International Journal of Semantic Computing
© World Scientific Publishing Company

Content-based Classification of Sensitive Tweets*

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

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Online Social Networks (OSNs), such as Facebook and Twitter, provide open platforms for users to easily share their statuses, opinions, and ideas, ranging from personal experiences/activities to breaking news. With the increasing popularity of online social networks and the explosion of blog and microblog messages, we have observed large amounts of potentially sensitive or private messages being published to OSNs inadvertently or voluntarily. The owners of these messages may become vulnerable to online stalkers or adversaries, especially considering that many online social network platforms (e.g., Twitter) provide open access to the public, including unregistered users and search engine bots. Studies show that users often regret posting sensitive or private messages. However, it is very difficult to completely erase such messages from the Internet, especially when the messages have been indexed by the search engines or forwarded (e.g., re-tweet in Twitter) by other users.

Therefore, it is critical to identify messages that reveal private/sensitive information, and warn users before they post the messages to the public. However, the definition of *sensitive information* is subjective and different from user to user. For example, some users may feel comfortable sharing political opinions, while others do not. 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 messages into topic categories, such as health, politics, family, relationship, religion, etc. In this paper, we make the first attempt to classify sensitive tweets into 13 pre-defined topic categories. In particular, we model the semantic content of tweets with term distribution features as well as users' topic-preferences based on personal tweet history. We also add domain-specific features, i.e., domain knowledge, to improve classification performance. Experiments show that our method can boost classification accuracy compared with the well-known Bag-of-Words and TF-IDF methods. [a]

Keywords: Online Social Networks; Privacy; Classification; Twitter.

1. Introduction

As the most popular open microblog platform, Twitter has 313 million monthly active users [2]. With this new socialization mechanism, users post tweets about every aspect of their daily life, ranging from professional and career development to personal and family

*This work was supported in part by the US National Science Foundation under NSF CNS-1422206 and NSF DGE-1565570.
[a]This paper is significantly extended from a previous conference report [1].

updates, from entertainment to political opinions. Besides the diversity of users' topics, Twitter accounts are public by default, which makes Twitter an ideal source for advertisers and adversaries to collect personal information. At the same time, it makes users' private information exposed unexpectedly.

For most of the users, their tweets are intended for friends that follow them, called "followers" on Twitter. However, Twitter is in fact an open platform – all messages posted on Twitter are accessible to the public, including unregistered users and search engine bots. When users are emotional and/or careless and want to post something, few of them remember this fact. Therefore, users sometimes become very vulnerable – private or sensitive information may be accidentally disclosed, even in tweets about trivial daily activities. Wang *et al.* [3] has shown that regret-tweets are very common. Most of them involve sensitive content and rich sentiments, such as alcohol and illegal drug use, sex, religion and politics, personal and family issues. Although users may imagine the audience before posting tweets, imagining is difficult, and the imagined audience is often inconsistent with the actual audience, as examined by Vitak *et al.* [4]. Especially considering that Twitter does not provide access control over tweets, the potential audience for every tweet is everyone on the Internet. Moreover, Luo *et al.* and Yang *et al.*, [5, 6] use information theory methods to examine users' identifiability, and quantify the amount of information leaked through user attributes from 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 sensitivity 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. Even though some users feel that certain topics are sensitive, such as obscenity content in tweets, they may treat them to different sensitive degrees. That said, 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* (i.e., 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. Moreover, after the classification of different tweets, different strategies for evaluating degrees of sensitiveness can be applied appropriately.

In this paper, we assume that a set of potentially sensitive tweets, with controversial or private content, has been identified already. Our goal is to properly classify these sensitive tweets into 13 pre-defined categories, as shown in Table 1. The reason why we did not use cluster to find these 13 topics is that the topics in tweets are broad and sparse and the occurrence frequency of some extremely sensitive tweets is low. To overcome these deficiencies, we manually list these topics. These 13 topics are selected because some of them are primary sensitive topics, such as drugs, racism and sexual orientation. Some of the topics might cause a controversial discussion or thoughts on Twitter, like religion and political attitudes. Some seemingly harmless topics, such as family information, can also compro-

mise a user's privacy. For example, the tweet about celebrating user's mother's birthday might be associated with the security question of Twitter owner's bank/shopping accounts. Meanwhile, a user's travel plan, from another perspective, is a prediction of the user's absence from home. Sometimes, users also want to have access control functions for their tweets. For example, the tweets about going to a bar is not intended to be shared with their bosses or teachers. In order to implement customized access control functions, accurate topic classification is expected. To the best of our knowledge, this is the first attempt to automatically classify microblog messages into a comprehensive set of potentially sensitive categories. Our baseline approach demonstrates relatively good performance. We further propose two kinds of boosting features: user topic preference and domain knowledge, in our classification model. Finally, the accuracy is improved from 78.8% to 89.2%.

In the rest of the paper, we will introduce the related work from two aspects: content-based tweet classification and Twitter users' privacy protection. Then the preliminary knowledge about the methods and algorithms we used in the paper will be described. The data collection and preprocessing will be followed. In this part, the process of data collection and labeling will be discussed in detail and preprocessing approaches will also be explained. The experiment results will be analyzed at the last part.

2. 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. However, due to the shortness of the tweet – up to 140 characters, the classification of tweets is full of challenges. 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. For example, Lee et al. [7] classify tweets about trending topics into 18 more general trending topics, such as sports, politics, and technology, which can make people understanding these topics easier and improve the performance of information retrieval. The contribution is by introducing network-based information, combining with text-based features, the accuracy of the result is improved. Takemura et al. [8] separate tweets into three kinds: 'expired', 'going to be expired', 'would not be expired', in which 'expired' means the information value has vanished when users read them. To filter out "expired" tweets and save users from outdated ones, some time-related features like the existence of bursty words and time expressions which are extracted from tweets to train the classifier. Sriram et al. [9] introduce eight features which extracted from the content of author's profile and tweet history. Their method outperforms the traditional BoW on classifying generic topics–news, sports, etc. Vosoughi et al. [10] developed a system for identification of rumors about real-world events. They first use syntactic and semantic features of tweets to classify a tweet as an assertion or not. Then, if the tweet is an assertion, the hierarchical agglomerative clustering method is used to identify rumor. Liu et al. [11] propose that user-posted content could implicitly reveal users' location context. And they use time-period and users' check-in history as boosting features to classify "check-in" tweets to different locations. Our work was partly inspired by these classification models, but it is still quite

Table 1. Tweet categories

Category	Example
1. Health & Medical	Seriously starting to regret this surgery...
	I'm waiting for this cough syrup to go down
2. Work	Nothing like being at work at 6 am! #ineedanewjob
	Pretty sure @XXX and I spend more time snapchatting each other at work than actually working.
3. Drugs & alcohol	Nothin' beats whiskey & coke
	I'm thinking he gets pure Columbian cocaine Pseudo Prof. #druglord
4. Obscenity	I hate fucking a skinny bitch!!! #ineedbigass
	This spring break was kind of trash.
5. Religion	Be strong and take heart and wait for The Lord.
	If you put the Lord first, everything else will fall into place.
6. Politics	If Obama wins I'm becoming a communist!
	I wish I was at that debate to ask obama questions. #debate #tearhimapart #romneyryan
7. Racism	I hate black people and gay people as well
	Nowadays these niggas always caught up in they feelings
8. Family & Personal	Grandma and papa flying in tonight!!
	Drinkin beer with future father in law and shondas uncle #buzzed # lovelife
9. Relationship	I have no problem flaunting my relationship.
	On a date with a pretty cute girl. Hope @XXX doesn't mind.
10. Sexual Orientation	Tayler just admitted to me that she is bisexual...
	One day I wanna convert a lesbo
11. Travel	I wish I could just leave and go on a long road trip
	4 more hours until a week of paid vacation
12. School life	3 hour class can suck my balls
	What's worse than immature freshman? Immature seniors.
13. Entertainment	Watching bad girls club while I wait for class #noshame
	Watched the series finale of Ally McBeal on #Netflix and now I'm all in my feels about everything in life ever #alldafeels #ALLOFTHEM

different from the literature, since our work focuses on sensitive tweets in the context of privacy protection.

From the previous introduction of tweet classification, we can find that the seemingly harmless and short tweets contain large information, especially when they are aggregated. And different kinds of informations can be extracted from tweets (e.g., gender [12, 13], location [14, 15, 16, 17], home [18, 19], sociodemographic/socioeconomic status [20, 21, 22], and many more). Thus, attention has been given to the users' privacy which can be leaked from their daily online information sharing. Sleeper *et al.* [23] find people often post regretted messages or reveal too much information online. Other methods have been proposed to examine regrettable social network posts and the possibility of avoiding posting such posts [24, 25, 3, 26, 27]. On the other hand, *Liu et al.* [28] propose a framework to compute the privacy score of a user. They take tweet owners' attitudes and rarity of information into consideration when they calculate the privacy score. Cristofaro *et al.* [29] propose a privacy-enhanced variant of Twitter – Hummingbird, which can provide encryption for pre-defined tweet content. However, few paper deal with how to separate tweets to several sensitive or private topics.

The papers most related to our work are about regret tweets. [25] 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 (i.e., keyword matching), whereas we use more complicate supervised classification to place tweets in a more comprehensive set of potentially sensitive categories. Another paper, [30], classifies whether a tweet about three topics – vacation, drinking, and disease, 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 [25, 30] did not directly classify tweets to different topics, they showed the necessity of classifying topics before identification.

3. Approach

In this paper, we define 13 topics of privacy content, which is shown in Table 1. Our intension is to include topics that some twitter users may not want to share with everyone, i.e., tweet messages that the owners may want to hide from a specific group of followers. While some topics are usually considered as highly sensitive by most of the users, some other topics are only sensitive to a smaller population. For instance, category *Politics* may appear to be far less sensitive than topics such as *Drugs & alcohol*, however, it is not uncommon that some people do not want to share political opinions with their supervisors, colleague, or clients. On the other hand, students may not want to share entertainment-related tweets (e.g., going to a party on a school day) with teachers, thus topic *Entertainment* is also considered sensitive for some people.

Our system consists of the following parts: data collection, labeling, data normalization, feature extraction and classification. The first three steps can be seen as preparation for data set, which will be introduced in the data collecting and preprocessing part. In the feature selection part, besides the content-based features, such as Bag-of-Words and TF-IDF, two kinds of boosting features are explored in our system. They are users' topic preference and domain knowledge. To compare the performance of classifiers trained by different feature sets, Bag-of-Words method is used as the baseline, and another four methods are tested respectively: TF-IDF, TF-IDF with proposed boosting features – users' topic-preference, TF-IDF with proposed boosting features – domain knowledge, TF-IDF with users' topic-preference and domain knowledge which is called "All" for short in the rest of the paper. The whole process of classification is shown in Figure 1. About the classification algorithm, the Naive Bayes model is selected, since it is commonly used in text classification and has good performance. The following part will describe these methods and models in detail.

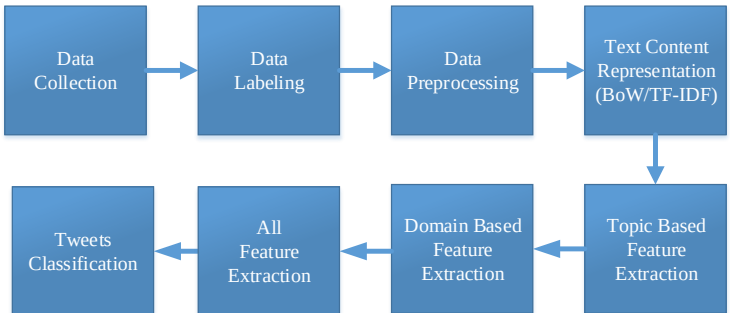

Fig. 1. Diagram for classification process

3.1. *Naive Bayes*

Naive Bayes is a popular algorithm in text categorization. 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 belong to the topic c can be demonstrated as

$$\begin{aligned} p(c|d) &= \frac{p(c)p(d|c)}{p(d)} \\ &= \frac{p(c)p(w_1, w_2, ..., w_n|c)}{p(w_1, w_2, ..., w_n)} \\ &\propto p(c)\prod_{i=1}^{n} p(w_i|c) \end{aligned} \tag{1}$$

In this expression, $p(c)$ is the prior probability of a tweet occurring in class c, defined as

$$p(c) = \frac{N_c}{N} \tag{2}$$

N_c is the number of tweets in the topic c, and N is the total number of tweets in the training set. $p(w_i|c)$ is the conditional probability of words distribution in category c, which can be calculated as

$$p(w_i|c) = \frac{N(w_i, c)}{\sum_{w_j \in V} N(w_j, c)} \tag{3}$$

where $N(w_i, c)$ is number of occurrences of word w_i from topic c, if Bag-of-Words method is applied. If TF-IDF is used, $N(w_i, c)$ will be the TF-IDF score in the certain topic which will be described in the following part. The tweet which is assigned to the best class c can be determined by

$$\arg\max_{c \in C} p(c) \prod_{1 \leq k \leq n_d} p(w_k|c) \tag{4}$$

3.2. *Bag-of-Words*

The Bag-of-Words model is a simplified representation used in a majority of information retrieval and natural language processing techniques. Essentially a text is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity. For example, both "John likes Mary" and "Mary likes John" can be represented as $\{"John'', "likes'', "Mary''\}$ in BoW model. The frequency of a term (TF), namely the number of times a term appears in the text is used as a feature for training a classifier. By using Bag-of-Words (BoW) model, a tweet can be treated as a set containing all the words appearing in the tweet.

However term frequencies are not the best representation for the text. Common words like 'a', 'the', 'to' are the terms with highest frequency in the text. Thus having a high raw count does not necessarily mean that the corresponding word is more important. Moreover, this method simply uses all words that have appeared in a topic as the feature set to represent each tweet in this certain topic, which will make data set sparse and large, and reduce the classification accuracy. Thus, we use this method as our baseline.

3.3. *TF-IDF*

To address the problems posed by the Bag-of-Words model, a widely used technique of normalizing the term frequencies (TF) is to weight a term by the inverse of document frequency (DF) or TF-IDF. TF-IDF can reduce feature dimension effectively, distinguish the importance of different words and reflect the importance of a word to a document in a corpus[31]. This scheme gives the word w in the document d the weight as

$$TF\text{-}IDF(w,d) = TermFreq(w,d) \cdot \log(N/DocFreq(w)) \tag{5}$$

where $TermFreq(w,d)$ 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 [7].

3.4. *Boosting with User Topic Preference*

Because of 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 13 features, which represent users' topic preferences for 13 categories. The motivation behind introducing boosting features is that different users would have different posting preferences according to these 13 topics. It is a very intuitive assumption that a user who likes traveling, for example, more frequently posts tweets about Travel instead of Drugs and Alcohol. So by adding features about their topic-preferences will improve the accuracy. A user's preference for a topic is estimated by two steps.

Table 2. Algorithm for user's own topic preference

Algorithm for a User's Own Topic Preference

Input:
tweets: List of n tweets from a Twitter user u
wordList: Related words list of Certain Topic
Output:
ownPreferences: Topic preferences for Twitter user u

```
1: words = preProcess(tweets)
2: for topic from 1 to 13 do
3:    prob[topic] = 0
4:    for word in words do
5:       if word in wordList do
6:          prob[topic] +=
7:    ownPreferences[topic] = prob[topic]/# of words in tweets
```

Firstly, we define topic-related words as the words which scored more than 1.5 after TF-IDF calculation for each topic. The threshold is selected as 1.5 because during our experiments, this score can decrease the number of features dramatically and still have a good classification effectiveness. Then each user's own topic preference can be calculated through counting the occurrence of topic-related words in each topic and comparing with the occurrence of the whole words in the user's tweet history. The algorithm is stated in the following table 2.

However, for some popular topics, such as 'Travel', lots of users have a high own topic preference in their tweet history. If we just utilize users' own topic preference as the topic preference, the always low score topics, like 'Racism' or 'Sexual Orientation' might be influenced. To avoid the influence of this evaluation method, the relative topic preference should be considered. The estimation of relative topic preference is our second step for users' topic preference features' extraction. In this step, all users' own topic preferences for a certain topic are sorted. We treat users whose own topic preferences are among top 50% as having a preference for this certain topic.

Now the classification problem becomes trying to maximize the conditional probability of a tweet belonging to a certain topic given its content and owner's topic preference, which can be defined as

$$\arg\max_{c \in C} p(c|d,t) \tag{6}$$

Consider a user with certain topic preferences. When she wants to post a tweet, it is probably that the content related to her topic preferences. Thus, the conditional independence is presumed in this situation. And it can be formally defined as follows in Naive Bayes.

$$p(c|d,t) = p(t)p(c|t)p(d|c) \tag{7}$$

For a tweet, its owner's topic preference can always be estimated. Therefore, the con-

ditional probability can be found by calculating

$$p(c|d,t) \propto p(c|t)p(d|c) \propto p(c|t)\prod_{i=1}^{n} p(w_i|c) \tag{8}$$

where $p(c|t)$ can be calculated as

$$p(c|t) = \frac{N_{ct}}{N_t} \tag{9}$$

3.5. *Boosting with Domain Specific Features*

Besides the content of tweets which are the important features for tweet classification, the background knowledge of the information can also play a role for accurate prediction. Therefore, we leverage the background knowledge of four specific topics – entertainment, work, religion and drugs – as four extra features for each tweet. The judgement of these four features are described in detail below.

3.5.1. *Entertainment*

The Entertainment topic can contain varied information. The tweets might have the information about users' preference for a particular kind of movie, music or artist. If a preliminary knowledge about entertainment has been obtained and implemented in the judgement of tweets, it will be helpful for the classification. Therefore, a script has been wirtten to look at individual tokens in a tweet and check whether any of them reference to entertainment content. IMDB's database, IMDB's API and the Google's API are used to identify whether tweets have any reference to entertainment features. For example, "Rachel is making me celebrate World Oceans Day with her by watching Finding Nemo. No complaints. #wildlifeconservation". In this tweet, the user mentions the movie "Finding Nemo", which can be found in IMDB's database.

3.5.2. *Work*

A common observation for the tweets in the work category includes their work load, salaries and professions. Therefore, a list of professions is created, which covers close to 1,100 job titles. Moreover, regular expressions are written to identify tweets where a monetary income is being discussed e.g. $10,000. And for work load regular expressions are written to identify description of a time like 5am, 6pm, 3 days etc. Take the following tweet as an example, "Act for 4 hours, then work for 7.. Tomorrows gonna kill me" contains a commonly used sentence "work for + number (hours)", which can be checked by our regular expression.

3.5.3. *Religion*

A lot of tweets that were categorized as religion had the name of a chapter in a religious text and the verse number. For example, "I praise you, for I am fearfully and wonderfully

made. Wonderful are your works; my soul knows it very well. Psalm 139:14". If a chapter name and a verse number are detected, it's definitely a Religion topic tweet.

3.5.4. *Drugs and Alcohol*

Tweets categorized as drugs had a mention of a drug or an alcoholic drink. So we created two lists. One list for all possible drugs that were illegal or used for recreational purposes and the other list for all kinds of alcoholic drinks. These information is sourced from Wikipedia. For example, in tweet "This song makes me want to do copious amounts of MDMA and cocaine", MDMA (methylenedioxy-n-methylamphetamine) and cocaine are on our list.

4. Data Collection, Labeling and Preprocessing

Before training classifier, the data set used for classification should be collected from Twitter. Once the tweets are crawled, randomly selecting and labeling them to 13 pre-defined topics related to privacy content is called data labeling. After these processes, tweets are still merely strings of text. To make the classifier understand the document, there is a need to represent the document in a more structured manner. Hence, the preprocessing of the data set and the text representation is necessary for the experiment. The above process is shown in Figure 1.

4.1. *Data Collection*

Data Collection is the process of collecting data that is relevant to our project. The data collected will then be preprocessed and used to make predictions and evaluate outcomes; thus, it is one of the most important steps. The better the data for training the classifier is, the better prediction results there will be. Therefore, some constraints are set during collection, which is demonstrated in the following part of this section.

In this part, we randomly selected a user as the valid seed user and the crawler began from a valid "seed user" by using Twitter Rest API, which provides programmatic access to read and write Twitter data [32]. For a valid user, the following constraints are applied:

- less than 500 followers or following count
- user's account language should be English

We think a user with more than 500 followers or followings means the user is extremely active, and their behaviors on the social network are quite different from "normal user". Our research does not target celebrities or public accounts, which are over-active and containing few private information. The seed user's recent 3,000 tweets (according to Twitter's limitation), seed user's followers' and followings' accounts are recorded during crawling. Then we check the seed user's followers and friends to find more potential seed users. If the seed user's friend or follower fulfills the criteria of a seed user mentioned above, then we can start crawling this new seed user's tweets. This method is called snowball crawling. We repeated the snowball crawling method twice using the Twitter Rest API.

More than 29,000 user accounts were crawled in our experiment, from March 10th to March 31th, 2016. From tweets obtained, we deleted tweets containing the term RT@ or URLs for better quality tweets, since most of them contained less personal information.

After obtaining tweets, the next logical step was to find the tweets most relevant to the project. In order to get tweets that were relevant, a rough list of keywords for each category is created. For example, keywords like hospital and surgery would be added to keyword list of 'Health & Medicine', keywords like vacation and road trip would be added to keyword list of 'Travel', so on so forth. To obtain these keywords, a seed word relevant to the category is selected and then fed the Urban Dictionary [33] which is an Internet dictionary containing lots of slang and shortenings. Then the 20 most related words of each seed word on the website are extracted. For each related word we found its related words. This process was repeated twice. After populating and proper cleaning, keywords of each category are obtained. These keywords were then used to fetch relevant tweets from the raw tweet data we collected.

4.2. *Data Labeling*

Once the tweets are filtered by keyword sets, the next step is to manually label the tweets into 13 pre-defined topics which might contain private information. The description and the criterion for judging are listed below. Examples of each category are shown in Table 1.

- Health & Medicine: Tweets that describe users' injury, pain, disease, medicine, surgeries or anything related to hospital visits, etc., are included in this topic category.
- Work: Tweets in this topic category pertain to users' work or employment. Tweets contain users' feeling towards his profession, the work environment, colleagues, work hours or wages, career, job hunt, etc.
- Drugs & Alcohol: Tweets are related to substance abuse or alcoholic consumption.
- Obscene & Abusive: Tweets that contain male and female intimate body parts, sex or porn related text are included in this category. Also, tweets where people use obscene language or complain about something fall in this topic category too.
- Religious: All tweets that indicate religious inclination or contain verses from holy books or talk about faith in God are included in this topic category.
- Politics: Tweets talk about a country's government, policies, elections, etc.
- Discrimination: Tweets that contain text related to discrimination against someone based on cast, color, creed, religion, sexuality, etc.
- Family & Personal Life: Tweets that tell us about the users' personal life. They include birthday tweets, anniversary tweets, marriage or engagement tweets, or pregnancy tweets about the user or his family members.
- Relationship: Tweets related personal relationship.
- Sexual orientation: Tweets that describes a person's sexual orientation.
- Travel: It contains tweets where a user talks about taking a vacation.
- School Life: Tweets containing text related to school like homework, assignments, grades, graduation etc., are included in this category.
- Entertainment: Tweets talk about movies, TV shows, books or music.

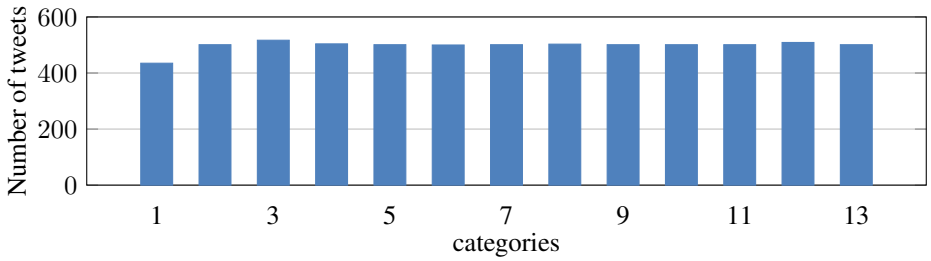

Fig. 2. Number of tweets in each category

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 500 tweets for each category to make the data set balanced. The distribution of tweets for each category is shown in Figure 2.

During labeling, one annotator first labels almost equal number of tweets for each category. Then the second annotator checks whether the labelings are correct. Only tweets agreed upon by both annotators remain in the data set. In case a tweet belongs to more than one category, 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 6,475 labeled tweets in our data set. Among these labeled tweets, 6,345 are distinguished and owned by 3,694 users. We also extract all the tweets of these 3,694 users to analyze their topic preferences, which will be used in the feature selection part of this study.

4.3. *Tweet Preprocessing*

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, a widely used text analyzing tool GATE is used to normalize them. Gate is open source free software for many types of computational task involving human language [34].

In GATE, there is a pipeline specifically developed to handle tweets, called TwitIE [35] which includes the components 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.

After dealing with the hashtags and shortenings by GATE, further processing is also needed. Tweets are tokenized to words through using python library nltk. For each word token, tokens starting with @ are removed and tokens containing non ascii elements like emojis are also deleted.

5. Experiment Results

The experiments are performed using the popular machine learning tool – Weka [36]. Weka supports many machine learning algorithms for data categorization, clustering, and feature selection. In our experiments, we implement the Naive Bayes model in Weka for five different feature-sets extracting from the labeled data set. The first feature-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 a TF-IDF score more than 1.5 are selected. The third one is based on the second one but adding features of users' topic-preferences. The fourth consists of the features in the second one and domain-knowledge features. The last one combines the TF-IDF features and two kinds of boosting features: users' topic preference and domain-knowledge, which is called "All" in our experiment for short. After classification, each tweet in the data set will be in only one category. We utilize 5-fold cross validation to evaluate the classification accuracy, and the final results are averaged over the five folds. The performances of these five different feature-sets are analyzed one by one in the following part. And the experiment results show that compared with the Bag-of-Words, the effectiveness of TF-IDF is obvious and after introducing boosting features, the accuracy of classification improved from 85.4% to 89.2%.

Table 8 presents the comparison of classifiers' accuracy, precision, recall and F-measure of the five different feature-sets. The classifiers' performances in each category are evaluated by F-measure and shown in Table 9. To visualize the Table 9 and observe the result more directly, Figure 5 is drawn.

As the baseline, the Bag-of-Words method achieves the accuracy of 78.8%. Table 3 shows the Confusion matrix of Bag-of-Words approach. From the table we can see, the misprediction happens in every category and distributes evenly compared with the other four methods. This is due to the mechanism of Bag-of-Words, which only counts the frequency of words, regardless of order and without distinguishing the importance between words. For example, some common words like "people" and "girls" will count in every category. Moreover, this method can not decrease the feature-dimension effectively, which leads to the sparsity of the feature-matrix. A more advanced representing method is needed for content-based feature extraction. Therefore, TF-IDF is used in our second experiment.

Table 3. Confusion Matrix with Bag-of-Words

GroundTruth \ Predicted	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1	334	18	5	20	3	0	1	15	9	2	10	10	8
C2	10	401	4	21	3	2	1	7	12	3	16	8	12
C3	12	14	381	22	2	5	6	13	6	7	7	10	16
C4	25	31	10	307	3	0	14	20	36	13	8	20	16
C5	3	6	0	8	463	1	1	1	8	1	2	1	6
C6	6	18	3	13	5	412	12	6	7	5	2	5	7
C7	7	5	3	17	3	9	390	7	13	36	3	5	3
C8	11	14	7	8	2	2	1	390	28	2	16	9	19
C9	5	15	3	23	5	0	5	16	394	4	3	7	21
C10	9	8	7	31	6	2	26	12	18	365	2	8	7
C11	4	6	1	9	3	1	2	9	9	3	455	9	6
C12	10	6	3	13	7	1	6	3	7	5	12	409	12
C13	8	15	2	14	2	1	3	20	17	3	11	5	400

In the second experiment, each category is treated as a document, and each tweet is a sentence in this document. Then, for each category, TF-IDF is utilized to dress the important words in each topic. Finally, we select 1.5 as the threshold of TF-IDF score, which decreases the number of features from a huge one – 10,107 – to an acceptable one – 2,369 – and increases the accuracy of classification from 78.8% to 85.4% effectively. Table 4 shows the Confusion Matrix of the classification results using the feature-set extracted by TF-IDF. We can see the results are quite different from that of the Bag-of-Words method. Most of the wrong predicted tweets are categorized as topic 13. This is because entertainment contains varied content. Except for the specific words like 'Netflix', 'cinema' and 'movie' with extremely high scores, words like 'club' and 'doctor' which might appear in other categories also have a score more than 1.5. For example, tweet labeled as work –"When a parent tells me I was a huge impact on their daughter's life when I worked at the boys and girls club" is hard to judge for the classifier and wrongly categorized as topic 13. Although the overall accuracy of TF-IDF improved a lot compared with Bag-of-Words, in some specific topics, such as 'Entertainment' and 'Obscenity' which share parts of important words with other categories, this method cannot work very well. For example, on the topics of 'Sexual Orientation' and 'Drug and Alcohol', dirty words are common. One of the mis-prediction shows this problem, "Anonymous is a lesbian but she had sex with hale so the makes her bisexual. This notch got it down". This tweet is labeled as topic 9 – sexual orientation, but with little dirty meaning.

Therefore, only using content-based features are not good enough for accurate classification. In the following part, two kinds of boosting features are added to the TF-IDF feature-set respectively. And finally, except for the Bag-of-Words, all methods are combined together, which improves accuracy more than 10% compared with the baseline.

The very beginning intuition behind adding users' topic preferences is that users are more likely to post the tweets about their interested topics. However, for some extremely sensitive topics such as racism or sexual orientation, though a few users might have a

Table 4. Confusion Matrix with TF-IDF

GroundTruth \ Predicted	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1	372	0	1	0	0	0	0	2	0	0	0	1	59
C2	0	427	3	2	0	0	0	2	1	0	17	2	46
C3	2	0	455	6	0	0	1	0	0	0	0	1	36
C4	1	0	9	312	0	0	0	3	4	0	4	0	170
C5	0	0	0	0	487	0	0	0	3	0	0	0	11
C6	0	0	2	0	6	450	1	0	0	1	5	5	31
C7	0	0	0	1	1	4	420	0	2	16	0	3	72
C8	1	0	0	0	3	1	0	463	0	0	16	5	20
C9	0	0	0	7	3	0	0	2	436	0	0	0	53
C10	0	0	0	11	0	0	7	4	1	374	0	0	104
C11	0	3	0	1	0	0	0	1	0	0	412	3	97
C12	1	0	0	0	0	0	1	0	0	0	20	458	24
C13	2	0	1	0	1	0	0	3	0	0	5	6	483

higher preference than that of the other users, these topics are still seldom mentioned, compared with some other 'popular' topics like 'work' or 'travel'. Thus, the user's own topic preference is not accurate enough for classification. The relative topic preferences are needed, which has been described clearly in section 3.4.

The Figure 3 shows the maximum of a user's own topic preference and the medium of a user's own topic preference respectively. Combined with Table 9, we find, this method has an obvious effect on category 4, 7, 10, 11 and 13. This is partial because, topic 4, 7 and 10 are seldom talked about on Twitter. But for the topic active users, they still have more tweets on these topics compared with the most users. For category 11 – travel, it is almost the most favorite topic for every user in our sample set. Then the relative preference makes more sense for the extreme popular topic. The improvement of topics 13 – 'Entertainment' is due to the decrease of mis-prediction. The Confusion Matrix in Table 5 can confirm the demonstration of improvement for various categories.

As mentioned before, through using domain knowledge, more accurate topic-related words will be grasped. By observing the Confusion Matrix in Table 6, we can see improvement in category 'Work' is obvious, due to the application of domain knowledge. So does the category 'Entertainment'. Compared with TF-IDF, the accuracy of domain knowledge method is not improved in category 'Religion' and 'Drug and Alcohol'. This phenomenon verifies that the topic related words in these two categories are very distinct among topics and ubiquitous in the specific topic, which makes TF-IDF a more powerful approach.

After combining TF-IDF, users' topic preferences and domain-knowledge together, we get the classification result as shown in Table 7. Compared with the previous results, fewer tweets are mis-predicted as category 13, due to the usage of domain knowledge. However, at the same time, the mis-prediction on some topics increased. This is because our domain-knowledge features just stress the characters on the four specific categories. For some categories with unobvious content-based traits, such as 'School life', the accuracy will be compromised.

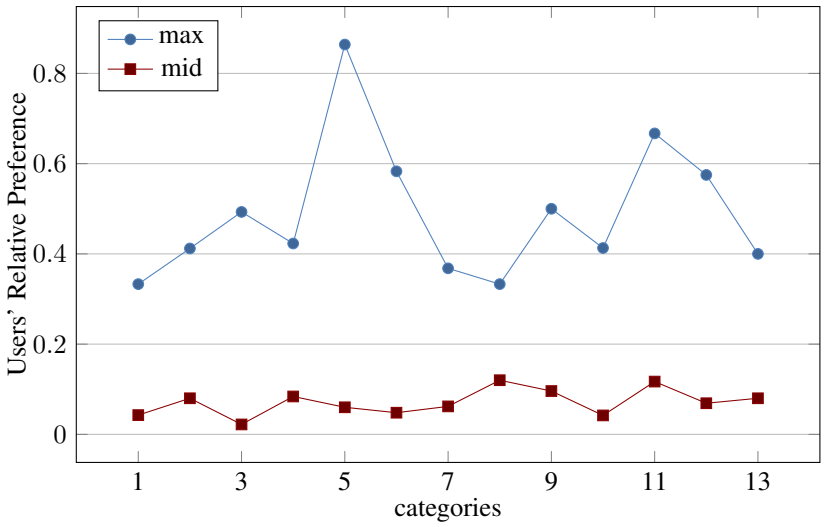

Fig. 3. Users' Relative Topic Preferences

Table 5. Confusion Matrix with TF-IDF + topic

GroundTruth \ Predicted	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1	374	0	15	1	7	0	0	2	0	0	0	8	28
C2	1	424	9	4	1	1	0	2	1	0	20	8	29
C3	2	2	460	7	1	1	1	0	0	0	0	5	22
C4	7	0	67	336	2	0	0	3	4	0	4	24	56
C5	1	0	0	0	486	0	0	0	3	0	0	0	11
C6	0	0	6	1	4	458	1	0	0	0	4	10	17
C7	0	0	18	5	7	5	414	0	1	15	0	7	29
C8	1	0	6	0	5	1	0	464	0	0	16	7	9
C9	2	1	17	10	2	0	1	2	435	0	0	4	27
C10	0	0	36	18	3	0	9	4	1	393	0	6	31
C11	0	3	17	0	3	0	0	1	0	0	430	6	57
C12	2	0	3	0	1	1	1	0	0	0	19	465	12
C13	2	0	5	0	0	0	0	3	0	0	5	9	477

To make the comparison more clearly, Table 8 has shown the accuracy, precision, recall and F-Measure for five different feature-sets respectively. And Table 9 lists the F-measure score for each category under five conditions. From the results, we can see that TF-IDF is an effective content-based feature extraction method, with significant improvement compared with the Bag-of-Words. However, for some categories whose topic-related words are also parts of other categories' topic-related words, this method performs bad, such as topic 'Obscenity' and 'Entertainment'. To make up this disadvantage, boosting features (i.e., users' topic preferences and domain knowledge) are introduced. After combining TF-IDF

Table 6. Confusion Matrix with TF-IDF + domain knowledge

GroundTruth \ Predicted	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1	394	0	5	0	0	0	0	2	0	0	0	33	1
C2	0	454	2	1	0	0	0	3	0	0	8	3	29
C3	2	17	472	5	0	0	1	0	0	0	0	4	0
C4	5	0	38	318	0	0	0	3	5	0	4	130	0
C5	1	0	1	0	486	0	0	1	3	0	0	5	4
C6	0	0	2	0	9	465	1	0	0	1	5	18	0
C7	0	0	1	1	1	4	413	0	2	15	0	64	0
C8	0	0	2	0	3	1	0	463	0	0	16	24	0
C9	0	0	2	8	3	0	1	2	446	0	0	38	1
C10	0	1	4	12	0	0	7	4	1	388	0	82	2
C11	3	2	2	1	0	0	0	1	0	0	451	53	4
C12	1	1	3	0	0	0	1	0	0	0	19	475	4
C13	2	0	1	0	0	0	0	2	0	0	5	20	471

Table 7. Confusion Matrix with All

GroundTruth \ Predicted	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1	389	0	12	5	1	1	0	4	1	0	0	21	1
C2	0	457	4	0	0	0	0	2	0	0	8	2	27
C3	3	16	475	3	0	0	1	0	0	0	0	3	0
C4	16	0	40	373	1	0	0	4	4	0	4	59	2
C5	1	0	1	0	486	1	0	2	3	0	0	1	6
C6	4	0	2	4	4	464	2	0	0	0	4	14	3
C7	5	0	8	7	1	9	424	1	1	15	0	27	3
C8	6	0	0	3	2	2	0	468	0	0	16	12	0
C9	12	0	3	22	1	0	1	2	445	0	0	13	2
C10	11	0	10	33	0	6	9	4	1	402	0	23	2
C11	9	2	7	3	0	1	1	2	0	0	451	36	5
C12	4	1	3	0	0	1	1	0	0	0	19	471	4
C13	4	0	2	1	0	1	0	2	0	0	5	14	472

Table 8. Comparison of Different Model

Models	Accuracy	Precision	Recall	F-Measure
BoW	0.788	0.793	0.788	0.789
TF-IDF	0.854	**0.915**	0.854	0.870
TF-IDF+topic	0.867	0.891	0.867	0.871
TF-IDF+domain	0.879	0.911	0.880	0.886
TF-IDF+topic+domain	**0.892**	0.902	**0.892**	**0.894**

and boosting features together, the classification accuracy has a notable improvement.

To make sure that our approach can be implemented in a real-time system, we also used Weka to evaluate the time for building model on training data and time taken to test each

Table 9. F-measure Score of Each Category

Category	BoW	TF-IDF	topic	domain	All
1. Health & Medical	0.760	0.914	0.904	**0.935**	0.865
2. Work	0.752	0.918	0.912	0.931	**0.936**
3. Drugs & alcohol	0.819	**0.936**	0.793	0.911	0.890
4. Obscenity	0.609	0.740	0.759	0.749	**0.780**
5. Religion	0.919	0.972	0.950	0.969	**0.975**
6. Politics	0.879	0.941	0.946	**0.958**	0.940
7. Racism	0.805	0.881	0.892	0.893	**0.902**
8. Family & Personal Info	0.759	0.936	**0.937**	0.935	0.936
9. Relationship	0.740	0.920	0.920	**0.931**	**0.931**
10. Sexual Orientation	0.768	0.839	0.865	0.857	**0.876**
11. Travel	0.855	0.827	0.847	0.880	**0.881**
12. School life	0.810	**0.927**	0.875	0.654	0.785
13. Entertainment	0.774	0.566	0.730	**0.926**	0.918

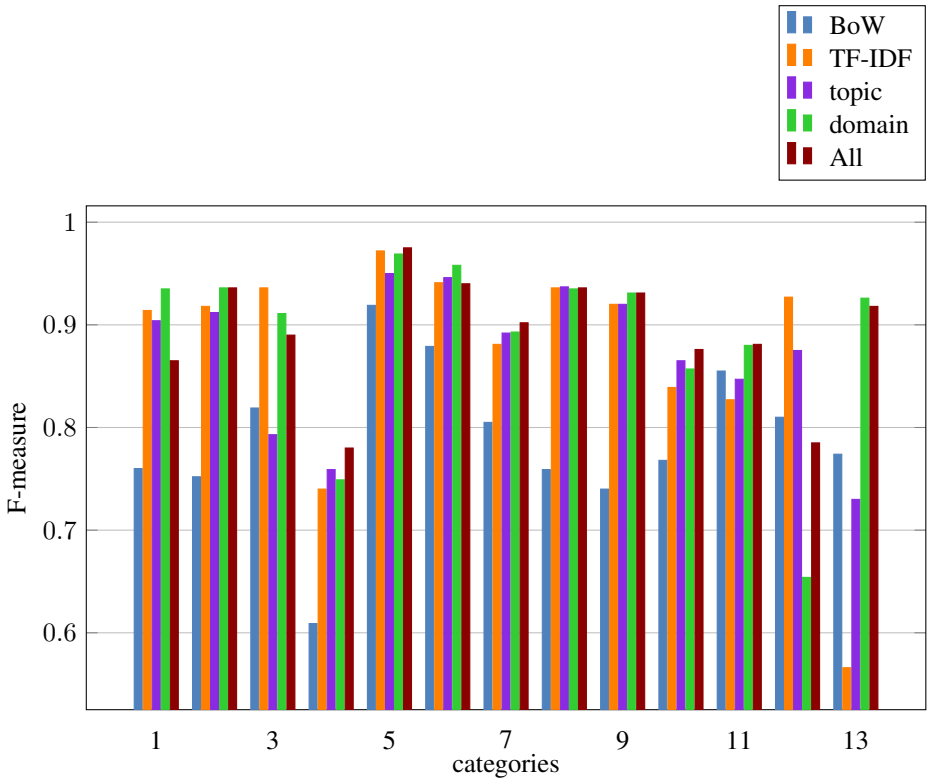

Fig. 4. F-measure Score of Each Category

tweet based on five different classifiers, shown in Table 10. From the results we can see, BoW takes the most time both of training and testing. This is because, the feature set of

Table 10. Training and Testing Time for five Different Models

Models	Module Build Time (second)	Testing Time on each tweet (second)
BoW	10.13	0.0395
TF-IDF	2.7	0.0096
TF-IDF+topic	2.6	0.0091
TF-IDF+domain	2.36	0.0096
TF-IDF+topic+domain	2.07	0.0093

BoW contains more than 10,000 words, which is five times that of the other four methods. The module building time for the other four methods are almost the same. The testing time on each tweet decides whether our proposed method is a real-time solution since the final goal of the project is to check users' posts in the real time and recommend users before they want to post something sensitive. The results in our table shows, except for BoW, the others' time consuming are around 0.01s, which is fast enough for real-time realization.

6. Analysis and Discussion

The previous experiment results show that our motivation of adding boosting features has an impact on the accuracy of classifiers. However, for the results, there are still several phenomena need to be explained in the following part of this section.

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 13 categories. This is partly due to the existence of bias in the dataset caused by the labeling process. The first annotator quickly scans through a 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 a significantly larger amount of data labeled through crowdsourcing platforms.

When compared with the result of TF-IDF, we might notice that the improvement after introducing users' topic preferences seems not ideal enough. This is due to the size of the data set. As we randomly selected the tweets from users' tweet history pool, the chance of getting several tweets from the same user is low. After checking, we find 30% of users have more than 1 tweets in our data set. And the chance that these same user's tweets belong to this user's preferred topics are even lower. With including more tweets in our data set and targeted on the specific users in the future work, we believe the impact of this method will have a significant improvement.

7. Conclusion and Future Work

In this paper, we study the problem of classifying private tweets into 13 different potentially sensitive topics based on the common TF-IDF method and boosting features – users' topic-preferences. The experiment results show that with users' topic-preferences and domain-knowledge, the accuracy of classification will increase notably. Our boosting features effectively boost the classification performance of each category, especially for the ones that BoW and TF-IDF are most inaccurate.

In the next stage of this research, we plan to optimize our experiment from three aspects. Firstly, the data set will be enriched by including more tweets and the possible bias in the data sets will be compensated through labeling tweets on crowd-sourcing platforms like Amazon Mechanical Turk. Moreover, semi-supervised learning will also be considered to automatically enrich our data set. Secondly, we will improve the algorithms to make the ad hoc classifiers more adapted to our sparse data set and at the same time, as well as take efforts to decrease the feature-dimension without the sacrifice of classification accuracy. Thirdly, the characters of each category will be further examined to identify better features and to improve the effectiveness of classifiers. Finally, the accurate classification of sensitive tweets will be employed to enable content-based access control for tweets.

References

[1] Q. Wang, J. Bhandal, S. Huang, and B. Luo, "Classification of private tweets using tweet content," in *IEEE International Conference on Semantic Computing (ICSC)*, 2017, pp. 1065–1074.

[2] "Twitter Usage company facts," https://about.twitter.com/company, accessed: 2016-07-20.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[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] S. Vosoughi and D. Roy, "A human-machine collaborative system for identifying rumors on twitter," in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 47–50.

[11] 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.

[12] W. Liu and D. Ruths, "What's in a name? using first names as features for gender inference in twitter." in *AAAI spring symposium: Analyzing microtext*, vol. 13, 2013, p. 01.

[13] M. Ciot, M. Sonderegger, and D. Ruths, "Gender inference of twitter users in Non-English contexts." in *EMNLP*, 2013, pp. 1136–1145.

[14] R.-H. Li, J. Liu, J. X. Yu, H. Chen, and H. Kitagawa, "Co-occurrence prediction in a large location-based social network," *Frontiers of Computer Science*, vol. 7, no. 2, pp. 185–194, 2013.

[15] Y. Ikawa, M. Enoki, and M. Tatsubori, "Location inference using microblog messages," in *Proceedings of the 21st International Conference on World Wide Web*. ACM, 2012, pp. 687–690.

[16] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 759–768.

[17] H.-W. Chang, D. Lee, M. Eltaher, and J. Lee, "@ Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, 2012, pp. 111–118.

[18] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of twitter users," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 3, p. 47, 2014.

[19] T. Pontes, G. Magno, M. Vasconcelos, A. Gupta, J. Almeida, P. Kumaraguru, and V. Almeida, "Beware of what you share: Inferring home location in social networks," in *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 571–578.

[20] D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras, "Studying user income through language, behaviour and affect in social media," *PloS one*, vol. 10, no. 9, p. e0138717, 2015.

[21] S. Volkova and Y. Bachrach, "On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure," *Cyberpsychology, Behavior, and Social Networking*, vol. 18, no. 12, pp. 726–736, 2015.

[22] V. Lampos, N. Aletras, J. K. Geyti, B. Zou, and I. J. Cox, "Inferring the socioeconomic status of social media users based on behaviour and language," in *European Conference on Information Retrieval*. Springer, 2016, pp. 689–695.

[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 2013 ACM annual conference on Human factors in computing systems*. ACM, 2013, pp. 3277–3286.

[24] J.-M. Xu, B. Burchfiel, X. Zhu, and A. Bellmore, "An examination of regret in bullying tweets." in *HLT-NAACL*, 2013, pp. 697–702.

[25] 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.

[26] K. Moore and J. C. McElroy, "The influence of personality on facebook usage, wall postings, and regret," *Computers in Human Behavior*, vol. 28, no. 1, pp. 267–274, 2012.

[27] S. Patil, G. Norcie, A. Kapadia, and A. J. Lee, "Reasons, rewards, regrets: privacy considerations in location sharing as an interactive practice," in *Proceedings of the Eighth Symposium on Usable Privacy and Security*. ACM, 2012, p. 5.

[28] K. Liu and E. Terzi, "A framework for computing the privacy scores of users in online social networks," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 1, p. 6, 2010.

[29] E. De Cristofaro, C. Soriente, G. Tsudik, and A. Williams, "Hummingbird: Privacy at the time of twitter," in *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, 2012, pp. 285–299.

[30] H. Mao, X. Shuai, and A. Kapadia, "Loose tweets: an analysis of privacy leaks on twitter," in *ACM WPES*, 2011.
[31] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.
[32] Twitter Developer Documentation, https://dev.twitter.com/rest/public.
[33] Urban Dictionary, http://www.urbandictionary.com.
[34] GATE Overview, https://gate.ac.uk/overview.html.
[35] 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.
[36] 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.