A PRIVACY PROTECTION SYSTEM FOR POTENTIALLY SENSITIVE TWEET BY USING CLASSIFICATION

ANIKET MALVADE, SWAPNIL MANDILKAR, KALURAM AURANGE, SHIVANI LATKAR
PDEA's COEM MANJARI (BK), PUNE.
Prof. S. V. Phulari, P.D.E.A's COEM, Manjari(BK), Pune.

Abstract

As we know 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, it is 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. This system develops 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. Hence,this system classifies private tweets into 14 categories, such as alcohol drugs, family information, etc by using naive bayes algorithm. System will model tweet semantic with term distribution features as well as users topic preferences based on personal tweet history.

***Keywords**: Online social networks (OSNs), Twitter,privacy protection mechanism,Tweet Normalization.*

I.Introduction:

With the growth of interactive websites, such as social media, forums, and crowd-sourced product review, it is now possible for people to spread their opinion on different subjects to strangers and many people choose to do so . The creation of such content opens up many possibilities to apply sentiment analysis and natural language processing on the data that is made available. Sentiment analysis can be very effective at identifying different sentiments on web pages , most commonly determining if a text, or part of text expresses a positive or negative sentiment. Typically, sentiment analysis does not consider the context of the sentiment, but some work exists where the causes of the sentiment are determined as well. This paper considers the context of sentiments by determining the topic of texts as well as the sentiment. The topic and sentiment analysis is performed on messages posted to the social media network Twitter. One of the defining characteristics of Twitter is the limitation that all messages must be short, no more than 140 characters long. Any such short message is called a tweet. The number of different topics that are analysed is by necessity limited, which is done by using only tweets from a certain group of users, with the common theme of discussing mainly scientific news and other issues relating to scientific research. The sentiment analysis as well as the determination of the topic of tweets is performed using multinomial naïve Bayes classifiers, since that method others a combination of accuracy and simplicity . The results of the sentiment and topic analysis are used as the basis to interact with human users on Twitter, using natural language processing with an approach similar to the ELIZA program which has been extended to work with parts-of-speech as well as with specific keywords.

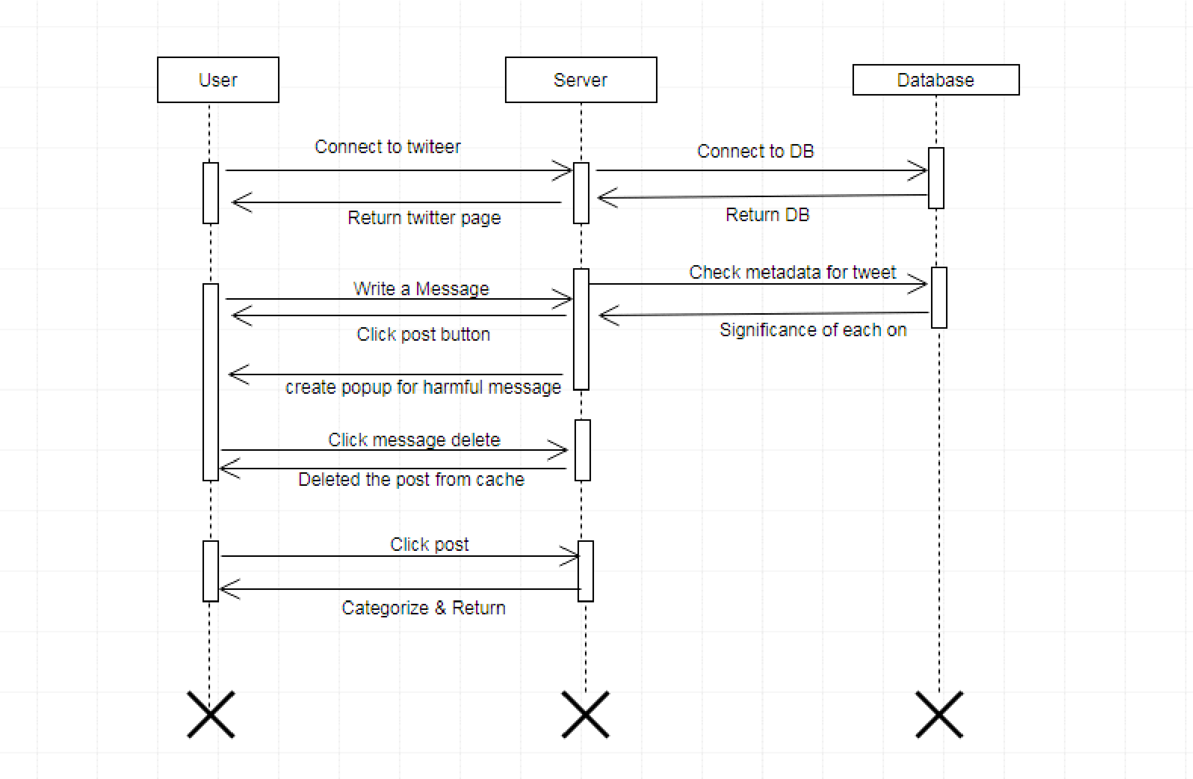

Fig1.Sequence Diagram of BTweet

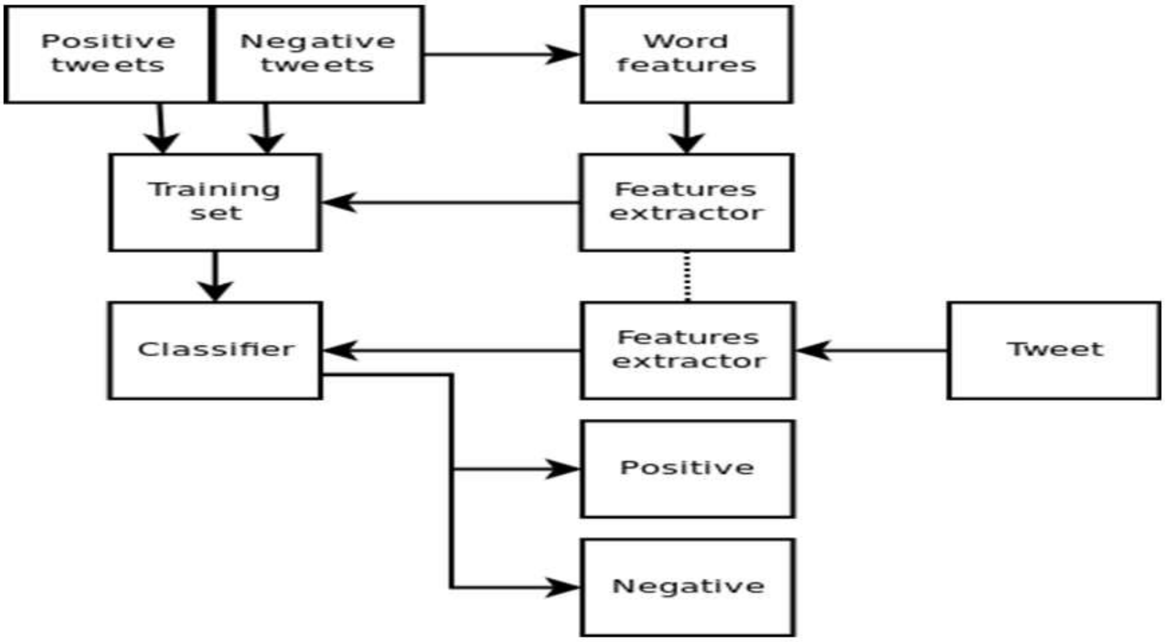

Fig2.Architecture Diagram of BTweet

II.METHODOLOGY

Short Text Classification

Earlier sections dealt with classification of text messages. By “text”, we referred
to documents housing the text. These documents are typically large and are rich with
content. Traditional techniques like Bag-Of-Words work well with such data sets since
the word occurrence is high and though the order is lost, word frequency is enough to
capture the semantics of the document. Alternate approaches like TF-IDF help to counter
some loop holes in the Bag-Of-Words approach by weighing the terms.
With the increase in popularity of online communication like chat messages, rich
information can be mined from concise conversation between groups of people. Some of
the other types of short text messages that are interesting to mine are shown below:

- SMS messages
- Image captions
- Code snippets
- Forum posts
- Product descriptions
- Reviews about various products
- Blog and news feeds (RSS)
- Twitter messages

However, when dealing with shorter text messages, traditional techniques will not
perform as well as they would have performed on larger texts. This matches our intuition
since these techniques rely on word frequency. Since the word occurrence is too small,
they offer no sufficient knowledge about the text itself.

The Natural Language Toolkit

The Natural Language Toolkit (NLTK) is a library for natural language processing written in Python . NLTK contains functions to perform a wide variety of tasks in the filed of natural language processing, we discuss the functions used in this project. One function often used as the first natural language processing stage is tokenization, available in NLTK using the function word_tokenize, which takes as input a string of text and returns a list of strings, where each element of the list is a word or a piece of punctuation. A tokenized text, i.e. a list of strings, can be passed to NLTK for part-of-speech tagging, either using a specialized tagger or the default offering. The default tagger offered by NLTK can be called using pos_tag, which uses the Penn Treebank corpus . The tagger willassign one of 36 part-of-speech tags (or one of 12 punctuation tags) to each element of the input list. The function pos_tag will return a list of tuples where each tuple has two elements, the first being the corresponding string from the input and the second being the tag assigned to the string. The full list of possible tags can be found in Table 2 in . NLTK contains a module, classify, which contains a framework for interacting with a classifiers in NLTK, as well as some built-in classification algorithms. The module also contains an interface that allows for seamless integration of classifiers offered by the Python libraryScikit-learn, alargerselectionthanwhatisavailableinNLTKitself. TheintegrationofScikit-learnclassifiersintoNLTKisdoneusingclassify.scikitlearn.SklearnClassifier, which takes as input a Scikit-learn pipeline. For example, the set up to use a multinomial naïve Bayes classifier from Scikit-learn using the NLTK interface is shown in Algorithm

```
import nltk
 import sklearn
 pipeline = sklearn.pipeline.Pipeline([
('nb', MultinomialNB()),
 ])
 classifier = nltk.classifier.scikitlearn.SklearnClassifier(pipeline)
```

Algorithm 1: How to set up a multinomial naïve Bayes classifier using the Scikit-learn pipeline that can be accessed using the standard NLTK interface for classifiers.
2.3 Datasets When running, the program automatically downloads and processes new tweets from a limited group of Twitter users, a group containing in large part scientists and some sciencerelated news sources.

Gathering this data relies on the Python Twitter Tools to fetch data from Twitter. Some tweets that are not relevant are removed from the dataset, using metadata attributes supplied by Twitter to determine the relevance of each tweet. The removed tweets have not spread much through the network of Twitter users, or are tagged as not being written in English. The spread of each tweet is measured by considering how many times the tweet has been marked as a favorite or has been retweeted by other Twitter users in relation to the age of the tweet and the number of followers that the posting user has. The language tag is supplied by Twitter as a piece of metadata, so the method used to determine the language is not known, may change at any point and should not be relied on to be perfectly accurate. While running, the program makes use of pre-trained naïve Bayes classifiers, which have been trained using a dataset of tweets. The dataset used for training the naïve Bayesian classifiers was collected from the same group of Twitter users, with a science-related focus. The dataset was collected during sessions downloading all new tweets posted by a user in the group while the collection program was running. The collection program was executed several times, taking breaks between executions, from February to April of 2015, resulting in a dataset containing all tweets from some time periods but wholly missing others. The dataset contains 7597 tweets. The dataset has not been made available in full as per the Twitter terms of service agreement, which does not allow distributing a copy of tweets (including for example usernamesand the full text of the tweets). One reason for this is the intention that the user who originally posted the tweet should be able to erase the tweet, which can't be done if it is allowable to make copies of tweets. Another smaller dataset has been collected, using the same users on Twitter as the sources. Each tweet in the smaller dataset has been manually tagged with a tag describing the sentiment of the tweet and a tag describing the content of the tweet. The sentiment of a tweet is described with one of the tags “positive”, “negative”, or “objective”. The content of a tweet is described with one of the tags “scientific news”, “scientific opinion”, “popular science”, “political”, or “other”. The tagged dataset contains 1308 tweets. A dataset of sentiment tagged tweets regarding four different topics was used to compare classification accuracy with the manually tagged dataset collected for this project. The external dataset is available at [13]. This dataset is available as a list of identification numbers of tweets, meaning that the full tweets (including the text of the tweets) must be downloaded from Twitter, which may not work if the tweet has been removed or is no longer publicly available. This limitation meant that only 4617 tweets from the dataset could be downloaded, rather than the 5513 tweets originally available in the dataset. The dataset tags the sentiment of each tweets as one of the following options: “positive”, “negative”, “neutral”, and “irrelevant”. This is similar to how the dataset collected for this project tags the sentiment of tweet, but has the addition of “irrelevant” tags. The dataset also contains tags of the subject discussed in the tweet but these tags do not consider the same idea of topic as considered in this project, as the topic of tweets is only based on direct mention of a Twitter user.

III.Conclusion

The work described in this thesis is a step towards efficient classification of short text messages. Short text messages are harder to classify than larger corpus of text. This is primarily because there are few word occurrences and hence it is difficult to capture the semantics of such messages. Hence, traditional approaches like “Bag-Of-Words” when applied to classify short texts do not perform as well as expected.

Existing works on classification of short text messages integrate messages with meta-information from other information sources such as Wikipedia and WordNet. Automatic text classification and hidden topic extraction approaches perform well when there is meta-information or when the context of the short text is extended with knowledge extracted using large collections. But these approaches require online querying which is very time-consuming and unfit for real time applications. When external features from the world knowledge is used to enhance the feature set, complex algorithms are required to carefully prune overzealous features. These approaches eliminate the problem of data sparseness but create a new problem of the “curse of dimensionality”. Hence efficient ways are required to improve the accuracy of classification by using minimal set of features to represent the short text.We have proposed a framework to classify Twitter messages which serve as an excellent candidate for short text messages because of their 140 character limit. In this framework, we have used a small set of features, namely the 8F feature set to classify incoming tweets into five generic categories – news, opinions, deals, events and private messages.

We have also extended this framework to allow users to define new classes based on their
interest and experiment with new features to improve the performance of our system

References:

[1] A. Java X. Song, T. Finin, and B. Tseng, 2007. Why we twitter: understanding microblogging usage and communities. In Procs WebKDD/SNA-KDD '07 (San Jose, California, August, 2007), 56-65.
[2] N. Cohen. Twitter on the barricades: Six lessons learned. http://www.nytimes.com/2009/06/21/weekinreview/21cohenweb.html, Pub. June 20, 2009.
[3] M. Milian. Twitter sees earth-shaking activity during SoCal quake. http://latimesblogs.latimes.com/technology/2008/07/twitter-earthqu.html, Pub. July 30, 2008.
[4] http://dictionary.reference.com/browse/event
[5] http://www.twitter.com
[6] http://www.time.com/time/magazine/article/0,9171,1044658,00.html
[7] http://en.wikipedia.org/wiki/Micro-blogging
[8] http://www.facebook.com/
[9] http://www.orkut.com
[10] J. Sankaranarayanan, H. Samet, B. E. Teitler, M.D. Lieberman, J. Sperling, TwitterStand: News in Tweets. In Proc. ACM GIS‟09 (Seattle, Washington, Nov. 2009), 42-51.
[11] 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.
[12] 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.
[13] 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.
[14] 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.
[15] 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.