Classification of Private Tweets using Tweet Content

Qiaozhi Wang*, Jaisneet Bhandal*, Shu Huang†, and Bo Luo*
*Department of EECS, The University of Kansas, Lawrence, KS; †Microsoft, Seattle, WA, USA
qzwang@ku.edu, Bjaisneet@ku.edu, shuang@microsoft.com, bluo@ku.edu

***Abstract*—Online social networks (OSNs) like Twitter provide an open platform for users to easily convey their thoughts and ideas from personal experiences to breaking news. With the increasing popularity of Twitter and the explosion of tweets, we have observed large amounts of potentially sensitive/private messages being published to OSNs inadvertently or voluntarily. The owners of these messages may become vulnerable to online stalkers or adversaries, and they often regret posting such messages. Therefore, identifying tweets that reveal private/sensitive information is critical for both the users and the service providers. However, the definition of sensitive information is subjective and different from person to person. To develop a privacy protection mechanism that is customizable to fit the needs of diverse audiences, it is essential to accurately and automatically classify potentially sensitive tweets. In this paper, we make the first attempt to classify private tweets into 14 categories, such as alcohol & drugs, family information, etc. We model tweet semantic with term distribution features as well as users' topic-preferences based on personal tweet history. Experiments show that our method can boost classification accuracy compared with the well-known Bag-of-Words and tf-idf methods.**

I. INTRODUCTION

As the most popular open microblog platform, Twitter has 310M monthly active users. With this new socialization method, users post tweets about every aspect of their daily life, ranging from professional and career development to personal and family updates. For most of the users, their tweets are intended for friends that follow them. However, Twitter is an open platform that everything posted to it are accessible to the public, which makes users very vulnerable – private or sensitive information may be accidentally disclosed, even in tweets about trivial daily activities. [1] have shown that regret-tweets are very common. Most of them involve sensitive content and rich semantics, such as alcohol and illegal drug use, sex, religion and politics, personal and family issues, etc. Although users may imagine the audience before posting tweets, imagining is difficult, and the imagined audience is often inconsistent with the actual audience [2], especially consider that Twitter does not provide access control over tweets. Moreover, [3], [4] use information aggregation to discover user identities and recover user attribute information from large amount of seemingly little and harmless data. With the development of data mining and user attribute extraction approaches, it is critical to automatically identify potentially sensitive tweets and alert users before they are posted.

However, the degree of sensitiveness and privacy is a subjective perception, which differs from person to person. For instance, some users are more conservative about health-related issues, while some others might be more protective on work-related information. That says, in developing a privacy protection mechanism for online social networks, we cannot use a uniform measure of privacy for all users. Classification of private tweets is necessary for customized privacy protection – we can alert users for the pre-set types of private information they want to protect. Meanwhile, we consider private tweet identification (automatically identify if a tweet contains private information) and private tweet classification as dual-problems. Progress towards one of them will eventually benefit the other.

In this paper, we assume that a set of potentially sensitive tweets, with controversial or private content, have been identified already. Our goal is to properly classify these sensitive tweets into 14 pre-defined categories, as shown in Table 1. To the best of our knowledge, this is the first attempt to classify microblog messages into a comprehensive set of potentially sensitive categories. Our baseline approach demonstrates relatively good performance. We further propose user topic preference in our classification model, which improves accuracy from 78.4% to 81.8%.

In the rest of paper, we introduce the related work, details of our data labeling and classification approach, experiment results, and analysis of the performance.

II. RELATED WORK

Since Twitter is an ideal platform with numerous tweets covering every aspect of daily life, tweets classification attracts lots of researchers to study. For example, [5] classifies tweets into 18 general trending topics, such as sports, politics, etc. Combining network-based information with text-based features, the accuracy of the result is improved. [6] separate tweets into three classes based on the time-dependency of tweets, which can filter out "expired" tweets during browsing. These classification models give us some inspiration, but still quite different from our model which intends to classify private tweets to pre-defined 14 categories.

The papers most related to our work are about regret tweets. [7] explore the features of deleted tweets and try to predict whether a tweet will be deleted later. The features they used for classification are ten topics about sensitive information and the sentiment of each tweet. Compared with our work, their topics mainly considered extremely sensitive topics, such as curses, drugs, etc. And tweets' topics are judged by checking the existence of any word in topics' word-bag, whereas the use of classifiers. Another paper, [8], classifies whether the tweet about three topics – vacation, drinking, and disease,

This work was supported in part by the US National Science Foundation under NSF CNS-1422206 and NSF DGE-1565570.

TABLE I
TWEET CATEGORIES

Category	Example
Health & Medical	Seriously starting to regret this surgery...
Work	Nothing like being at work at 6 am! #ineedanewjob
Drugs & alcohol	Nothin' beats whiskey & coke
Obscenity	I hate fucking a skinny bitch!!! #ineedbigass
Religion	Be strong and take heart and wait for The Lord.
Politics	If Obama wins I'm becoming a communist!
Racism	I hate black people and gay people as well
Family & Personal	Grandma and papa flying in tonight!!
Complaints, Curses	This spring break was kind of trash
Relationship	I have no problem flaunting my relationship.
Sexual Orientation	Tayler just admitted to me that she is bisexual...
Travel	I wish I could just leave and go on a long road trip
School life	3 hour class can suck my balls
Entertainment	Watching bad girls club while I wait for class #noshame

is sensitive or not. Tweets are first filtered to three topics via keyword matching. Then for each topic, specific features are used, such as time information in vacation, to classify sensitiveness. This paper demonstrates the identification of sensitive tweets should be based on different topics, which corresponds to our motivation. But we believe more topics have the potential containing sensitive content. Although both papers did not directly classify tweets to different topics, they showed the necessity of classifying topics before identification.

Due to the shortness of the tweet – up to 140 characters, the classification of tweets is full of challenges, and many articles try to improve the accuracy not only relying on Bag-of-Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF), but also introducing more features to boost the classifier. [9] introduces 8 features based on the content of tweets, which outperforms the BoW on classifying generic topics– news, sports, etc. When trying to classify "check-in" tweets to different locations, [10] uses time-period and users' check-in history as boosting features. Inspired by this boosting method, we introduce users' topic-preferences as our boosting features to improve classification performance.

III. APPROACH

In this paper, we define 14 topics related to privacy content, which is shown in Table 1. Our system can be separated into the following parts: data collection, labeling, data normalization, feature selection and classification. The first three steps can be seen as preparation for data set, which will be introduced in data collecting and preprocessing part. In the feature selection part, we use three methods respectively. They are Bag-of-Words, tf-idf, and tf-idf with our proposed boosting features – users' topic-preference. About the classification algorithm, Naive Bayes model is selected. The following part will describe these methods and models in detail.

A. Naive Bayes

Naive Bayes is a popular algorithm in text category. The motivation of the algorithm can be described as, if each tweet is treated as a document d and d is composed of a bag of words $w_1, w_2, \ldots, w_n$, then the posterior probability that the tweets belongs to category c can be demonstrated as

$$p(c|d) \propto p(c) \prod_{1 \leq k \leq n_d} p(w_k|c)$$

In this expression, $p(c)$ is the prior probability of a tweet occurring in class c, defined as the number of tweets in category c divide the total number of tweets in training set. $p(w_k|c)$ is the conditional probability of words distribution in category c. The tweet is assigned to the best class determined by

$$\arg\max_{c \in C} p(c) \prod_{1 \leq k \leq n_d} p(w_k|c)$$

B. Bag-of-Words

By using Bag-of-Words (BoW) model, a tweet can be treated as a bag containing all the words appearing in the tweet, disregarding grammar or order. For example, both "John likes Mary" and "Mary likes John" can be represented as $\{"John'', "likes'', "Mary''\}$ in BoW model.

This method simply uses all words in a tweet as features to represent each tweet, which will make data set very sparse, and reduce the classification accuracy. Thus, we use this method as our baseline.

C. TF-IDF

Comparing with BoW, tf-idf can reduce feature dimension effectively and distinguish the importance of different words. TF-IDF is short for term frequency-inverse document frequency, which is intended to reflect the importance of a word to a document in a corpus. This scheme gives the word w in the document d the weight as

$$TF\text{-}IDF(w,d) = TermFreq(w,d) \cdot \log(N/DocFreq(w))$$

where $Weight(w,d)$ is the frequency of the word in the document, N is the number of all documents, and $DocFreq(w)$ is the number of documents containing the word w.

In our system, we first remove stop words from tweets. Then, for each category, they are treated as a document, and the importance of each word in tweets belonging to a category can be calculated based on tf-idf. Most frequent words and their tf-idf weights are used to represent each tweet and build data set for classification [5].

D. Boosting Features

Since the limitation of tweet-size – 140 characters, each tweet contains very few features compared with all the word-features, which makes accurate classification hard. To improve the accuracy of a classifier, not only should semantic feature selection methods be used, such as tf-idf, but also features from other perspectives should be considered. In this paper, we add 14 features, which represent users' topic preferences for 14 categories. The motivation behind introducing boosting features is, different users would have different posting preferences according to these 14 topics. It's a very intuitive assumption that a user who likes traveling, more frequently

posts tweets about travel, whereas Drugs&Alcohol. So by adding features about their topic-preferences will improve the accuracy. A user's preference for a topic is estimated by the relative frequency as

$$P(topic \mid userID) = \frac{\#keywords\ of\ this\ topic}{\#keywords\ of\ all\ topics}$$

where keywords of each category are generated from Urban Dictionary [11] which is an Internet dictionary containing lots of slang and shortenings. Firstly, we give several "seed words" to Urban Dictionary, and collect 20 most related words of each seed word on the website. After populating and proper cleaning, keywords of each category are gotten. Total appearance times of keywords in a user's tweet history divided by frequency of keywords on each topic is users' topic-preference. By introducing users' topic-preferences, the result improves 3.4%.

IV. Data Collection and Preprocessing

Before training classifier, the data set used for classification should be prepared first. This process includes data collection, data labeling, tweet normalization.

A. Data Collection

Firstly, we randomly selected a user as the seed user, whose following and follower numbers were less than 500. We think a user with too many followings and followers means the user is extremely active, and their behaviors on the social network are quite different from "normal user". Then seed user's followers' and followings' accounts were checked. If the language of an account was English and met our criteria of the normal user, this account will be treated as a new seed user, and crawling would begin again. We repeated this crawling process twice, using Twitter rest API. More than 29,000 users' accounts were crawled in our experiment, from March 10th to March 31th, 2016. For all the crawled tweets, we deleted tweets containing URL or "RT @", since most of them contained less personal information. After data cleaning, we randomly selected parts of tweets to label.

B. Data Labeling

We assume 14 topics might contain private information. They are *health & medical, work, Drugs & alcohol, obscenity, religion, politics, racism, family & personal information, complains & curses, relationship, sexual orientation, travel, school life, and entertainment*. Examples of each category are shown in Table 1. For each category, the distribution of tweets is different. For example, it's easy to find a tweet about work, while hard to find one about illegal drug use. To make sure our classifier can distinguish different topics correctly, we select around 200 tweets for each category. During labeling, one annotator first labels almost equal number of tweets for each category. Then the second annotator checks whether labelings are correct. Only tweets agreed by both annotators remain in the data set. In case a tweet belongs to more than one categories, the tweet is saved in all relevant documents. For example, tweet like "anonymous yoo baby how's that sexy ass of yours? Just sitting here thinking about it while I'm working." is about both work and obscenity. Finally, there are 2,857 labeled tweets in our data set. Among these labeled tweets 2,709 are distinguished and owned by 1,859 users. We also extract all the tweets of these 1,859 users, to analyze their topic preference, which will be used in feature selection part.

TABLE II
Comparison of Different Model

Models	Accuracy	Precision	Recall	F-Measure
BoW	0.727	0.733	0.727	0.728
tf-idf	0.784	0.823	0.784	0.794
topic	**0.818**	**0.836**	**0.818**	**0.821**

C. Tweet Normalization

Tweets have the traits of shortness, full of slang and shortenings, and widely usage of hashtags, which makes it hard to understand for computers if we don't normalize it. Before doing natural language processing for these tweets, TwitIE is used to normalize them.

TwitIE [12] is software focusing on the normalization of tweets, including component used for recognizing the words in long hashtags and changing normal shortenings to complete words. For example, "#lifeisbeautiful" will become "# life is beautiful" and "lol" will become "laugh out loud". This process is important, since hashtags usually contain very important words for classification.

V. Experiment Results

In our experiments, we used the popular machine learning tool – Weka [13]. Weka supports many machine learning algorithms for data categorization, clustering, and feature selection. In our experiments, we implement Naive Bayes model in Weka to three data sets. The first data set consists of labeled tweets processed by Bag-of-Words model. The second one is data set processed by tf-idf. And only words with tf-idf score more than 2 is selected. The third data set is based on the second one but adds features of users' topic-preferences. After classification, each tweet will be in only one category. We utilize 5-fold cross validation to evaluate the classification accuracy. Table 2 presents the comparison of classifiers' accuracy, precision, recall and F-measure for the different data set. The classifiers' performances in each category are evaluated by F-measure and shown in Table 3. We also draw Table 3 as Figure 1.

VI. Analysis and Discussions

Our experiment results in Table 2 show that, from four aspects – accuracy, precision, recall and F-measure, boosting features work better than BoW and tf-idf, which means our motivation of adding boosting features is correct.

From the results, we can see that even the simple bag of words model produces accuracy higher than 70%, which is relatively high, especially considering that there are 14 categories. This is partly due to the existence of bias in the dataset caused by the labeling process. The first annotator quickly

TABLE III
F-MEASURE SCORE OF EACH CATEGORY

Category	BoW	tf-idf	topic
1. Health & Medical	0.698	0.691	0.748
2. Work	0.629	0.885	0.885
3. Drugs & alcohol	0.632	0.749	0.773
4. Obscenity	0.606	0.751	**0.704**
5. Religion	**0.888**	**0.926**	**0.949**
6. Politics	0.842	**0.582**	0.738
7. Racism	0.806	0.833	0.824
8. Family & Personal Info	0.821	0.899	0.901
9. Complaints & Curses	**0.512**	0.777	0.782
10. Relationship	0.730	0.860	0.884
11. Sexual Orientation	0.770	0.741	0.801
12. Travel	0.848	0.811	0.845
13. School life	0.700	0.825	0.834
14. Entertainment	0.689	0.809	0.833

scans through large number of tweets, and labels tweets into a category when certain keywords are spotted. For example, when the annotator sees terms like "drunk", "intoxicated", the tweet is labeled as *Drugs & alcohol*. If a tweet contains terms that are weakly associated with this category, e.g. "a cup of beer before dinner", the tweet is labeled as "not sensitive", and eliminated from the dataset. As a result, each category only contains tweets with strong indicator words. That is, to some extent, inadvertent word filtering is made during the human cognitive process in data labeling. In our future work, we will include significantly larger amount of data labeled through crowdsourcing platforms.

We can see from Figure 1 that categories 4 and 9 produce very low performance. This is because there are many ambiguous tweets that could be labeled into either category. For example, "I'm not saying shes a slut....but...her vagina should be in the NFL hall of fame for greatest wide reciever...". In many cases like this, it is hard to distinguish the tweets even for human annotators. The reason that topic-preferences get a worse result than tf-idf in topic 4 is that our data set has almost same number of tweets in each category. However, in reality, most users have fewer tweets in category 4. Under this condition, topic-preferences have side effect in accuracy, which we will try to compensate in the future work.

For Category 6, tf-idf produces the worst performance in both precision and recall. This is because, in topic 6, there's a lot of rare words, like the names of people. These words have very high tf-idf scores (due to high inverse document frequency), compared with more general words like "government", "republicans". This results in decreased sensitivity for the classifier for many widely used words in this topic. But after introducing topic-preferences, the accuracy improves a lot. When we check the data set of this part, we find there are three users who contribute more than 3 tweets in this categories, and they have more interest in politic than other topics. Using this pattern, we notice that most users who like posting topics about work and school life are more active, and their topics covering many aspects without an obvious difference. This is why our model has little improvement in topic 2 and 13, compared with tf-idf model.

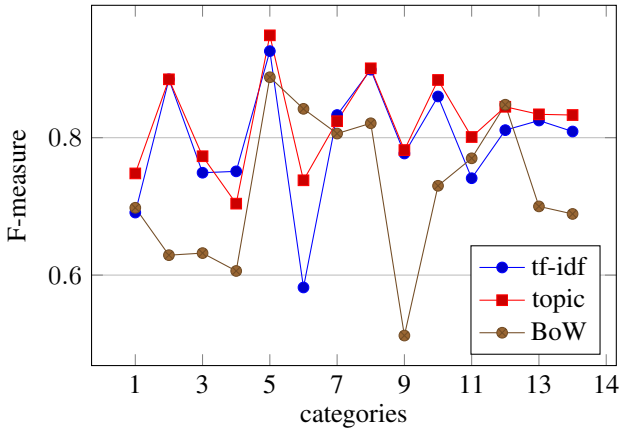

Fig. 1. F-measure Score of Each Category

VII. CONCLUSION

In this paper, we study the problem of classifying private tweets into 14 different potentially sensitive topics based on common tf-idf method and boosting features – users' topic-preferences. The experiment results show that with users' topic-preferences, the accuracy of classification will increase. Users' topic-preferences also effectively boost the classification performance of each category, especially for the ones that Bow and tf-idf are most inaccurate.

REFERENCES

[1] 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 Seventh Symposium on Usable Privacy and Security*. ACM, 2011, p. 10.

[2] J. Vitak, S. Blasiola, S. Patil, and E. Litt, "Balancing audience and privacy tensions on social network sites: Strategies of highly engaged users," *International Journal of Communication*, vol. 9, p. 20, 2015.

[3] B. Luo and D. Lee, "On protecting private information in social networks: A proposal," in *IEEE ICDE Workshop on Modeling, Managing, and Mining of Evolving Social Networks (M3SN)*, 2009.

[4] Y. Yang, J. Lutes, F. Li, B. Luo, and P. Liu, "Stalking online: on user privacy in social networks," in *Proceedings of the second ACM conference on Data and Application Security and Privacy*, 2012.

[5] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," in *IEEE ICDM Workshops (ICDMW)*. IEEE, 2011, pp. 251–258.

[6] H. Takemura and K. Tajima, "Tweet classification based on their lifetime duration," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 2367–2370.

[7] L. Zhou, W. Wang, and K. Chen, "Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones," in *International Conference on World Wide Web*, 2016, pp. 603–612.

[8] H. Mao, X. Shuai, and A. Kapadia, "Loose tweets: an analysis of privacy leaks on twitter," in *ACM WPES*, 2011.

[9] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *ACM SIGIR*. ACM, 2010, pp. 841–842.

[10] H. Liu, B. Luo, and D. Lee, "Location type classification using tweet content," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol. 1. IEEE, 2012, pp. 232–237.

[11] Urban Dictionary, http://www.urbandictionary.com.

[12] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, and N. Aswani, "Twitie: An open-source information extraction pipeline for microblog text." in *RANLP*, 2013, pp. 83–90.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.