Identifying nature of tweets regarding Privacy

Ankur Saxena
asaxena3@ncsu.edu

Vidhisha Jaswani
vjaswan@ncsu.edu

I. Abstract

Our proposal is the sentiment analysis of privacy-related tweets in order to prove or disprove the hypothesis that tweets regarding privacy are predominantly positive.

II. Introduction

Social media is a forum for expression of public interest, so it forms a relevant data source to mine public perception. Social media has been proven to reflect public opinion in areas such as stock markets, banking sector. So, perception garnered from mining of data from social media can be reflective of public opinion as well. Privacy aware development benefits from such studies. Our aim therefore is to do text mining of privacy related news on Twitter in order to test whether the sentiments regarding privacy related tweets are predominantly positive or not [1].

III. Methodology Overview

A. Data set creation

We found there there are no publicly available data sets of tweets related to privacy. So for the purpose of this project, we plan to collect tweets using the Twitter SearchAPI, and manually label them using Cohen's Kappa Score as a measure of inter-rater agreement. The Cohen's Kappa, defined below, helps maintain homogeneity of the classification task.

$$\frac{P_o - P_e}{1 - P_e}$$

where P_0 is the observed agreement among the raters and P_e is the hypothetical probability of chance agreement, using the observed data to calculate the probability of each observer randomly seeing each category.
Maintaining a Kappa score of 0.8 or greater ensures a good agreement between the raters.

B. Sentiment Analysis: Naive Bayes Classification

The aim of the project is to perform the sentiment analysis on privacy related tweets. A part of this project will involve testing the performance of a naive classifier on privacy data set. We plan to build a Naive Bayes classifier that uses ngrams as features. The naive Bayes assumes the independence of each word occurring in the tweet and uses the Bayes theorem [2] to calculate the posterior probability as described below:

$$P(c|d) = \frac{P(c)(\sum_{i=1}^{n} P(f|c)^{n_i(d)})}{P(d)}$$

Here, f represents a feature and $n_i(d)$ represents the count of feature f_i found in tweet d. There are a total of n features. Parameters $P(c)$ and $P(f|c)$ are obtained through maximum likelihood estimates.
The results will set a benchmark for the performance of naive Bayes classifier on privacy related data sets.

C. Sentiment Analysis: Vader/ Textblob

We will be using an existing python package called VADER (for Valence Aware Dictionary for sEntiment Reasoning) to perform sentiment analysis on our data set. It uses a combination of qualitative and quantitative methods to produce, and then empirically validate, a gold-standard sentiment lexicon that is especially attuned to microblog-like contexts. Incorporating these heuristics improves the accuracy of the sentiment analysis engine across several domain contexts (social media text, NY Times editorials, movie reviews, and product reviews) [3].
To verify the results of the sentiments from Vader, we will use another python package named Textblob which uses the StanfordNLP library to analyze sentiments. Also, to maintain consistency of results, a small data set will be manually labelled by us to verify the distribution of tweet sentiments of results from these libraries. This will also be used to perform hypothesis test of our original hypothesis that privacy tweets are predominantly negative.

D. Trend Analysis

Once the sentiment analysis is complete on the privacy data set, we will segment our data set into several time slots and perform topic modelling on it and analyze the trend results with Google Trends data. It is interesting to identify whether the trends in privacy tweets are correlated or not with the trends on google.

IV. Data set Creation

The data set creation was performed in two stages as described below.

- **Data Collection:** We needed to collect a data of tweets relevant to privacy. The key step to this was to finding the keywords that could fetch us desired tweets for our data set.
- **Manual Labelling:** Manually labelling the data set to classify tweets as privacy-related or random. To ensure good agreement a Kappa Score of 0.8 or more was targeted.

A. *Keywords selection*

We initially collected around 18000 tweets using the keyword '#privacy'. This data set contained random as well as privacy related tweets. A Count plot of all the hashtags contained in this tweet data set is shown in figure 1. We manually went through the list of hashtags to select the most relevant ones. The list of keywords selected are shown in the table below.

#privacymatters	#databreach	#cybersecurity
#identitytheft	#onlinesecurity	#cyberaware
#makeprivacygreatagain	#freedomofinformation	#surveillance

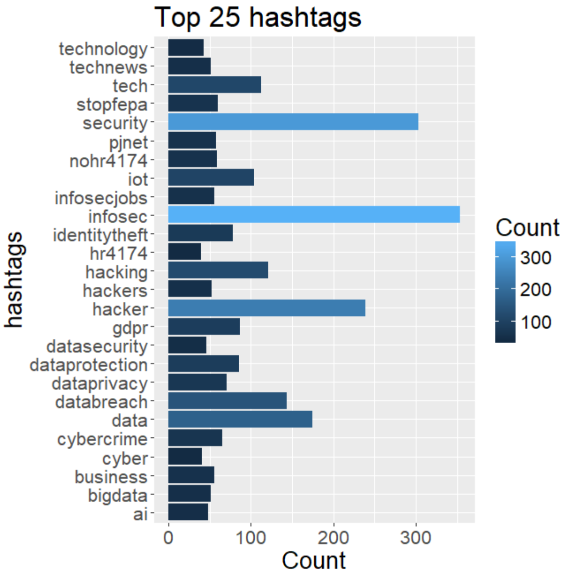

Fig. 1. Most frequent 25 hashtags

With these keywords as our search query and using R's rtweet library, we collected 6480 tweets of English language from Twitters SearchAPI to collect tweets related to Privacy. Due to restriction posed by SearchAPI on the oldest tweet that can be queried, there were limited tweets we could collect. However, we excluded collection any retweets which assures that each tweet in data set was a unique text.

B. *Manual Labelling*

Having gathered a huge data set of privacy related news, our challenge was to classify them as either being actually privacy related or random. Otherwise we could run the risk of including noisy data in our data set. Also, our views of privacy had to agree. To achieve this, we adopted following steps as instructed by Dr. Singh.

1) Select and arbitrary small number of tweets (100, in our case).
2) Manually label the data set.
3) Calculated the kappa score (was approximately 55% the first time).
4) Share our ideas or privacy related news again and attempted to label the tweets, until Kappa of 80% or higher is achieved (90.5%, in our case).
5) Increase the data set size and continue steps 2-4 until tweets exhaust.

We maintained an average kappa score of 82.77% on the entire data sets. The graph for the kappa score throughout the iteration are shown in figure 2. The red line shows kappa score for the iteration while blue line is the cumulative average. Black line is the cut-off line of 80%. After the labelling was completed, we consolidated all the tweets that were labelled as privacy related by both of us into a privacy data set. This data set had 3881 tweets in all.

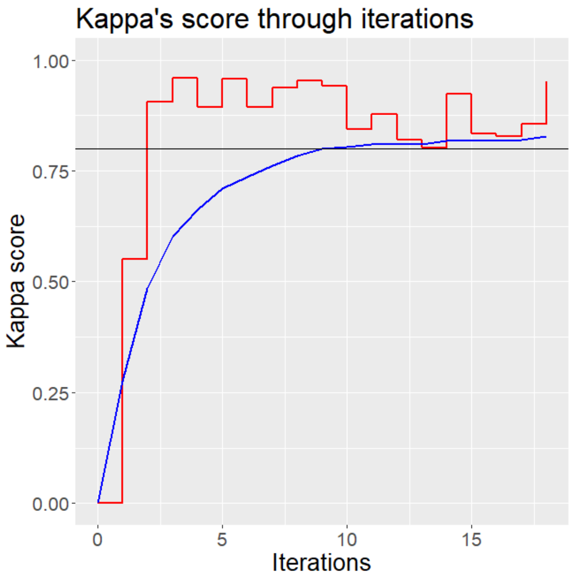

Fig. 2. Kappa score through the iterations

V. Sentiment Analysis: Naive Bayes

We decided to test performance of Naive Bayes classifier on the privacy data set. The steps involved in this task are explained in following subsections.

A. *Cleaning data set*

Before performing any type of modelling on the data set, we cleaning the data set to remove several unwanted patterns and words from each of the tweet. A comprehensive list of which includes:

- removing URLs
- removing non-ascii symbols
- removing mentions beginning with
- converting &amp; to &
- removing punctuation's and white spaces
- lemmatizing tweets to get the root form

B. *Data split and labelling of training data*

We used the VADER sentiment analysis library [3] to label the cleaned data set. The library provides a polarity score in range [-1,1] where -1 being extremely negative and 1 being extremely positive. These scores were then divided into 5 groups based on their scores:

1) *very negative:* scores in range [-1,-0.4)
2) *somewhat negative:* scores in range [-0.4,-0.1)
3) *neutral:* scores in range [-0.1,0.1)
4) *somewhat positive:* scores in range [0.1,0.4)
5) *very positive:* scores in range [0.4,1]

Once the data was divided into these 5 classes, we were ready to perform the classification task. The data set was split into train and test data using $trainingsize = x * totaldatasize$, where x was varied between [0.5,0.8].

C. Naive Bayes classification with unigram feature selection

We performed unigram feature selection on the data converting the data set into a Document-Term Matrix indication presence or absence of each unigram in every tweet. However, we only used the top n most frequent unigrams to build the classifier and perform training. This prevented over-fitting and was computationally inexpensive. The result of prediction on 5 classes present in the data set are shown in the figure 3. The dashed lines are testing accuracy while solid lines are training accuracy. We ran the classifier for n=20 and n=50 which are in blue and red respectively. The results clearly shows that simple Naive Bayes classifier with unigram feature generation, giving an overall accuracy of 32%, is not particularly effective for our privacy data set.

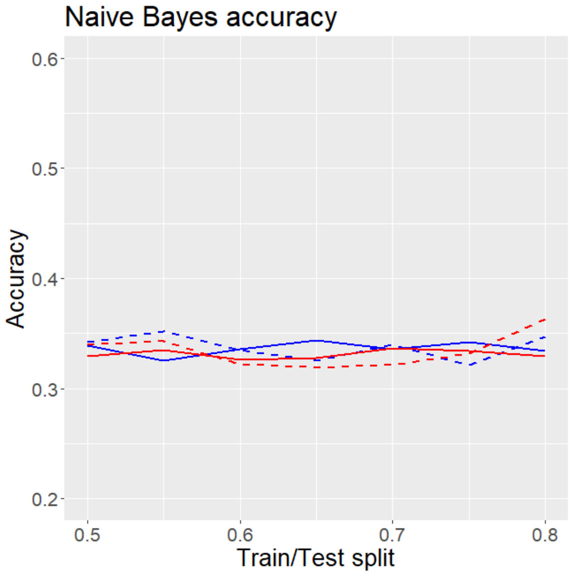

Fig. 3. Accuracy for Naive Bayes classifier

VI. Sentiment Analysis: Vader/ Textblob

We saw that the naive classifiers fail to accurately perform sentiment analysis on privacy data set. Hence, we proceeded further to perform the complete analysis of the data set using existing libraries that used advanced techniques to identify sentiments from tweets.

Vader is a package in python which performs sentiment analysis using a Parsimonious Rule-based Model for social context. The library provides a polarity score in range [-1,1] where -1 being extremely negative and 1 being extremely positive. The results obtained after getting the polarity scores for our data set

The histogram of tweet sentiments obtained from Vader is shown in figure 4. And when they were divided into groups, the count is shown in 5

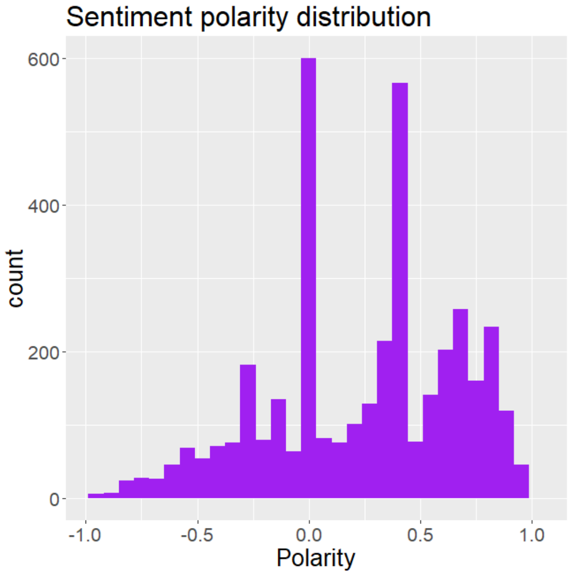

Fig. 4. Histogram of sentiments in Privacy tweets

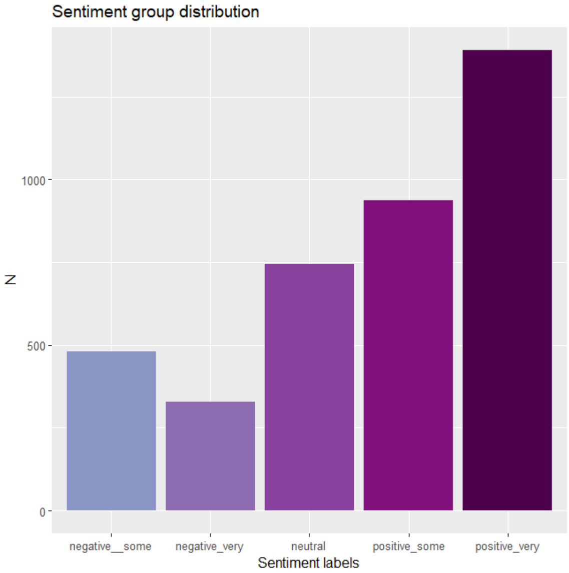

Fig. 5. Count of sentiment groups in Privacy tweets

We generated word clouds of frequent term appearing in each of these groups. Some of them are demonstrated in figure 6, and 7.

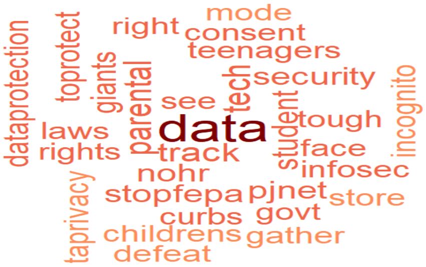

Fig. 6. Word cloud: Somewhat Negative

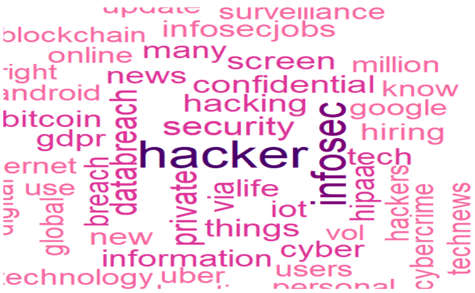

Fig. 7. Word cloud: Neutral

A. *Hypothesis Test*

Similar to the procedure mentioned above apart from privacy related tweets, we have also gathered random tweets using some random keywords such as 'the' and applied sentiment analysis over a data set of 400 tweets in order to check the distribution of sentiments in random tweets as compared to privacy tweets. We obtained a result of 65.25% positive+neutral tweets in the data set. This is corroborated in Figure 8, which shows the frequency distribution of polarity of random tweets and privacy tweets. Blue line is privacy tweets and red line is random tweets. It is clear that privacy tweets are more positive in sentiments than random ones.

VII. Trend Analysis with Google Trends

Having identified the nature of our privacy data set, we now go one step further in describing the nature privacy tweets. We perform topic modelling on our data set and compare the trends with Google Trends

We implemented Topic Modelling using Latent Dirchlet Allocation with the genSim package in python to identify topics occurring our data set. The data set spans over 10 days, so we segmented the data set into intervals of one day and searched for topics in every interval. The topic length was restricted to 3 words each which gave a list of probable topics for each day. Top LDA results included

1) *uber databreach millions*
2) *infosec security cisco*
3) *parent databreach protection*
4) *cybersecurity online data*

For all the topics suggested, we compared the results with google trends within the same time span. Figure 9 shows the trend correlation of the topic *uber hack, equifax, gdpr, and hipaa*. We can be inferred from the graph that there exists positive correlation between the popularity of topics on Google Trends and Tweet data set.

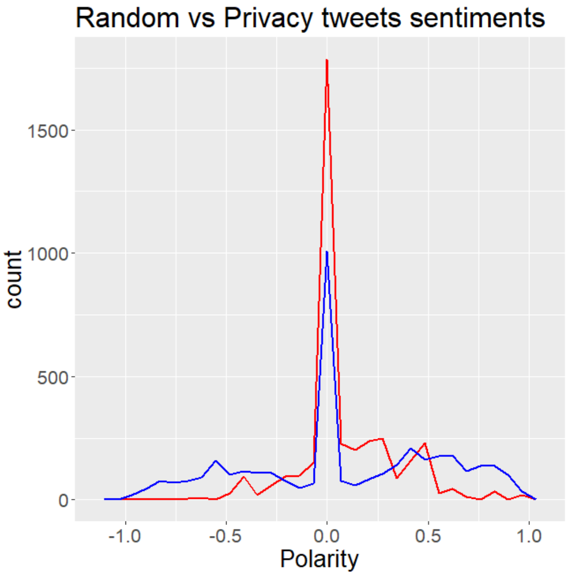

Fig. 8. Sentiment distribution: Random vs Privacy tweets

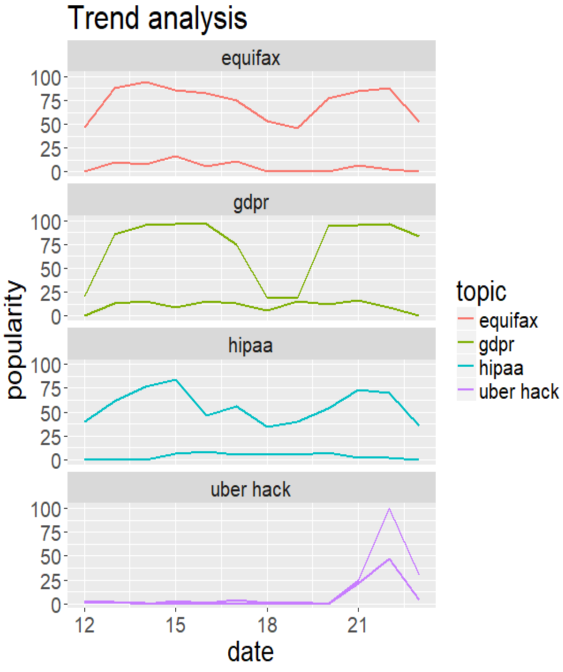

Fig. 9. Trend Analysis

VIII. Conclusion

We can conclude that our initial hypothesis that privacy related tweets are positive can be proven after this sentiment analysis. This seems likely since, people are more aware of their privacy and people tweet about privacy proactively, rather than after being at the receiving end of privacy violation, as was initially hypothesized by us. Positive tweets on privacy can be result of people promoting privacy-aware features, habits, products on social media. Which is generally more active phenomenon than occasional negative tweets which occur only upon violation of privacy of people. Also, comparison with Google Trends show that privacy related topics show positive correlation in popularity over multiple platforms of social media (Google and Twitter in our case).

IX. Identifying Limitations

Our problem statement was to identify the nature of the tweets related to privacy. During formulation, we originally assumed that people speak up about privacy only when theirs is violated. We created a data set from scratch by collecting tweets using privacy related keywords as were seen fit by us. This association, of privacy with individuals voicing their grievance, must have led us to insert bias in our labels while creating the data set. Hence the creation of the data set is prone to plenty of subjection.

Our hypothesis was testing whether privacy tweets are predominantly positive or not. The hypotheses seemed rather concrete at the time of proposal, however, after doing the project, we feel that it was too general and could have been more specific. We could have worked specifically to identify sentiments of just individual tweeters. This specification while problem formulation could have made the labeling spot-on.

We performed Naive Bayes text classification using unigrams to generate feature matrix. This model didn't perform well owing to lack of evenly distributed samples. Presence of large amount of neutral and positive tweets made the classifier biased towards those values. Extensive processing could have improved the performance of classifier. Other than that, we used existing packages like Textblob and Vader which use state of the art techniques for sentiment analysis.

Towards the end, we ventures into trend analysis of topics in our data set and compared the results with Google Trends results. The results were fairly plausible. Any privacy-related news, which impacts the masses showed correlation in their trends. This can't be said for other topics. Unfortunately we had a limited amount of data to work upon, owing to which we couldn't perform a long term analysis of these trends which would have been interesting to learn.

X. Prerequisites for project application

To apply the findings of this project, or to improve the results of the project, it is important to understand that privacy is not only violation of personal space. There can be various aspects to privacy that we learnt through the course of this project. Privacy includes customer support advising deletion of tweets, new technology helping the society to stay alert on internet, people commenting about the privacy related regulations and much more. Once we understand the vast scope of privacy contexts, performing the correct processing on the data set to remove stopwords for privacy can be another challenge. Having done this much, it is easy to verify hypotheses and analyze correlation in trends.

References

[1] Karthik Sheshadri, Nirav Ajmeri, and Jessica Staddon. No (Privacy) News is Good News: An Analysis of Privacy News in the U. S. and U. K. from 2010-2016. 2017.

[2] McCallum, Andrew, and Kamal Nigam. ”A comparison of event models for naive bayes text classification.” AAAI-98 workshop on learning for text categorization. Vol. 752. 1998.

[3] Hutto, C.J. Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.