for each regrettable reason, which form a set of user preference features.

With these content and user preference features, we develop classifiers to effectively distinguish the regrettable tweets from non-deleted ones. Our study shows that these features are better than the generic language features such as Unigram, Bigram, and Part-of-Speech (POS) features.

In summary, our research has the following contributions:

- To remove deleted tweets that are unlikely regrettable, we develop a user-clustering method to analyze user tweeting behaviors and exclude the users who are less likely to produce regrettable tweets. The result allows us to eliminate a significant portion of noisy deleted tweets.

- Considering personal preferences on publishing regrettable contents, we design a set of content and history-based features for classifier modeling. The result shows that these features can be used to effectively distinguish regrettable tweets from published tweets.

The remaining part of the paper will give the background and definitions first (Section 2), present the user filtering analysis (Section 3), analyze sample tweets from normal individual users, and develop classifiers to identify a subset of regrettable tweets (Section 4).

2. BACKGROUND AND DEFINITIONS

Twitter is a popular microblogging service where users can post *tweets/statuses* of up to 140 characters. Newly posted tweets will show up in the timelines of their authors as well as the timelines of users who *follow* these authors. A user can re-post a tweet (i.e., retweet) from another user or delete his/her tweets. In the following, we define the terms that will be used in later discussions.

Deleted tweets. A tweet can be deleted in both active and passive fashions. If a user clicks the 'delete tweet' button to delete one of their tweets, the tweet will be removed from the user's timeline as well as the followers' timelines. The retweets of this deleted tweet will also be automatically deleted - the passive deletion. In both fashions, Twitter will send out a "Status Deletion Notice" via its streaming API to notify third-party clients. We use the deletion notice to mark the deleted tweets. Deleted tweets in our study refer to the tweets that were deleted in a given time window. These deleted tweets may be published and deleted during the given time window, or published before this time window but deleted during the given time window.

Tweets are actively deleted for many reasons, such as typos, rephrasing, and spamming, which are excluded from our study. We classify the remaining ones into two classes.

- ***Content-identifiable regrettable tweets*** refer to the tweets that were deleted for certain identifiable regrettable reasons. These reasons can be understood based on *the content only*. Clearly, such regrettable tweets compose only a subset of all the regrettable tweets. Table 1 lists a few such regrettable tweets.

- ***Unsure tweets*** are the deleted tweets whose contents do not indicate any identifiable regrettable reasons. Some of them can still be regrettable tweets. However,

[1] `https://support.twitter.com/articles/64986-reporting-spam-on-twitter`

to fully understand them, we need to explore complex contextual information beyond contents. Table 5 lists some unsure tweets.

Non-deleted tweets. If a tweet has been kept and not deleted after a certain time window, we call it a non-deleted tweet. In this study, we set the time window to be seven days, and tweets deleted out of this time window is beyond the scope of this paper.

Published tweets. Published tweets are all tweets which are posted in a given time window, including all the non-deleted tweets and a part of deleted tweets that are published and deleted in the same window. Note that the remaining deleted tweets in this time window were published before this window, and thus not counted as a part of the published tweets. As a result, the number of published tweets may be smaller than the sum of deleted tweets plus non-deleted tweets. It is meaningful to include the concept of published tweets to understand users' publishing and deleting patterns in a specific time window.

3. UNDERSTANDING DELETION BEHAVIORS WITH USER CLUSTERING

We clean noisy deleted tweets to exclude deletions from certain categories of users because these deletions are less likely regrettable deletions. There are many categories of users on Twitter, such as spammers, corporation users, celebrity users, and normal individual users. Our focus is on normal individual users, who make up the majority of users and have distinct publishing and deleting patterns. These normal individual users cannot be easily identified by their Twitter profiles, because profiles can be fake (unless they are verified) and there is no effective way to identify whether a profile is authentic. However, we believe that normal individual users should share certain characteristics of tweeting behaviors, which can help us to identify the group of highly likely normal individual users. To locate such users, we propose an effective user clustering method based on a set of features describing a user's tweeting behaviors.

In the following, we will describe data collection, feature design, and clustering analysis. Since verified users are typically celebrities, politicians, and organizations, we exclude them from our dataset before clustering.

3.1 Collecting and Cleaning Data

We used Twitter's sample streaming API [2] to collect a random sample of published tweets for one week from May 12th 2014 to May 18th 2014. Meanwhile, we also collected "status deletion notices" from Twitter that indicate which published tweets were deleted. To retain only English tweets, we kept only the users who specified "en" as their language in their profiles. To deal with multilingual users who post in different languages, we applied Google Chrome Browser's embedded language detector to remove non-English tweets. In total, we gathered about 1.25M published tweets, 440,431 deleted tweets (323,768 of them were published and deleted in our time window, 116,663 of them were published before our time window, but are deleted in our time window), and 929,558 non-deleted tweets, from about 279k distinct user accounts. Table 2 provides statistics of the collected tweets.

[2] `https://dev.twitter.com/streaming/reference/get/statuses/sample`

Table 2: Statistics on the collected random sample of tweets

Total number of users	279,360
Total number of published tweets	1,253,326
Total number of non-deleted tweets	929,558
Total number of deleted tweets	440,431
Number of users who deleted at least 1 tweet	252,528
Number of users who posted at least 1 tweet	279,360

3.2 Clustering Users

To identify different categories of users, we apply a clustering-based approach to partition users. In the following, we will introduce our feature design and clustering analysis.

Feature Design. To conduct user clustering, we represent each user with a five-dimensional vector that is derived from the one-week sample data and users' Twitter profile:

- ***Number of published tweets*** is the total number of published tweets that users post in this one-week period, including the ones that were published in this period but were deleted later.
- ***Number of deleted tweets*** is the total number of deleted tweets that users delete in this period. Besides deleting tweets that were published in this period, a user can also delete tweets that were published before this period. Thus, the number of deleted tweets may exceed the number of published tweets.
- ***Deletion ratio*** is defined as the number of deleted tweets divided by the number of published tweets.
- ***Fano factor*** measures the dispersion of a probability distribution of a Fano noise [8], which is used to identify possible bulk deletions. Bulk deletions should be excluded from our study as these deleted tweets are less likely regrettable ones. Fano factor is defined as σ_d^2/μ_d for the seven-day period, where μ_d is the mean of deleted tweets per day, and σ_d is the standard deviation of seven days.
- ***Reputation*** is defined by Thomas et al. [25] as (the number of followers)/(the number of followers + the number of followings) for a Twitter user. The followers of the user receive the user's tweets; the user are also following other users to receive their tweets. They observed that normal users are likely to follow back when others follow them, and their reputation values are around 0.5.

For each dimension, each value x is normalized by $(x-\mu)/\sigma$, where μ is the mean of the dimension, and σ is the standard deviation of the dimension. This step is necessary for clustering to avoid large-value dimensions dominating the clustering results.

Clustering Analysis. Intuitively, the normal individual users should be an overwhelming proportion of the users, which incurs a unique difficulty in clustering. Most existing clustering algorithms such as the k-means algorithm cannot handle such heavily skewed cluster distributions [11]. Other valid candidates such as CURE [11] are too expensive to handle the scale of our dataset - they often require $O(N^2)$ memory and $O(N^2 \log N)$ time. Thus, we decide to use self-organizing map (SOM) [16] to reduce the dataset but preserve the skewed clustering structure, and then apply the complete-linkage hierarchical clustering algorithm

[15] on the reduced data. Specifically, SOM maps the vectors from the high-dimensional space to a two-dimensional grid, say 30×30, while approximately preserving the clustering structure. The 2D grid encodes the density of the distribution, the cells of which are further clustered with the hierarchical clustering algorithm. We find this approach is effective as the found clusters can be well understood.

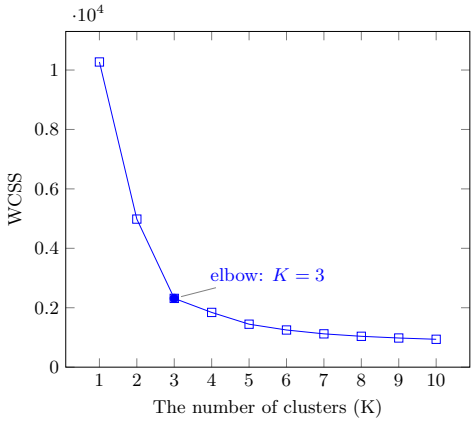


Figure 1: Relationship between WCSS and the number of clusters (K)

We use the plot of the within-cluster sum of squares (WCSS) to find the optimum number of clusters [26]. In Figure 1, the elbow criterion indicates that the optimal number of clusters should be 3, with which we derive the three clusters and their centroids as listed in Table 3.

By analyzing the centroids in Table 3 we can characterize three clusters. The users in Cluster 1 are more likely normal individual users, who publish and delete much smaller numbers of tweets. The users in Cluster 2 delete an extraordinarily large number of their tweets in a short time, which are more likely due to bulk deletions. The users in Cluster 3 publish and delete a large number of tweets. They are typically not normal users, some of whom might be spammers. We will discuss each cluster in detail in the following paragraphs.

Table 3: Centroids of clusters

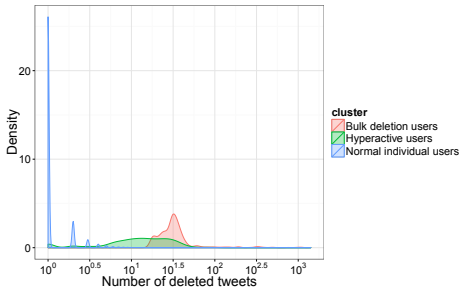
Cluster	1	2	3
Total number of users	275102	524	3001
Total number of published tweets	4.232 ± 0.621	1.786 ± 0.245	27.078 ± 8.854
Total number of deleted tweets	1.490 ± 0.565	49.916 ± 19.002	13.259 ± 13.304
Deletion Ratio	0.700 ± 0.815	29.261 ± 14.364	4.109 ± 4.059
Fano factor	0.969 ± 0.382	36.265 ± 18.982	8.441 ± 8.892
Reputation	0.544 ± 0.182	0.625 ± 0.191	0.686 ± 0.211

3.2.1 *Likely Normal Individual Users*

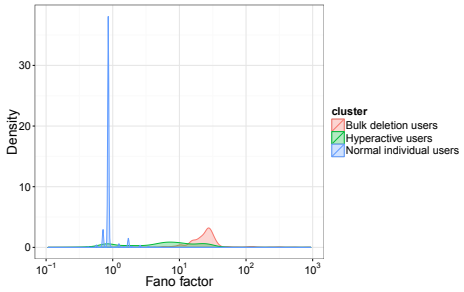
Cluster 1 is the biggest cluster. 98.735% of sample users are in this cluster. They publish and delete a much smaller number of tweets compared with users in other clusters. The number of published tweets ranges from 1 to 138, while the number of deleted tweets ranges from 0 to 10. The average numbers of published tweets and deleted tweets are only 4.2 and 1.5 respectively.

We focus on the deletion behaviors to understand the users in the clusters. Figure 2a plots the kernel density estimation [19] of the number of deleted tweets. Kernel density estimation is a non-parametric method to estimate the probability density function of a random variable. From this figure, we find that normal individual users (blue color) have the smallest number of total deleted tweets. In the contrary, both bulk deletion users (cluster 2) and "hyperactive" users (cluster 3) mainly locate at the higher numbers (10^1 to $10^{1.5}$).

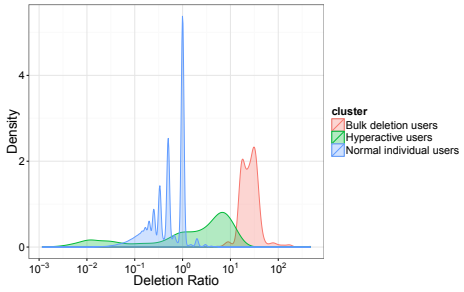
As Thomas et al. [25] described, normal users have reputation values close to 0.5, which matches the feature of users in Cluster 1.



(a) Kernel density estimation of total deletion among different clusters



(b) Kernel density estimation of Fano factor among different clusters



(c) Kernel density estimation of deletion ratio among different clusters

Figure 2: Kernel density estimation

3.2.2 *Bulk Deletion Users*

Users in Cluster 2 delete a very large number of tweets, which are probably caused by bulk deletions. Even though Twitter does not have bulk deletion mechanism, some third party applications allow bulk deletions such as Tweeteraser and TwitWipe. A few users may like to periodically clear their historical tweets in a short time as observed previously [1].

Bulk deletions can be captured by the bursting factor: Fano factor [8]. Cluster 2 has an average Fano factor of 36.265 that is far higher than those in other clusters. Figure

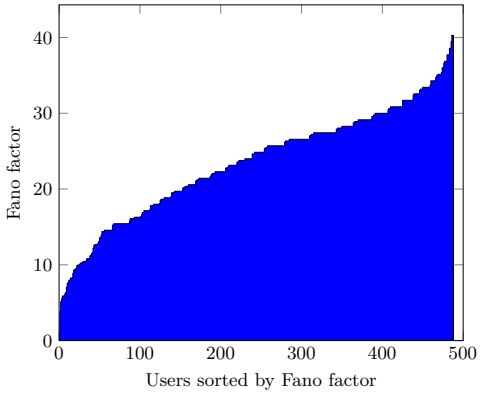


Figure 3: Sorted Users' Fano factors in Cluster 2. This figure covers 488 of 524 users with Fano factors in the range [3, 41). The remaining 36 users have very large Fano factor in the range (41, 937].

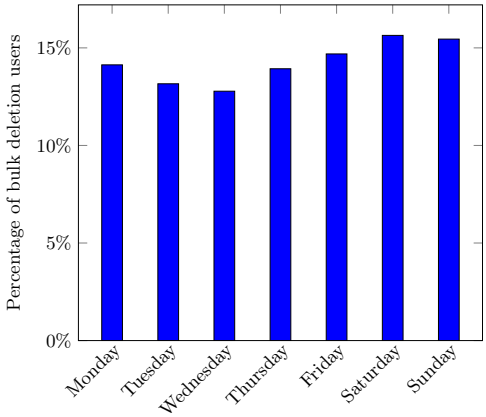


Figure 4: Bulk deletion users per week day for Cluster 2

2b shows Kernel density estimation of Fano factor in Cluster 2. The red cluster, which is bulk deletion users, has a different pattern compared with other clusters. In contrast, the average Fano factor of normal users is 0.969, indicating they have a low probability of bulk deletion. Figure 2c on deletion ratio shows a similar pattern.

By examining their deletion behaviors closely, we show the distribution of Fano factor values in Cluster 2 in Figure 3. All users have the Fano factors greater than 1. Because the highest Fano factor reaches 937, for readability, the figure cuts the value at 40. In addition, Figure 4 also gives the percentage of bulk deletion users for each week day. For each user in this cluster, we find the weekday that this user deletes the most number of tweets. By aggregating the users' peak deletion days, we get Figure 4 that gives the percentage of users conducting bulk deletion in each week day. The weekend days show slightly higher numbers as many users have more time to browse social networks during weekends.

Correspondingly, bulk deletion users often have higher deletion ratios. Cluster 2 has the highest mean deletion ratio (29.261) compared to other clusters. It is reasonable that the number of deleted tweets is larger than the number of published tweets due to cleaning history tweets. We plot Kernel density estimation of deletion ratio which is showed in Figure 2c. It is also reported that spammers regularly delete tweets to mimic regular users [14]. Twitter's spammer detection algorithm may also identify spammers and delete their tweets in batch. Since bulk deletions normally do not involve regrettable reasons, we should exclude them from our study.

3.2.3 "Hyperactive" Users

Users in Cluster 3 have an unusually large number of published tweets and a proportionally large number of deleted tweets. A deletion ratio of 4.109 and a Fano factor of 8.441 indicate that users in Cluster 3 have bulk deletion behaviors. In Figure 2c, even though "hyperactive" users' deletion ratios have a wide range from 10^{-3} to $10^{1.5}$, the peak population locates between $10^{0.5}$ to $10^{1.5}$. We randomly sample users from Cluster 3 and find that the majority comes from the two groups of users, namely "Suspicious users" and "Entertainment users".

Suspicious users. After manually examining their accounts, we find that 73.3% of sample users are fan/parody accounts of celebrities such as Frank Ocean, Channing Tatum, and Beyoncé. These users have about 320k followers and 93k followings on average, resulting the highest reputation value in the three clusters. The reason these users have so many followers might be that many fans mistakenly followed these parody users instead of authentic celebrity users. By further examining these suspicious users, we find that they share the same behavior of retweeting messages from spam users. For example, many users retweet posts containing aggressively abused reply (@) and hashtag (#) functions to catch attention. These users are likely created for propagating spam tweets.

Entertainment users. Besides suspicious users, we also find another group of users. We believe they are most likely entertainment accounts that are operated by a team or a bot program, rather than by a normal individual user. For example, "Funny Pinoy Quotes" posts tweets to inspire or entertain people, or promote products, such as: "True love and loyal friends are two of the hardest things to find". Apparently, it is not meaningful to include such a user in our study.

3.3 Summary of Clustering Users

In summary, we identify three groups of users by clustering. After further analysis, we find that the two small clusters with less than 2% of sample users contribute 11.1% published tweets, and 17.1% deleted tweets. As their deleted tweets are out of our study scope, we can eliminate a large portion of noisy deleted tweets by excluding these users. In the next section, we will continue to focus on analyzing deletion reasons of normal individual users.

4. EXPLORING, UNDERSTANDING, AND PREDICTING REGRETTABLE TWEETS

In this section, we focus on tweets that are deleted by the likely normal users that we have identified in the last section and conduct two sets of experiments. (1) The first set of experiments is to find how many percentage of tweets in deleted tweets can be identified with regrettable reasons through content analysis. We will manually label 4,000 deleted tweets and study the specific regrettable reasons and their distribution. (2) We will design content and user-

preference features based on these regrettable reasons and users' history to develop classifiers for automatically distinguishing this subset of regrettable tweets from non-deleted tweets. The result shows that we can effectively identify such regrettable tweets from non-deleted ones with the content and user-preference features.

4.1 Collecting and Cleaning Data

We select a random set of 30,000 users [3] out of the likely-normal users identified in the previous section, and apply Twitter filter streaming API [4] to continuously collect all their published and deleted tweets for two months. These deleted tweets are further filtered using the following procedure.

- We exclude retweets in our study for they are not likely regrettable tweets. A retweet can be deleted in two scenarios. i) The author of the original tweet deletes the original tweet, and all the retweets of this original tweet will be passively deleted as well. ii) The user who retweeted the original tweet decides to delete the retweet. The first scenario apparently does not involve any regrettable reason for "retweeters". Even though the user may actively delete retweets for any regrettable reason, the potential damage of the published retweets might be little to retweeters - on the contrary, more severe damage on the original author. For these reasons, we can exclude them from our data collection.

- In some cases, Twitter users delete tweets and repost similar ones due to typography, rephrasing, or missing mentions/hashtags [1]. For simplicity, we use "rephrasing" to cover all these reasons. These deletions are apparently not caused by regrets and thus should be excluded from our dataset. To eliminate such tweets, we built a classifier to automatically identify them. The training data is created based on a set of randomly selected deleted tweets. For each selected deleted tweet, the tweets published in the subsequent one-hour period by the same author are also retrieved with the Twitter API and manually examined. If this deleted tweet is *very similar* to any subsequently posted tweet, we label it as a deletion caused by rephrasing. The critical problem is to define the similarity measure and find the appropriate threshold for identifying rephrasing. We tested four candidate measures: Jaccard distance, edit distance, Levenshtein ratio, and time difference (in minutes). These measures are formulated as four features for each pair of tweets. In total, we generate a dataset of 58 positive tweet pairs and 512 negative tweet pairs. We then apply Decision Tree classifier J48 [22] with 10-fold cross-validation to identify the most important features and thresholds. It turns out that J48 uses only Levenshtein ratio to build the decision tree with a threshold of 0.6818, and the F1-measure is 99.8%, which works perfectly for our purpose. We applied this classifier to exclude deleted tweets caused by rephrasing.

[3]Fano factor=1 is used as a conservative threshold [8] to detect and remove bulk-deletion users, who did not conduct bulk deletions in the seven-day window.

[4]`https://dev.twitter.com/streaming/reference/post/statuses/filter`

Table 4: Statistics on the collected tweets of 30,000 users

Total number of users	30,000
Total number of published tweets	17,587,816
Total number of non-deleted tweets	14,325,871
Total number of deleted tweets	3,261,945
Number of users who deleted at least 1 tweet	26,543
Number of users who posted at least 1 tweet	28,778

In total, we collected about 17.6M published tweets, including 14.3M non-deleted tweets and 3.2M deleted tweets, from 28,778 distinct users. Table 4 introduces the details.

Table 5: Examples of deleted tweets with unsure reasons. Links and @s are masked with xxxxxx to preserve privacy.

You'll,let me known
The world as we know it is over #whynik http://t.co/xxxxxx
my birthdays in 5 more days
@xxxxxx man, when ya doing another show in ny?
Yasss my hair came in @xxxxxx #bleuribbon
would any sophomores buy this? needs to know http://t.co/xxxxxx
Yay! My cousins are visiting me next weekend for the @Dodgers game!
A guy in my clan somehow managed to do this :3 http://t.co/xxxxxx
Feeling blessed! #godisgood
Beyond over it

4.2 Mining and Understanding Reasons for Regrettable Tweets

After excluding retweets and rephrasing tweets, we randomly sample and manually examine 4,000 deleted tweets from the second month to understand the reasons for regrettable tweets. Consulting the regrettable reasons identified by Wang et al. [29] for Facebook posts, three annotators read the tweet contents and manually annotate each tweet with the corresponding possible regrettable reason. Later, these regrettable reasons are grouped into ten major categories (negative sentiment, cursing, sex, alcohol, drug, violence, health, racial and religion, job, relationship). If annotators are not able to identify any regrettable reason for a tweet, we call the tweet an *unsure* tweet; otherwise, we call it an *regrettable* tweet. The agreement measure [13] "Fleiss' kappa" is 0.62 among three annotators, which is considered a substantial agreement. In the end, the disagreed examples are discussed by the annotators and unified. We focus on the content-identifiable regrettable tweets because their meaning can be easily captured by readers, which creates more damages compared with the ones without identifiable reasons.

Based on only tweet contents, only about 18% of the tweets can be labeled with specific regrettable reasons, while the remaining 82% cannot be explained by simply examining the tweet contents. For example, the contents of the following deleted tweets do not indicate any regrettable reason: "Lol I love my dad", and "Captain America with my boo through last night it was gooooooddd!" Bauer et al. [3] has a similar discovery: 6.0% of the deleted Facebook posts are so important that they need to be deleted, while 89.0% of the deleted Facebook posts were deleted for unknown reasons. Their discovery corroborates ours that only about 18.0% of the tweets have content-based identifiable regrettable reasons.

Some unsure tweets are due to disguised contents. About 18% of unsure tweets contain links, and 76% of these links are user-posted photos that may contain sensitive information. Unfortunately, these photos had been deleted by Twitter, and we cannot retrieve these photos to figure out regrettable reasons. Table 5 contains more examples, the contents of which do not provide sufficient clues for deletion. We focus on the content-identifiable regrettable tweets because their meaning can be easily captured by readers, which creates more damages compared with the ones without identifiable reasons.

The distribution of the ten identifiable reasons is highly imbalanced as shown in Figure 5. Cursing, relationship, sex, and negative sentiment are four dominating reasons, covering about 85% of the identifiable tweets. The other reasons (alcohol, drug, health, job, violence, racial and religion) contribute to the remaining 15%. In addition, a tweet might be labeled for multiple reasons. For example, "Ugh I hate working till 1am!! I always come home full of energy" is labeled by both job and negative sentiment.

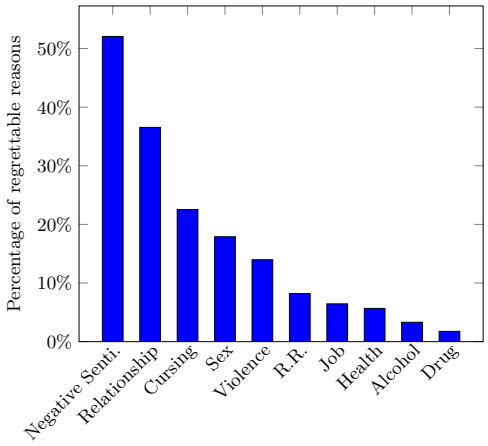


Figure 5: Distribution of regrettable reasons (R.R. represents Racial and Religion).

4.3 Automatically Identifying Regrettable Tweets Based on Contents and User Preferences

Due to the significance of the regrettable tweets, we try to build classifiers to automatically identify regrettable tweets based on their contents and user preferences in publishing/deleting such tweets.

Feature Design. We design two sets of features for classifier modeling: content-based and user-preference features.

- ***content-based.*** Based on the ten identified reasons, we define ten features correspondingly for each tweet. These features are all binary features: 1 indicates that the corresponding reason is present, and 0 otherwise. For negative sentiment feature, we apply SentiStrength [24] to detect the sentiment of each tweet. The outputs from SentiStrength are two real values representing the strengths of the tweet being positive and negative. If the strength has "negative" > "positive", we set the feature to 1, otherwise 0. For other features, to determine whether a reason is present, we define a function $f_i(t)$ for the i-th feature, where t represents a bag of words of a tweet after removing stopwords and stemming with tool NLTK [4]:

$$f_i(t) = \begin{cases} 1 & \text{if } t \cap S_i \neq \emptyset \\ 0 & \text{otherwise} \end{cases}$$

where S_i is the set of keywords related to the feature.

Multiple methods are used for defining the keyword set S_i for different features. For cursing feature, we adopt a comprehensive list of cursing words collected by Wang et al. [28]. For defining each of the remaining features, we start with a seed word that is the name of the reason category and then expand it by looking up their synonyms and related words in a bootstrapping fashion. The sources for word expansion include WordNet [9] and Urban Dictionary [20], where Urban Dictionary is a rich source for understanding tweet language because it includes many slangs from the Internet.

For example, "alcohol" is the seed word for the category *alcohol*, "drunk" is one of related words to "alcohol" in Urban Dictionary, and "liquor" is one of the synonyms of "alcohol" in WordNet.

Table 6 lists the definition and some sample words of each feature. The complete list of lexicons for generating these features can also be downloaded from (`http://bit.ly/1LQD22F`).

With these sets of keywords, we can derive the content-based features for each tweet. Let's take the previous sentence, "Ugh I hate working till 1am!! I always come home full of energy", to show the resultant feature vector. This tweet is clearly about a job complaint. The job-related keyword is "working". Besides, SentiStrength returns a negative sentiment. Therefore, the job and negative sentiment features are set to 1, while other features are set to 0.

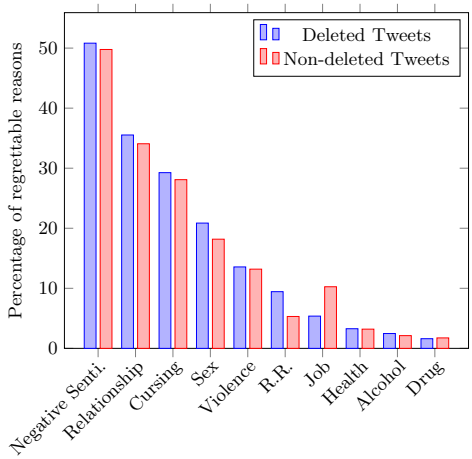


Figure 6: Comparison of the feature distributions for deleted tweets and non-deleted tweets that contain the regrettable contents. R.R.: Racial and Religion

- ***user-based*** Users' historical publishing and deleting patterns can be explored to provide personalized features, indicating their preferences in publishing/deleting sensitive tweets. For example, users who frequently

Table 6: Feature and Lexicon Design

Features	# of Words in Lexicon	Examples
Negative Sentiment	apply SentiStrenge[24]	N/A
Relationship (strong emotions about relationship)	95	breakup, dating, lover, ...
Cursing (containing cursing words)	796	ass, bitch, fuck, ...
Sex (any sexual activity and orientation)	148	blowjob, cock, dick, ...
Violence (contents about violence and war)	149	attack, fight, kill, ...
Racial, Religion (discrimination about race and religion)	91	nigger, negro, coon, ...
Job (anything about work and company)	64	fired, unemployed, loser, ...
Health (contents contain personal complains about diseases)	126	fat, disorder, ugly, ...
Alcohol (drunk activities or feelings)	95	drunk, hangover, vodka, ...
Drug (drug products or using drugs)	54	cocaine, marijuana, weed, ...

publish tweets that contain cursing words may not regret about publishing another cursing tweet. However, for users who have no history of cursing tweets, a tweet containing cursing words may raise a red flag. We generate 12 such features by analyzing one-month history tweets of the targeted users. (1) For each of these ten identifiable regrettable reasons, we calculate ratios of deleted tweets to the published (deleted and non-deleted) tweets that contain the specific regrettable reason. For example, if the numbers of deleted and non-deleted tweets that contain cursing words is n_i and m_i, respectively, the ratio of deleted cursing tweets to the published cursing tweets is $n_i/(m_i + n_i)$. In this way, we generate the first ten user-preference features based on the users' tweeting and deleting history. (2) The last two ratio-based features are defined as follows. The ratio of regrettable deletions to the total deleted tweets that contain any of the ten reasons is $\sum_i n_i/(\sum_i (m_i + n_i)$. The ratio of the total deleted tweets N to all published (N deleted + M published) tweets is $N/(N + M)$.

Classifier Design and Evaluation. Our problem is to predict whether an input tweet containing any of the regrettable reasons is likely to be deleted or kept. To build a classifier for this prediction, we randomly collect training data from the second-month tweets of users who have published at least one deleted tweet and one non-deleted tweet. The training data consists of 10,000 deleted tweets and 10,000 non-deleted tweets, each of which contains at least one of the ten reasons, from the same set of users (5,000 users). For each user, we collect 2 deleted tweets, and 2 non-deleted tweets respectively. This design allows us to focus on the specific problem: once the likely regrettable content is created, predict whether the author will delete it or not.

We model this problem as an information retrieval problem: finding the deleted (thus, implicating regrettable) tweets from the mixed set. The precision, recall, and F1 measures are used to evaluate the quality of retrieving the deleted tweets. We use 10-fold cross-validation to evaluate four different classifiers (Naive Bayes, SVM, J48, AdaBoost).

Table 7: Classifiers trained with our proposed features. NB: Naive Bayes.

		Content-only	Content+User-history
NB	Precision	0.552 ± 0.031	$\mathbf{0.775 \pm 0.046}$
	Recall	0.349 ± 0.078	0.486 ± 0.055
	F1-Score	0.427 ± 0.043	0.598 ± 0.035
SVM	Precision	0.536 ± 0.041	0.753 ± 0.042
	Recall	0.478 ± 0.049	0.626 ± 0.054
	F1-Score	0.505 ± 0.041	0.683 ± 0.034
J48	Precision	0.537 ± 0.027	0.711 ± 0.072
	Recall	0.593 ± 0.030	$\mathbf{0.716 \pm 0.081}$
	F1-Score	0.563 ± 0.019	$\mathbf{0.714 \pm 0.081}$
AdaBoost	Precision	0.541 ± 0.055	0.731 ± 0.048
	Recall	0.434 ± 0.077	0.696 ± 0.068
	F1-Score	0.482 ± 0.048	0.713 ± 0.055

Table 8: Top 10 features in different models. R.R.: racial and religion.

	Content-only	Content+User-history
1	job	total deletion ratio
2	R.R.	total regrettable deletion ratio
3	sex	negative senti. deletion ratio
4	drug	cursing deletion ratio
5	alcohol	sex deletion ratio
6	negative senti.	R.R. deletion ratio
7	health	job deletion ratio
8	violence	alcohol deletion ratio
9	relationship	violence deletion ratio
10	cursing	drug deletion ratio

The third column of Table 7 shows the result of using only the ten content features to train four classifiers. The precisions are all around 0.5, which indicates that we cannot distinguish the two sets of tweets by using only the content features. This result suggests that the same regrettable content can be either kept by one set of users or deleted by another set of users. Thus, using only content features, we cannot effectively distinguish these two sets of tweets. Figure 6 shows the distributions of the features in deleted tweets and non-deleted tweets, respectively. Their distributions are very close in all the different reason categories.

The fourth column of Table 7 shows the result of applying both content-based and user-history-based features. We find both precision and recall are significantly improved for all classifiers. Overall, Naive Bayes achieves the highest precision of 0.775. J48 has the highest recall of 0.716, and the highest F1-score of 0.714. We also compare the feature importance in these two sets of modeling. Table 8 lists the top 10 ranking features in each set according to J48 output in Weka. All history features come into the top positions, which indicates that user preferences play important roles in modeling classifiers. Although all top-ranked features are user-history features, the content-based features cannot be simply removed from modeling. Our study shows that the F1 measure of J48 is reduced to 0.549, if we use only the user-history features in modeling. This result supports our understanding that different users may have different perceptions on "regrets"; whether deleting possibly regrettable tweets or not is a personal choice. By appropriately modeling the regrettable contents and users' history activities, we show that the personal preferences can be effectively captured.

To show the advantages of the proposed features, we also train models with common NLP features, such as Unigram, Bigram, and POS (part of speech) features, which are gen-

erated with the TagHelper tools [5]. Unigram (and Bigram) features are derived from initial processing, which are ranked and selected with the Information Gain (IG) method [18]. It turns out the threshold IG=0.004 gives us the best performance for the Unigram features. However, Bigram features failed in classification modeling due to the sparsity of feature space. As a result, the learned classifiers label almost all testing examples with "regrettable tweets", giving ∼100% recall and ∼50% precision for the balanced training data. We thus exclude them from the report. Table 9 shows the performance of these features in developing classifiers. The Unigram+POS features give best precision 0.554 for AdaBoost classifier, and all the results are significantly worse than the classifiers trained with our proposed features.

Table 9: Classifiers trained with NLP features. NB: Naive Bayes.

		Unigram	Unigram+POS
NB	Precision	0.529 ± 0.031	0.533 ± 0.046
	Recall	0.460 ± 0.078	0.490 ± 0.055
	F1-Score	0.492 ± 0.043	0.511 ± 0.035
SVM	Precision	0.533 ± 0.041	0.533 ± 0.042
	Recall	0.433 ± 0.049	0.485 ± 0.054
	F1-Score	0.478 ± 0.041	0.508 ± 0.034
J48	Precision	0.533 ± 0.027	0.534 ± 0.072
	Recall	0.393 ± 0.030	0.474 ± 0.081
	F1-Score	0.452 ± 0.019	0.502 ± 0.081
AdaBoost	Precision	0.523 ± 0.055	0.554 ± 0.048
	Recall	0.463 ± 0.077	0.333 ± 0.068
	F1-Score	0.491 ± 0.048	0.416 ± 0.055

5. RELATED WORK

Social media data is becoming more and more popular for research with different focuses, e.g., privacy [12], search ranking [7] and sentiment analysis [10]. Here, we focus on the literature that is most relevant to our work.

Posting regrettable content is not uncommon in social media. Pew Internet project survey [17] shows that 11.0% of social network users had the experience of posting content that they regret later. Some studies also show that people sometimes disclose *personally embarrassing information* in Twitter when they interact with friends and forget that tweets are publicly visible [2, 5]. Besides regrets, posting inappropriate things can have other serious consequences, including getting fired by companies [6].

A few studies adopt survey-oriented approaches to study this phenomenon by asking users to recall recent regret posts and answering designed questions. Wang et al. [29] study different types of regrettable posts and reasons why people post them in Facebook. Sleeper et al. [23] conduct a similar study by comparing regret incidents in Twitter and these in real-world conversations, regarding types of regret and how people become aware of these incidents.

Instead of interviewing users and asking them to recall recent regret incidents, few studies take an "in-field" approach by collecting users' deleted posts from social media. Almuhimedi et al. [1] examine aggregated properties between deleted tweets and non-deleted tweets. No substantial difference is found between deleted tweets and non-deleted tweets, except a few dimensions, e.g., posting clients, sentiment vocabularies. Bauer et al. [3] collect deleted Facebook posts and ask users how important it is to delete these posts. They find out that 6.0% of deleted posts are so important that these posts need to disappear from Facebook while 89.0% of the deleted posts were deleted for not-at-all important reasons. This discovery corroborates that in our study on tweets: only a small portion of tweets were deleted because of regrets, and deleted tweets are not necessarily regrettable tweets. Petrovic et al. [21] apply machine learning techniques to classify deleted tweets from non-deleted tweets, and as we mentioned above, their target, deleted tweets, are different from our target, regrettable tweets.

There are also studies on understanding the characteristics of people who like to delete social posts. Boyd and Crawford illustrated that teens like to delete tweets to avoid the negative consequence instead of setting up the privacy control [6]. Tufekci found out that women are more likely to delete Facebook posts than men for privacy protection [27].

To summarize, existing studies on regret posts in social media [1, 3, 23, 29] have focused on exploring different aspects of regrettable tweets, e.g., regret types and reasons, but left the problem of how to computationally tackle regret posts unexplored. To the best of our knowledge, our study is the first data-driven attempt to automatically identify regret posts in social media for normal users.

6. CONCLUSION AND FUTURE WORK

Inappropriate tweets, once published, can cause permanent damages to the author's reputation or privacy. However, it is very challenging to identify such tweets before authors publish them. In this paper, we address this problem by studying the regrettable tweets that can be identified solely based on their contents and users' publishing/deletion preferences.

Our study focuses on regrettable tweets created by normal individual users who are identified by a user clustering method. Based on five features describing each user, we can identify three clusters of users: likely normal users, bulk deletion users, and "hyperactive" users, after excluding the verified users. We find that about 98.7% of sample users are likely normal users, while the 1.3% of users from the other two clusters contribute 17% deleted tweets and 11% published tweets.

The tweets from likely-normal users are further cleaned by removing retweet deletions and rephrasing deletions. Then, we manually analyze 4,000 cleaned tweets to identify ten candidate regrettable reasons. Based on these reasons and users' historical publishing and deleting patterns, we design a set of features to distinguish regrettable tweets from non-deleted tweets. We show that these features are more effective than the generic NLP features in constructing classifiers.

Future Work. We outline some future work that will further enhance the current study. In particular, we will explore and develop more features for more accurately locating likely normal individual users, and distinguishing regrettable tweets from non-deleted tweets. We also plan to develop a prototype system that applies the developed classifiers to detect and flag the potentially regrettable tweets, allowing users to withdraw them before publishing. It will be valuable to explore the users' implicit feedback from the prototype system.

[5] http://www.cs.cmu.edu/~cprose/TagHelper.html
[6] http://tinyurl.com/k9mlxsw

7. ACKNOWLEDGMENTS

This material is based upon work partially supported by the National Science Foundation under Grant 1245847 and a DAGSI/AFRL Grant.

8. REFERENCES

[1] H. Almuhimedi, S. Wilson, B. Liu, N. Sadeh, and A. Acquisti. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proceedings of CSCW*, pages 897–908. ACM, 2013.

[2] J. Bak, C. Lin, and A. H. Oh. Self-disclosure topic model for classifying and analyzing twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1986–1996, 2014.

[3] L. Bauer, L. F. Cranor, S. Komanduri, M. L. Mazurek, M. K. Reiter, M. Sleeper, and B. Ur. The post anachronism: The temporal dimension of facebook privacy. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*, pages 1–12. ACM, 2013.

[4] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

[5] D. Boyd. Making sense of privacy and publicity. `http://www.danah.org/papers/talks/2010/SXSW2010.html`, 2010. Accessed: 2014-12-02.

[6] D. Boyd and K. Crawford. Six provocations for big data. In *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011.

[7] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: improving recency ranking using twitter data. In *Proceedings of the 19th international conference on World wide web*, pages 331–340. ACM, 2010.

[8] U. Fano. Ionization yield of radiations. ii. the fluctuations of the number of ions. *Physical Review*, 72(1):26, 1947.

[9] C. Fellbaum. *WordNet*. Wiley Online Library, 1998.

[10] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

[11] S. Guha, R. Rastogi, and K. Shim. Cure: an efficient clustering algorithm for large databases. In *ACM SIGMOD Record*, volume 27, pages 73–84. ACM, 1998.

[12] S. Guo and K. Chen. Mining privacy settings to find optimal privacy-utility tradeoffs for social network services. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 656–665. IEEE, 2012.

[13] K. L. Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[14] D. Irani, S. Webb, C. Pu, and K. Li. Study of trend-stuffing on twitter through text classification. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[15] A. Jain, M. Murty, and P. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31:264–323, 1999.

[16] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[17] M. Madden. Privacy management on social media sites. Pew Research, `http://tinyurl.com/d9xwzsh`, 2012. Accessed: 2014-11-03.

[18] T. M. Mitchell. Machine learning. wcb, 1997.

[19] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, pages 1065–1076, 1962.

[20] A. Peckham et al. *Urban dictionary: Fularious street slang defined*. Andrews McMeel Publishing, 2009.

[21] S. Petrovic, M. Osborne, and V. Lavrenko. I wish i didn't say that! analyzing and predicting deleted messages in twitter. *arXiv preprint arXiv:1305.3107*, 2013.

[22] J. R. Quinlan. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann, 1993.

[23] M. Sleeper, J. Cranshaw, P. G. Kelley, B. Ur, A. Acquisti, L. F. Cranor, and N. Sadeh. I read my twitter the next morning and was astonished: a conversational perspective on twitter regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3277–3286. ACM, 2013.

[24] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[25] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 243–258. ACM, 2011.

[26] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[27] Z. Tufekci. Facebook, youth and privacy in networked publics. In *Proceedings of International AAAI Conference on Weblogs and Social Media*, pages 36–7, 2012.

[28] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth. Cursing in english on twitter. In *Proceedings of CSCW*, pages 415–425. ACM, 2014.

[29] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. I regretted the minute i pressed share: A qualitative study of regrets on facebook. In *Proceedings of the Symposium on Usable Privacy and Security*, page 10. ACM, 2011.