Identifying Regrettable Messages from Tweets

Lu Zhou[*], Wenbo Wang[+], Keke Chen[*]
[*]Data Intensive Analysis and Computing Lab, [*+]Kno.e.sis Center
Department of Computer Science and Engineering
Wright State University, Dayton, OH 45435
kbzhoulu@gmail.com, wenbo@knoesis.org, keke.chen@wright.edu

ABSTRACT

Inappropriate tweets may cause severe damages on the authors' reputation or privacy. However, many users do not realize the potential damages when publishing such tweets. Published tweets have lasting effects that may not be completely eliminated by simple deletion, because other users may have read them or third-party tweet analysis platforms have cached them. In this paper, we study the problem of identifying regrettable tweets from normal individual users, with the ultimate goal of reducing the occurrences of regrettable tweets. We explore the contents of a set of tweets deleted by sample normal users to understand the regrettable tweets. With a set of features describing the identifiable reasons, we can develop classifiers to effectively distinguish such regrettable tweets from normal tweets.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering

Keywords

Social Media, Regret, Deleted Tweets, Classification

1. INTRODUCTION

Twitter is a popular online social network, where users can post their thoughts, share photos, and have conversions with others publicly in a real-time fashion. While it is convenient to communicate with others on Twitter, people sometimes mistakenly post tweets that they will regret later. For example, people may feel inappropriate after venting out frustrations about friends or managers. Moreover, people may feel awkward after posting a secret about themselves or other people unconsciously. Table 1 lists a few tweets that were deleted because of regrets after posting. On Twitter, most tweets are public and can be rapidly spread. Inappropriate tweets may be read and spread by lots of people, before authors delete them. As a result, tweet deletion does not eliminate the risk of privacy disclosure or self-image destruction.

Table 1: Sample regrettable tweets

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

Inspired by the observation that users will delete regrettable tweets when they start regretting, we start with collecting users' deleted tweets and manually label regrettable tweets based on their contents. We study how to distinguish such regrettable tweets from normal (i.e., not deleted) tweets as a first step towards the goal of identifying all regrettable tweets. Being aware that feeling regret is context-dependent, in this paper we will focus only on tweet content to explore the effectiveness of content-based features to identify regrettable tweets. Based on different categories of regret reasons, we apply a bootstrapping approach to construct lexicons by retrieving relevant words of seed words from WordNet, Urban Dictionary, and other sources [2], which will be used to extract regret-specific features. Experimental results show that the proposed features can be used to effectively distinguish regrettable tweets from normal ones.

2. DISTINGUISHING REGRETTABLE TWEETS FROM NORMAL TWEETS

We select a random set of 553 normal Twitter users (excluding non-normal users such as spammers, corporate users, and celebrity users) and apply Twitter filter streaming APIs to continuously collect all their published and deleted tweets for one month. We exclude retweets in the dataset to reduce duplication. Consider that Twitter users sometimes delete tweets and repost similar ones (a.k.a, rephrasing)[1], we also exclude tweets that were deleted because of the rephrasing purpose. The "rephrasing tweets" are automatically identified as follows. For each deleted tweet in the training dataset, we examine the tweets published in the subsequent one hour by the same author. If this deleted tweet is *very similar* to any tweet in the subsequent tweets, we label it as a deletion caused by rephrasing. We applied three string similarity measures:Jaccard distance, edit distance, and Levenshtein ratio, and the time difference feature (in minutes), to train a J48 Decision Tree classifier on a dataset of 58 rephrasing pairs and 512 non-rephrasing pairs. The F1-measure with 10-fold cross validation reaches 99.8%, which indicates highly reliable identification of rephrasing tweets.

To understand the reasons for regrettable tweets, we manually label 4,000 randomly sampled deleted tweets after preprocessing, based on the possible regret reasons extended

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742735.

from Wang et al. [3]. We were able to identify the specific regrettable reasons for only 700 (17.5%) based on tweet contents, while the reasons for the remaining 82.5% cannot be explained by simply reading the content of the tweets - we name them "unsure tweets". For example, the content of the following deleted tweets does not indicate any regrettable reason: "Lol I love my dad." and "Captain America with my boo through last night it was gooooodddd!" Interestingly, 18.0% of unsure tweets contain links, 75.7% of these links are photos posted by users which may contain sensitive information. Unfortunately, we are unable to trace these deleted photos to figure out reasons.

The distribution of the content-identifiable reasons is highly imbalanced as shown in Figure 1: cursing, relationship, sex, and negative sentiment are 4 dominating reasons, covering about 85.0% of regrettable tweets, while the other reasons (alcohol, drug, health, job, violence, racial and religion) cover only about 15.0% of regrettable tweets. A tweet might be labeled with 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.

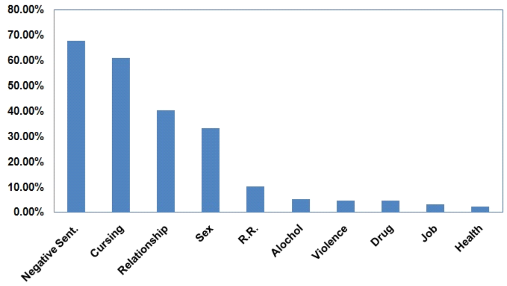

Figure 1: Distribution of regrettable reasons (R.R.: Racial and Religion).

We design ten features correspondingly for the above-mentioned reasons. These features are all binary features: 1 to represent the corresponding reason is presented, and 0 otherwise. Except the negative sentiment feature, to determine whether a reason is presented, we define a function $f_i(t)$ for the i-th feature, where t represents a bag words of the tweet after removing stopwords. We apply SentiStrength (sentistrength.wlv.ac.uk) to extract the negative sentiment feature.

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

where S_i is the set of keywords for the feature. Multiple methods are used to define the word set S_i of different features. For the cursing feature, we adopt a comprehensive list of cursing words collected by Wang et al. [2]. To define each of the remaining features, we start with a few seed words related to that particular reason and then expand it by looking up their synonyms and related words from WordNet and Urban Dictionary in a bootstrapping fashion. While WordNet has a good coverage for formal words, Urban Dictionary includes a lot of Internet slangs, a rich source for understanding the tweet language. For example, "alcohol" is a seed word for reason alcohol, and "drunk" is one of related words to "alcohol" in UrbanDictinary.

We use the manually labeled 700 regrettable tweets from 553 distinct users as positive examples. To achieve balanced training data, for each of the 553 users, we count the number of regrettable tweets first, and then randomly select the same number of normal tweets (exclude retweets) of this user from the same window as negative examples. Collecting normal tweets in this way can avoid possible biases brought by different users.

Table 2: Classifiers trained with different types of features. NB: Naive Bayes.

		Our Features	Unigram	Unigram+POS
NB	Precision	0.840 ± 0.031	0.761 ± 0.046	0.609 ± 0.039
	Recall	0.789 ± 0.078	0.592 ± 0.055	0.802 ± 0.031
	F1-Score	0.812 ± 0.043	0.664 ± 0.035	0.691 ± 0.025
SVM	Precision	0.796 ± 0.041	0.765 ± 0.042	0.743 ± 0.046
	Recall	0.850 ± 0.049	0.616 ± 0.054	0.631 ± 0.052
	F1-Score	0.822 ± 0.041	0.681 ± 0.034	0.681 ± 0.032
J48	Precision	0.774 ± 0.027	0.836 ± 0.072	0.709 ± 0.053
	Recall	**0.940 ± 0.030**	0.408 ± 0.081	0.559 ± 0.047
	F1-Score	**0.849 ± 0.019**	0.545 ± 0.081	0.623 ± 0.032
AdaBoost	Precision	**0.858 ± 0.044**	0.828 ± 0.048	0.781 ± 0.052
	Recall	0.669 ± 0.067	0.514 ± 0.068	0.533 ± 0.065
	F1-Score	0.751 ± 0.054	0.631 ± 0.055	0.631 ± 0.053

In Table 2, we summarize the 10-fold cross-validation results with different types of features and classifiers. Since it is more like a retrieval problem to identify regrettable tweets in a set of tweets, we use precision, recall, and F1-Score to evaluate the results. Our proposed 10 features are compared to the common NLP features (Unigrams, Bigrams, and POS) that are extracted with TagHelper (http://www.cs.cmu.edu/∼cprose/TagHelper.html). With our features AdaBoost has the highest precision of 0.858; J48 has the highest recall of 0.940 and the highest F1-score of 0.849. Naive Bayes and AdaBoost provide marginally better precision than the other two, while J48 gives statistically significantly better recall than the other three. The result shows that the proposed features work effectively on identifying regrettable tweets.

Thousands of Unigram (and Bigram) features are derived from initial processing, which are ranked and selected with the Information Gain (IG) method. It turns out the threshold IG=0.004 gives us the best performance for the Unigrams features. However, Bigram features totally failed in classification modeling due to the extreme sparsity of feature space - the classifiers label almost all examples with "regrettable tweets", resulting ∼100% recall and ∼50% precision for the balanced training data. The Unigram features give good precision 0.836 and 0.828 for J48 and AdaBoost classifier respectively, but their recall 0.408 and 0.514 are significantly lower than our best result. We also find that the additional POS features do not help much in classification modeling.

Acknowledgments This material is based upon work partially supported by the National Science Foundation under Grant No. 1245847.

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

[3] 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.