Poster - Using Twitter to Identify Privacy Information

Esha Sharma[1] and Jessica Staddon [2]

***Abstract*— We are building the first comprehensive database of privacy incidents [6]. Such a database helps identify the common elements in privacy incidents that enable technology and policy improvements. Previous work uses news articles to populate the database [15]. We find Twitter to be a good source of information about new privacy incidents. We created data sets of positive examples (privacy-related tweets) and negative examples (non-privacy-related tweets) and used them to train 6 different classifiers. The classifier based on a linear SVM has the best recall on our testing data set (i.e., fraction of privacy-related tweets that are identified): 91.26%.**

I. Introduction and Background

Currently, there is no comprehensive database of privacy incidents available. While there are databases on security incidents that include some privacy incidents (particularly, data breaches), privacy incidents that do not include a security incident are not be represented in these databases. A comprehensive database of privacy incidents is essential for identifying the patterns in incidents. Identification of patterns can lead to changes in privacy policy and technological changes that can improve system privacy.

We are building the first such privacy incidents database [6]. Previous work uses New York Times [16] and Guardian [13] APIs to find privacy incidents [15] for populating the database. We chose Twitter [7] to augment the database both because it is often used for "breaking" events and it is active in technology areas like privacy.

Twitter has around 313 million active users [10] and is accessed by a number of people from different social and interest groups [17]. Because of the free format of tweets a number of people can use Twitter to express their opinions [17]. People use Twitter to tweet or retweet about what they find interesting [14]. People also may tweet about specific things they find interesting than tweeting general comments [20]. A number of entities like Edward Snowden [3] and Wikileaks [11] predominantly use Twitter accounts to communicate about privacy. This makes using Twitter an interesting source for privacy related incidents and information, events, communication. A lot of privacy advocates (e.g., [4], [1], [2]) use Twitter to share information and opinions making it a good source for privacy related discussion, news and articles.

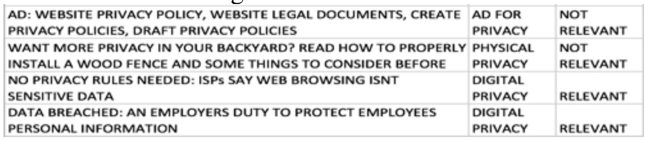
Twitter can be used to find information about incidents in real time [20]. This is helpful since we would like to detect incidents as they occur for our database and find patterns and trends among incidents as they get detected. Updates made by twitter users in real time can be used to keep track of what people are doing at a given time and to gather new information [20]. Twitter makes it possible for users to keep track of large number information updates [20]. It has been used in the past for tracking news and to detect important events [19] , for real time detection of earthquakes [18], to predict changes in the stock market [12]

II. What is a Privacy Incident ?

For our database, we define a "privacy incident" as *an event involving accidental or unauthorized collection, use or exposure of sensitive information about an individual,* ***or*** *an event that creates the perception that unauthorized collection, use or exposure of sensitive information about an individual may happen or is happening,* ***and*** *the event involves data in digital form* [6]. We aim to use twitter to find privacy incidents and any discussion and comments related to these privacy incidents.

III. Identification of Privacy Related Discussion from Twitter

We aim to find incidents or discussion related to privacy involving data in digital form on twitter. Twitter statuses (Tweets) are limited to 140 characters which makes automated semantic analysis difficult. The number of privacy articles found on Twitter is low as compared to other topics like politics. Also, privacy and security are very closely related and many privacy articles are about physical privacy (not online privacy, the focus of our database), further reducing the number of relevant tweets. The table below shows 4 tweets which have keyword "privacy". Of these, the third and the fourth entries fall in the category of the tweets we are looking for.



AD: WEBSITE PRIVACY POLICY, WEBSITE LEGAL DOCUMENTS, CREATE PRIVACY POLICIES, DRAFT PRIVACY POLICIES	AD FOR PRIVACY	NOT RELEVANT
WANT MORE PRIVACY IN YOUR BACKYARD? READ HOW TO PROPERLY INSTALL A WOOD FENCE AND SOME THINGS TO CONSIDER BEFORE	PHYSICAL PRIVACY	NOT RELEVANT
NO PRIVACY RULES NEEDED: ISPs SAY WEB BROWSING ISNT SENSITIVE DATA	DIGITAL PRIVACY	RELEVANT
DATA BREACHED: AN EMPLOYERS DUTY TO PROTECT EMPLOYEES PERSONAL INFORMATION	DIGITAL PRIVACY	RELEVANT

A. Training Data

Since there were no readily available collections of privacy tweets, we created our own data sets using the Twitter streaming API [9]. Data can be pulled using this API in the form of JSON files. From the files pulled we extracted the original tweet, the retweeted status (if available) and the quoted status (if available), the Twitter handle, the expanded URL associated with the original tweet. These data sets were created:

[1] Esha Sharma is with NC State University, 890 Oval Drive, Raleigh, NC 27695-8206, `esharma2@ncsu.edu`

[2] Jessica Staddon is is an Associate Professor of Computer Science and Director of Privacy at NC State University, 890 Oval Drive, Raleigh, NC 27695-8206, `jessica.staddon@ncsu.edu`

Data set 1: This was created using the filter “privacy” keyword with the Twitter streaming API over 4 hour period. This data set was labeled manually for privacy or not privacy. This data set had 285 non privacy tweets and 418 privacy tweets.

Data set 2: This was a randomly generated data set of 43,253 tweets created using Twitter streaming API created over 5 hour running period. All these tweets found using this were labelled “negative”. Since the incidence of privacy related tweets is very low, we assumed that a randomly generated set would consist of non-privacy tweets. To gauge the accuracy of this assumption, we reviewed all tweets in this set containing the keyword “privacy” classified them manually as privacy and non-privacy tweets. We found 4 privacy related tweets in this data set and labelled them as “positive”.

Data set 3: This data set consists of 50,000 tweets which include URLs from our existing privacy incidents database. These tweets were found by searching for the URls in Twitter search and exporting those results using a Twitter search scrapper for this [8]. All these tweets are labelled as “positive”.

Testing data set: This data set was created for testing. It was created using the Twitter streaming API. 100 of these were pulled by not using any filters in the Twitter streaming API and 239 of them were pulled using the “privacy” filter in the API. All these tweets were the manually classified. Of the random sample 99 were not privacy related, 1 was privacy related. Of the privacy sample 102 were privacy related and 137 were not.

B. *Feature Engineering*

The tweets in the data sets created were cleaned and the relevant words were filtered and transformed and the filtered text was represented as feature vectors.

1) Data Preprocessing: The data sets were cleaned before using them for feature extraction. Punctuation and stopwords [5] were removed. URLs starting with “https://twitter.com/”, words of length less than 2, @ references and these words “and” , “http”, “com”, “https”, “www”, “Twitter”, “status”, “utm_medium”, “utm_source”, “sharefromsite”, “utm_campaign”, “utm_content” were removed since during initial training we found that these did not provide any useful information and added noise.

2) TF and TF-IDF Scores: The TF-IDF scores of the words in the cleaned data sets were used to represent the tweets as feature vectors.

C. *Classification*

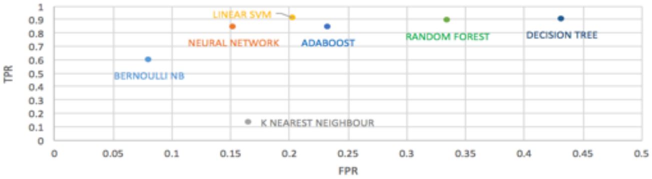
Decision Trees, Random Forest, Linear SVM, Adaboost, Neural Network, Bernoulli Naive Bayes, and k-nearest neighbor classifiers were used to identify privacy related tweets. In the initial experiments, some parameters of the classifiers were tuned to improve accuracy. In the next stage the classifiers were used with the parameters which gave the best accuracy to classify the tweets. The recall of the classifiers was compared. In one set of experiments we tweaked the data sets to remove the Twitter handles and quoted statuses to see how it made a difference to the precision and recall. We also created a data set in which the handles were included and quoted statuses were not and a data set in which just the quoted statuses were included and the handles were not included.

D. *Baselines*

The Twitter privacy filter was used to establish the baseline. We labelled Data set 1 manually to identify tweets as privacy related or not privacy related. The labelling was done by two coders. This data set had 285 non privacy tweets and 418 privacy tweets which is a precision of 59%. We could not establish a baseline for recall since given the large volume of the tweets generated it was difficult to manually classify how many of the tweets not pulled by the streaming API filter could be privacy related.

IV. Evaluation and Results

The classifiers are intended to be used to identify privacy-related tweets in ongoing tweet streams (and so, to help find privacy related articles for the database) hence it is more important to reduce the number of false negatives than number of false positives, hence, we used recall to judge the performance of a classifier. The ten fold cross-validation results indicated that the precision was best for Random Forest (99.21%) and Linear SVM (99.14%) without using the quoted status and the Twitter handles and the recall was best for Neural Networks (86.88%) on a data set with handles and quoted statuses both included. The classifiers were then evaluated on the test data set. The results obtained by the classifier were compared against the manual labels on the test data set to calculate the precision and recall. The precision was best for Naive Bayes (76.54%) and recall was best for Linear SVM (91.26%). The performance of the classifiers went down when the handles and quoted statuses were removed. The performance was worse for data sets with just the handles or just the quoted statuses included.



These are the ROC curves for the classifiers representing true positive rate or TPR (Y axis) and false positive rate or FPR (X axis). In this context since we are interested in getting highest recall and getting a high TPR is more important than getting a low FPR, the best performance is exhibited by Linear SVM which has highest TPR and reasonably low FPR.

Acknowledgment

We are grateful for the support of NCSU’s Science of Security Lablet and the Laboratory of Analytical Sciences. We thank Chinmaya Dabral for his help with creating data set 3.

REFERENCES

[1] Arvind Narayanan's Twitter Account. https://twitter.com/random_walker.

[2] Ashkan Solanki's Twitter Account. https://twitter.com/ashk4n.

[3] Edward Snowden's Twitter Account. https://twitter.com/Snowden.

[4] Jonathan Meyer's Twitter Account. https://twitter.com/jonathanmayer.

[5] Ntlk stopwords corpus. nltk.corpus.stopwords.

[6] Privacy incidents database. https://sites.google.com/site/privacyincidents database.

[7] Twitter. https://twitter.com.

[8] Twitter search scrapper. https://github.com/Jefferson-Henrique/GetOldTweets-python.

[9] Twitter Streaming API. https://dev.twitter.com/streaming/overview.

[10] User statistics from Twitter. https://about.twitter.com/company.

[11] Wiki Leaks's Twitter Account. https://twitter.com/wikileaks.

[12] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[13] T. Guardian. Guardian open platform. http://open-platform.theguardian.com/. Accessed: 2016-3-3.

[14] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pages 73–80. ACM, 2010.

[15] P. K. Murukannaiah, C. Dabral, K. Sheshadri, E. Sharma, and J. Staddon. Learning a privacy incidents database. In *Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp*, pages 35–44. ACM, 2017.

[16] NYTimes. The New York Times API. http://developer.nytimes.com/. Accessed: May 2016.

[17] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, 2010.

[18] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[19] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pages 42–51. ACM, 2009.

[20] D. Zhao and M. B. Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proceedings of the ACM 2009 international conference on Supporting group work*, pages 243–252. ACM, 2009.