Privacy Detective: Detecting Private Information and Collective Privacy Behavior in a Large Social Network

Aylin Caliskan-Islam
Drexel University
ac993@drexel.edu

Jonathan Walsh
Drexel University
jdw74@drexel.edu

Rachel Greenstadt
Drexel University
greenie@cs.drexel.edu

ABSTRACT

Detecting the presence and amount of private information being shared in online media is the first step towards analyzing information revealing habits of users in social networks and a useful method for researchers to study aggregate privacy behavior. In this work, we aim to find out if text contains private content by using our novel learning based approach 'privacy detective' that combines topic modeling, named entity recognition, privacy ontology, sentiment analysis, and text normalization to represent privacy features. Privacy detective investigates a broader range of privacy concerns compared to previous approaches that focus on keyword searching or profile related properties.

We collected 500,000 tweets from 100,000 Twitter users along with other information such as tweet linkages and follower relationships. We reach 95.45% accuracy in a two-class task classifying Twitter users who do not reveal much private information and Twitter users who share sensitive information. We score timelines according to three privacy levels after having Amazon Mechanical Turk (AMT) workers annotate collected tweets according to privacy categories. Supervised machine learning classification results on these annotations reach 69.63% accuracy on a three-class task. Inter-annotator agreement on timeline privacy scores between various AMT workers and our classifiers fall under the same positive agreement level. Additionally, we show that a user's privacy level is correlated with her friends' privacy scores and also with the privacy scores of people mentioned in her text but not with the number of her followers. As such, privacy in social networks appear to be socially constructed, which can have great implications for privacy enhancing technologies and educational interventions.

Keywords

privacy; detecting private information; sensitive information; text classification; privacy behavior; social network

1. INTRODUCTION

Numerous organizations, from corporations to governments to criminal gangs, are actively engaged in the collection of personal information released on the Internet. Generally, this pervasive collection is performed without the user's knowledge. Internet users need an increased ability to realize how they are influenced to reveal privacy and the amount of sensitive information they are exposing.

In this work, we will focus on text submitted online, particularly user timelines on Twitter, which expose user information through tweets. A Twitter user might share her text with another party that she trusts but this user may not know how her information will be redistributed on the Internet. The user might also not realize how much private information she is exposing. In such cases, understanding how risky other users are by assigning a privacy score to those users' timelines can help a user decide how much sensitive information she is willing to share with users of certain privacy scores. In order to study and understand privacy behaviors in aggregate, especially as they are embedded in social networks, we propose our method 'privacy detective' to attribute a privacy score to a Twitter timeline using a learning based approach.

Privacy varies from individual to individual and each user may have differing views of privacy. Nonetheless, there is an imperfect and non-negligible societal consensus that certain material is more private than other material in the general societal view. We captured this societal consensus by having AMT workers annotate tweets as private or not according to Table-6 to calculate the privacy scores of Twitter users.

Privacy scores within a user's network could be used to understand how social interactions influence users' privacy behaviors. We need a reliable method for associating users to privacy levels to analyze how privacy behavior is influenced. Do the people a user follows or mentions in tweets influence her sensitive information-sharing behavior? Does the number of followers a user has affect her privacy habits? Our method 'privacy detective' can classify Twitter users' timelines according to the amount of private information being exposed and associate each user with a privacy score.

Outliers in timelines are important since a privacy preserving user can all of a sudden decide to reveal a very rare disease or homeland security information. 'Privacy detective' is not trying to catch such extreme cases and it is not

designed for self censoring. Such outliers do not have an adverse effect on collective privacy behavior analysis, since the focus of the study is on population level effects.

We attempt to confirm a hypothesis that may simply be stated as, those who follow or reply to users who frequently divulge private information are at a higher risk for having their private information exposed. For example, the user may release private information directly, or the release of private information may occur by an encouragement effect in which a user replies to a post from another user revealing private information which they would not have otherwise posted publicly. Intuitively, we believe that certain users will be more likely to reveal private information. We will attempt to discover if users are more likely to reveal private information on their own, or by the influence of their friends, or after prompting from another user.

The benefit for a user having the ability to detect this type of effect is twofold. First, if we are able to provide users with a measure of the full extent of their contacts' release of private information they may take steps to safeguard themselves. Second, if we are able to identify a relationship between users providing private information in replies, users of these types of systems will be more aware of the risks in such situations.

We can learn new things about aggregate privacy behaviors by using 'privacy detective'. The loss of privacy has become prevalent as online social networks expand and privacy behaviors seem to be socially constructed. We perform quantitative analysis of the extent of the user-to-user influence in sensitive information revealing habits as a possible factor contributing to the loss of personal and online privacy. Our goal is to translate this analysis to improve privacy enhancing technologies and educational interventions. For example, a user can apply this on friends' status messages to get a sense of their privacy scores and build friends lists accordingly.

Our analysis has been influenced by the study on the collective dynamics of smoking in a large social network [10] and the spread of obesity in a large social network over 32 years [9]. Christakis and Fowler used network analytic methods and statistical models to derive results from these studies. They examined whether weight gain in one person was associated with weight gain in her friends, siblings, spouse, and neighbors. They concluded that obesity appears to spread through social ties. They also examined the extent of the person-to-person spread of smoking behavior and the extent to which groups of widely connected people quit together. They concluded that network phenomena is relevant to smoking behavior and smoking cessation. These findings had implications for clinical and public health interventions to reduce and prevent smoking and to stop the spread of obesity.

'Privacy detective' detects the presence and amount of private content given text input using topic modeling, a privacy ontology, named entity recognition, and sentiment analysis. Tweets are preprocessed to make better use of natural language processing techniques. This preprocessing is important given our source text, as Twitter has evolved a language which is challenging for natural language processing tasks. For topic modeling we use Latent Dirichlet Allocation method by Blei et al. [5]. The privacy ontology is based on the privacy dictionary contributed by Gill et al. [15]. Named entities consist of names, location, date, time, organization, money, and percentage. Sentiment analysis classifies sentences as either private or not private. Private information can fall under one or more of the following 9 categories: location, medical, drug/alcohol, emotion, personal attacks, stereotyping, family or other associations, personal details, and personally identifiable information. We extract features with the mentioned techniques to train machine learning classifiers on various timelines with varying degrees of privacy in order to come up with a privacy score for a user's timeline of unknown privacy score.

The learning based approach 'privacy detective' is our key contribution for three reasons:

1. Privacy detective detects a broad range of privacy categories. Previous work focuses on certain types of privacy such as location privacy, medical privacy, or writing under the influence.

2. Privacy detective adopts a learning based approach whereas previous methods focus on keyword and regular expression based detection.

3. Privacy is socially influenced and this is demonstrated by the positive correlation between a user's and her friends' privacy scores.

Detecting private information is a hot topic since a lot of personal information is being exposed online. It is difficult to manage private information and friends lists on various social media sites such as Twitter, Facebook, and Google+, which are frequently changing their privacy policies and, at times, sensitive information is being redistributed without the owner's knowledge. 'Privacy detective' can be adapted to assist users in privacy preferences about friend lists, sharing choices, and exposed content. 'Privacy detective' also presents an invaluable research platform for privacy researchers since it makes it possible to study how private information is revealed over time, what affects sensitive information sharing habits, and where people expose personal information.

Text preprocessing, topic modeling, privacy ontology, named entity recognition, and sentiment analysis will be explained in detail in section 6.

2. RELATED WORK

Mao et al. [20] study privacy leaks on Twitter by automatically detecting vacation plans, tweeting under the influence of alcohol, and revealing medical conditions. Their study focuses on analyzing these three specific privacy topics by creating filters to analyze content and automatically categorizing tweets into the three categories. They investigate who divulges information. Their study is followed by a cross cultural study that detects these three types of privacy leaks in the US, UK, and Singapore. They discuss how their classification system can be used as a defensive mechanism to alert users of potential privacy leaks.

Sleeper et al. [26] survey 1,221 Twitter users on AMT and discover that users mostly regret messages that are critical of others, cathartic/expressive, or reveal too much information. They also show that regrets on Twitter reached broader audiences and were repaired more slowly compared to in-person regrets. The privacy categories that we used in our annotations, explained in Table-6 in the appendix, were partly influenced by Sleeper et al.'s Twitter regret categories, which are: blunder, direct attack, group reference, direct criticism, reveal/explain too much, agreement changed, expressive/catharsis, lie, implied criticism, and behavioral edict.

Wang et al. [29] survey 569 American Facebook users to investigate regrets associated with posts on Facebook. They show that regrets on Facebook revolved around topics with strong sentiment, lies, and secrets, which all have subcategories. Privacy categories used in our annotations were also partly influenced by Wang et al.'s regret list. Their survey results revealed several causes of posting regrettable content. They report how regret incidents had serious implications such as job loss or breaking up relationships. They also discuss how regrets can be avoided in online social networks.

Thomas et al. [27] explore multi-party privacy risks in social networks. They specifically analyze Facebook to identify scenarios where conflicting privacy settings between friends reveals information that at least one user intended to remain private. This paper shows how private information can be spread unwillingly when a risky user in the network gets access to other users' personal information. To mitigate this threat, they present a proof of concept application built into Facebook that automatically ensures mutually acceptable privacy restrictions enforced on group content.

Cristofaro et al. [13] present a privacy preserving service for Twitter called 'Hummingbird'. Hummingbird is a variant of Twitter that protects tweet contents, hashtags, and follower interests from the potentially prying eyes of the centralized server. It provides private fine grained authorization of followers and privacy for followers. Hummingbird preserves the central server to guarantee availability but the server learns minimal information about users.

Hart et al. [16] classify enterprise level documents as either sensitive or non-sensitive with automatic text classification algorithms to improve data loss prevention. They introduce a novel training strategy, supplement and adjust, to create an enterprise level classifier. They evaluate their algorithm on confidential documents published on Wikileaks and other archives and get a very low false negative and false discovery rate. A support vector machine with a linear kernel performs the best on their test corpora. Their best feature space across all corpora is unigrams such as single words with binary weights. They eliminate stop words and the number of features is limited to 20,000.

Liu et al. [19] propose a framework for computing privacy scores for users in online social networks based on sensitivity and visibility of private information. The privacy score in this study indicates the user's potential risk caused by her participation in the network.

Chow et al. [8] design a text revision assistant that detects sensitive information in text and gives suggestions to sanitize sentences. Their method involves querying the Internet for detections and recommendations.

There have been numerous studies on topic modeling [18], named entity recognition [25], and sentiment analysis [6] on Twitter as well as normalizing micro-text [30] though not focusing on tweets in particular.

3. PROBLEM STATEMENT AND THREAT MODEL

The main problem we investigate in this work is: *'Does the given text contain any private or sensitive information and if it does, how much of the text reveals private content?'* We want to control the type of information we reveal in our text that is submitted online. We also want to know the private information sharing habits of people in our network in order to make sharing decisions based on their privacy scores. This also helps us understand social influences for revealing private information. Detecting private information is crucial for analyzing textual content and privacy behavior embedded in social networks.

We can assume that, in the worst case, an adversary will have access to all content posted by a user to the social network. Any publicly posted information may be captured by an adversary who is constantly monitoring public portions of the social network. For our study we are analyzing Twitter feeds, which are either entirely public or private, and thus we can focus on users with knowledge that we have captured their full set of activity. For purposes of our study, we assume that adversaries do not have supplemental information to associate with each particular user that is not available through the Twitter system.

User social behavior can impact privacy. An online social network member Alice may be influenced by her friends to release more information than she might otherwise and then some third party observer Bob, who might be an advertiser, a potential employer, or a social enemy, uses this information to harm or embarrass her.

4. DATA COLLECTION

We use randomly selected Twitter users and posts in this study primarily due to the open nature of the posts on that social network. We collect both the relationships between users and their activity on the social network. Furthermore, on Twitter, unlike a social network such as Facebook or LinkedIn, users do not have an array of built in fields or requests for personal data. For example, on Facebook, users are routinely requested to divulge further information to the social network which may include private information such as organizational association, current location, and specific relationship information. Twitter simply requests a username and, optionally, a location. Thus we have the benefit that any private information found within the service is likely to be shared without prompting from the service itself.

The process of data collection emphasizes collection of a continuous stream of a conversation on Twitter. The result of this approach is that tweets of users that are more than

a single degree away from the initial user are collected and considered. In doing so, we consider the complete chain of a conversation, which may have led to the release of private information.

Each tweet is analyzed for metadata within the content of the message. This metadata includes both hashtags and user references. By associating hashtags directly to tweets, we can group tweets that are posted by users who are not connected by a following-type relationship, but may be related in content.

For purposes of experimental data collection, we begin with a seed user. We then select up to 1,000 followers of the seed user and download the tweets for each of these followers. For any tweet which is in reply to another tweet, we also download the originating tweets. We repeat this process until we reach the initial originating tweet. The initial originating tweet is a tweet that has been replied to, but is not a reply to any other tweet. Due to time delays with the Twitter API, this process is time consuming. Thus the automated process developed was essential in data collection.

All tweet data was collected over a period of approximately three weeks in November 2013. Twitter does not present demographic information on its users, thus it is difficult for us to predict age and gender. Although Twitter permits users to enter location information, many users do not, and we did not consider these directly for our study. Since we chose our initial user as a local news sportscaster from Philadelphia, the majority of users live in the Philadelphia area. Up to 200 of the most recent tweets for each user were downloaded. The data collection is designed so that it cannot impact the results because ground truth is provided by AMT annotations to represent a societal consensus which is explained in detail in section 5.

Item	Count
User	95,264
Tweet	426,464
Follower Relationships	4,620
Referenced Users	19,123 (not included in user)
Unique Hashtags	180,186

Table 1: Dataset Information

Data is stored in an SQL database for easier access following collection. A Java API for accessing the data and performing queries was also developed. Table-1 illustrates the total number of entities captured for the dataset. Due to delays caused by the Twitter API, we were unable to collect the complete set of tweets for all followed users in a reasonable amount of time. Thus, one of the intermediate goals is to determine if there is a minimum tweet count which will give a significant chance of evaluating the likelihood of a user releasing or encouraging the release of private information.

5. AMAZON MECHANICAL TURK ANNOTATIONS

The Amazon Mechanical Turk (AMT) is a crowdsourcing Internet marketplace that enables individuals and businesses to use human intelligence for tasks that computers cannot currently accurately perform. The goal of AMT annotations is to obtain ground truth about how much private information Twitter users reveal. Turkers annotate the publicly available Twitter data which is used for calculating the privacy scores of Twitter users. These scores are later used in supervised machine learning to classify timelines based on privacy scores. AMT is used only for annotation purposes on data that's publicly available.

We randomly selected 270 users from the Tweet collection dataset. Then, we randomly selected tweets that total to 500 words from each of these 270 users' timelines. We show that 500 words of random tweets have a good representation of information sharing habits on Twitter and result in reasonable topic ratios in topic modeling.

We asked AMT masters to label whether each tweet is private or not according to Table-6, which is in the appendix.

AMT masters achieve the 'master' distinction by completing work requests with a high degree of accuracy across a variety of AMT requesters. We selected this work to be performed by AMT masters that have demonstrated accuracy in data categorization. Additionally, we placed 10 random quality check tweets, that have been manually labeled in advance, in a user's timeline and used these as an inter-annotator agreement checkpoint. If the worker correctly interpreted the privacy category of 80% of the quality check tweets, we accepted their submission to be used in our experiments. If not, we resubmitted a work request for that timeline.

We tried to categorize tweets that were identified as generically private in order to give guidance to AMT workers. The categories in Table-6 were influenced by related work, primarily the participant reported types of regret in 'Twitter Regrets' [26] and regret categories in 'Regrets on Facebook' [29]. We calculate the privacy score of a user's timeline by calculating the percentage of tweets that fall under one of the 9 privacy categories in Table-6.

- Privacy score-1: If more than 70% of the tweets are not private, the user is assigned a privacy score of 1.
- Privacy score-2: If 30% or more and less than 60% of the tweets are private, the user is assigned a privacy score of 2.
- Privacy score-3: If 60% or more of the tweets are private, the user is assigned a privacy score of 3.

According to this calculation, 185 users had a score of 1, 57 users had a score of 2, and 28 users had a score of 3, as shown in Figure-1.

Having a tool that can detect the sensitivity of a timeline relative to the societal consensus on private information is useful and interesting, especially for population-level effects. The difference between the privacy levels of exposing having the flu and the presence of a rare disease is not weighted in the privacy score calculations. Excluding such exceptions does not have an adverse effect on the analysis since the population-level privacy revealing habits on social network users can be captured without such outliers. This approach

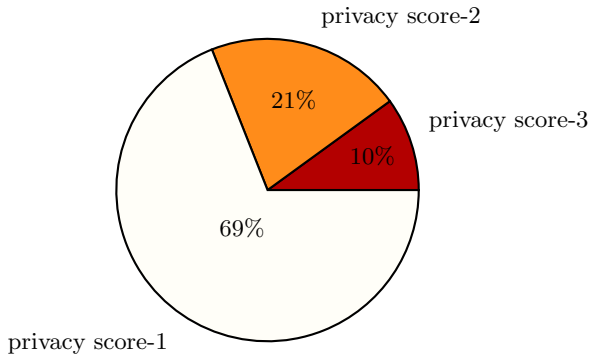

Figure 1: AMT Annotation Results

enables us to focus on aggregate privacy behavior which is a reflection of sensitive information revealing patterns as opposed to discovering important secrets.

A second set of annotations were requested to measure the variance among the first set of annotations, supervised machine learning results, and this second set of annotations. We randomly selected a subset of 100 timelines from the first set of 270 work requests on AMT and had master workers annotate these tweets. We calculated the privacy scores of 100 users the same way we did for the first set of annotations. According to the calculation, 75 users had a score of 1, 15 users had a score of 2, and 10 users had a score of 3. Inter-annotator agreement results are discussed in section 7.

6. APPROACH

We consider a supervised machine learning problem and train classifiers on timelines of users with known privacy scores of 1, 2 and 3 to predict the privacy scores of timelines of interest. We calculated the privacy scores of the users with known privacy scores based on ground truth obtained from AMT annotations. A timeline of a user with unknown privacy score is preprocessed to normalize micro-text and after that, features are extracted to be used in machine learning. Timelines are classified with privacy scores by using AdaBoost [14] with Naive Bayes classifier as a weak learner. Test data is limited to 500 words of randomly selected tweets from each users' timeline for the reasons explained in section 5. The process is shown in Figure-2. The code is available at https://github.com/calaylin/privacy-detective.

Naive Bayes is a popular method to provide baseline text categorization results such as ham or spam classification. Naive Bayes can outperform support vector machines (SVM) with appropriate preprocessing. In our experiments, boosted Naive Bayes significantly outperformed sequential minimal optimization [24], a type of SVM. AdaBoost is a machine learning meta-algorithm that stands for 'Adaptive Boosting'. AdaBoost trains one base Naive Bayes classifier at a time which is tweaked in favor of instances that were misclassified by the previous classifiers, and weights this classifier according to how useful it is in the ensemble of classifiers. As long as the the base learners perform even slightly better than random chance, the boosted ensemble converges to a strong classifier by majority voting.

6.1 Text Preprocessing

In general, informal communication on the Internet does not tend to follow proper English conventions such as proper sentence structure. Furthermore, such communications tend to include significant amounts of abbreviations, slang, and iconography. Since users on Twitter are restricted to 140 characters, there is an increased likelihood that such shorthand will be used. This is especially true when hashtags are considered. Since hashtags are metadata contained within the tweet itself, they are important to consider for both grouping tweets and also for the release of private information.

Tweets contain text that is specific to Twitter and contain micro-text of slang and unstructured sentences. For example, they can include hashtags to tag a certain topic and user handles to refer to another Twitter user. The average number of words per tweet in our sample is 15 and the average number of words per sentence in our sample is 11. These properties of tweets make them challenging for topic modeling, named entity recognition, and many other common natural language processing tasks. In order to create meaningful topic models and detect present entities, we need to clean up tweets and convert the English to a more formal form.

Tweets contain slang words and hashtags that are hard to process as vocabulary words. In order to get rid of these, we replace them with cluster keywords from Twitter word clusters. We use the 1000 hierarchical Twitter word clusters from the Twitter NLP project [22], which were formed by Brown clustering [7] from 56,000,000 English tweets that had over 217,000 words. We manually reviewed the clusters and selected a keyword that describes the words in the cluster. If any of the words in the timeline were present in the clusters, we replaced that word with the cluster keyword.

After converting the words to cluster keywords, we removed non-ASCII characters to reduce non-English language and pictographic characters. User handles (e.g @johnsmith) were replaced with the word *he*, URLs were replaced with the keyword *URL*, and misspellings were corrected based on an English dictionary. These text preprocessing steps are shown in Figure-3.

6.2 Feature Extraction

A list of extracted features which reflect presence of sensitive information are shown in Table-2. The reason behind extracting these particular features and methods used to obtain the feature values are explained one by one in the following sections.

6.2.1 Feature Normalization

All features used in the experiments were calculated either on a normalized scale or normalized during the classification process. The majority of classifiers calculate the distance between two points by using a distance metric. If one feature's values fall under a broad range, then that feature will govern the distance measurements and mislead the classifier [3]. Features are normalized to fit individual samples in the same scale so that they have unit norm and contribute proportionately to classification distance calculations.

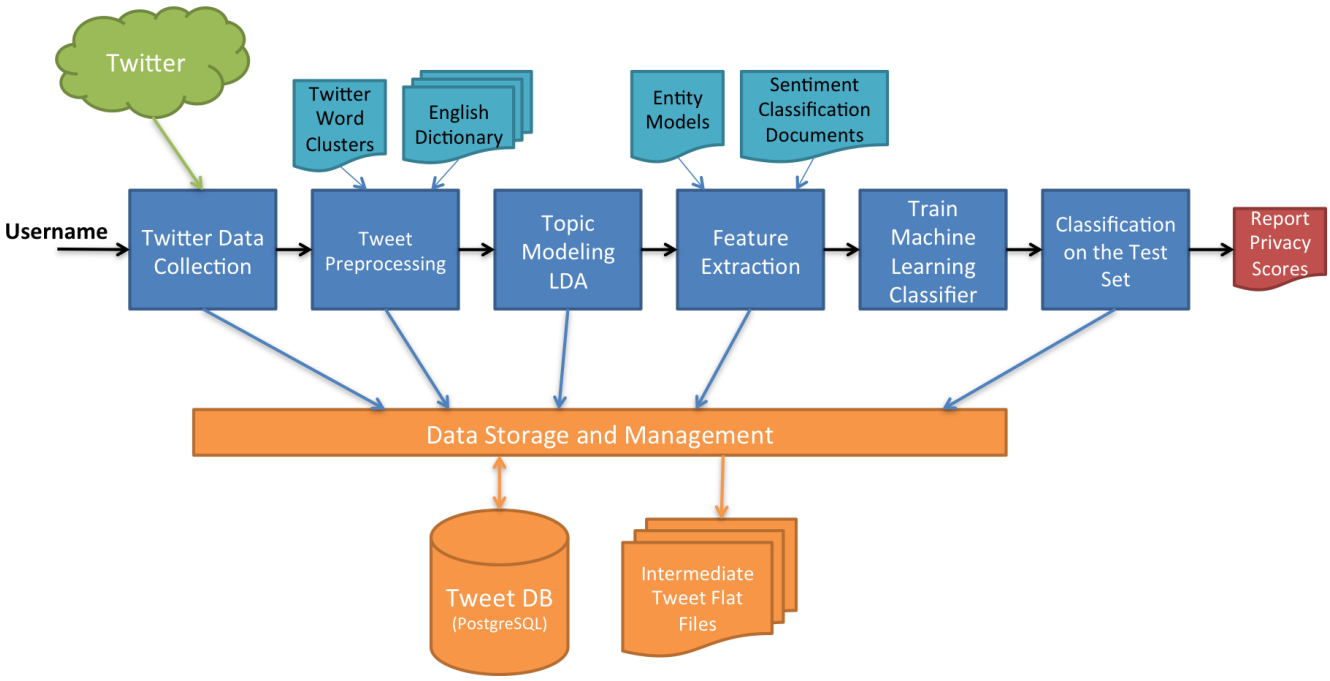

Figure 2: Workflow

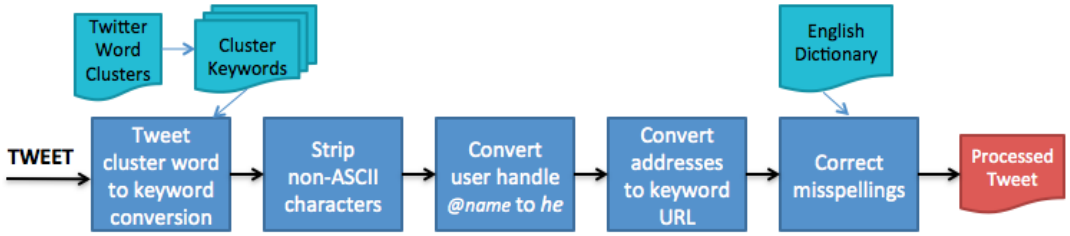

Figure 3: Tweet Preprocessing

Feature	Count
Topic Probabilities	200
Privacy Dictionary Matches	1
Name Entity Count	1
Location Entity Count	1
Date Entity Count	1
Time Entity Count	1
Organization Entity Count	1
Money Entity Count	1
Percentage Entity Count	1
Private Sentiment Count	1
Not-Private Sentiment Count	1
Quote Count	1
URL Count	1
Handle Count	1
Retweet Count	1
Hashtag Count	1

Table 2: Privacy Feature Set

6.2.2 Topic Ratios

Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents according to the discovered themes [4]. Latent Dirichlet Allocation (LDA) [5] is used to discover topics. This model allows you to consider each document in a set of documents as a collection of topics. Topic modeling assumes that when a document is created, the topics that make up that document and their proportions are selected according to the dirichlet distribution. Then, the document is created by repeatedly selecting a topic according to its proportion and a word from the vocabulary for that topic until the document is completed. Although this is somewhat convoluted, if we estimate the posterior probabilities of this process using Gibb's sampling, we can determine the topics discussed in a set of documents and the proportion of those topics present in each document.

We use MALLET [21] to train a topic model on tweets that we collected from 27,293 Twitter users 267,026 tweets through the Twitter API. MALLET topic modeling toolkit contains an efficient and sampling-based implementation of 'Latent Dirichlet Allocation' [5] as well as routines for transforming text documents into numerical representations and

removing stop words.

Some topics of discussion are more likely to reveal private information while other topics remain neutral privacy-wise. Following this intuition, we trained a topic model from the tweet dataset and used this model to infer the topic ratios in given user timelines. Topic modeling and inferencing proved more effective on preprocessed text. We used the inferred topic ratios for each topic as a feature for machine learning.

In order to find the optimum number of topics, we divided the data into two parts: training set (90% of the data) and testing set (10% of the data). We then conducted 20 runs of LDA by changing the number of topics from 20 to 400. On each run, we built an LDA model on the training set and calculated the perplexity (Eq. 1) of the testing set. Perplexity of an LDA model is defined as,

$$Perplexity(D_{Test}) = exp\left(-\frac{\sum_{d=1}^{D} log\ p(w_d|\alpha,\beta)}{\sum_{d=1}^{D} N_d}\right) \quad (1)$$

where, D_{Test} = tweet dataset,
$\sum_{d=1}^{D} N_d$ = total number of tokens in the tweet dataset,
$p(w_d|\alpha,\beta)$ = probability of an entire timeline belonging to a topic.

Lower perplexity scores represent a more robust model. We chose the number of topics as 200 since it produced the most robust model with the lowest perplexity measure.

Table-3 shows 6 topics that fall under private or neutral categories. We extracted top 20 terms from each topic to better assess contents of the topics.

Topic	Top 20 terms
Private: Inappropriate	fuck bad fucking female person i'm people inappropriate shit ass laugh appeal funny man holy fun real hell hate talking
Private: Religious	god love jesus life bless lord give respect man world good heart christ people day job family sex hope peace
Private: Marijuana	marijuana reveals legal medical law philly sam protest call pot country american story smoke white prohibition hunkie smoking horror qld
Public: Sports	sixers game heat tonight win season team people bynum andrew year order nba games flyers play classify ers mention night
Public: News	change africa climate service food news storm year jobs geez location weather job adaptation direction duce shows calls japan tornado
Public: Entertainment	job song music great video love rank listening watching movie channel i'm make favorite show country making talking dance cool

Table 3: Some Private and Public Topics

6.2.3 Privacy Dictionary Matches

We count the number of matches between the 'privacy dictionary' and a user's timeline to be used as a feature in machine learning. Since the timelines are limited to 500 words, this feature is normalized across users' feature vectors.

'Privacy dictionary' [28] is a tool for performing automated content analysis of privacy. The privacy dictionary allows us to automate the content analysis of privacy related text. Using methods from corpus linguistics, Vasalou et al. [28] constructed and validated eight dictionary categories on empirical material from a wide range of privacy-sensitive contexts. They show that these dictionary categories detect privacy language patterns within a given text.

The dictionary is compatible with Linguistic Inquiry and Word Count (LIWC), a text analysis software program developed by Pennebaker et al. [23]. We use the privacy dictionary to calculate details on the usage of categories of words across heterogenous types of text. The eight categories for privacy-sensitive contexts are Law, OpenVisible, OutcomeState, NormsRequisites, Restriction, NegativePrivacy, Intimacy, and PrivateSecret. Each linguistic category contains words and phrases, which can be used to gain an understanding of the types of content contained within the text and in relation to other content.

6.2.4 Named Entity Recognition

The more specific wording a user has, the more entities are found in text. Following this intuition, the higher the specificity is the higher the chances of revealing private information. We use OpenNLP's [1] named entity recognizer to extract the number of name, location, date, time, organization, money, and percentage entities. Again, since the timelines are limited to 500 words, this feature is normalized across users' feature vectors.

6.2.5 Sentiment Analysis

Sentiment analysis is generally used to extract subjective information in text. It can be used to infer whether the source is subjective or objective, or whether the tone is positive, negative, or neutral. We use sentiment analysis to help us differentiate private tweets from neutral or objective tweets. Therefore, the sentiment of interest is the state of revealing private information which can be used as a feature on a tweet by tweet basis.

We train a sentiment classifier on 9 privacy categories: location, medical, drug/alcohol, emotion, personal attacks, stereotyping, family or other associations, personal details, personally identifiable information, and a not private category that contains objective and neutral tweets. These 9 categories are influenced by related work and are explained in more detail in section 5. Each category contains at least 6000 words of training data made up of manually labeled tweets that represent the privacy content. We use Lingpipe's *n-gram* based sentiment classifier [2] to extract the number of tweets in a timeline classified as private or not private. This feature is normalized across users because of the timeline word length limit.

6.2.6 Quote, URL, Handle, Retweet, Hashtag Count

Twitter users tend to place retweets or sentences written by others in quotes. We use the number of quotes and retweets in timelines as a feature that represents not private content. The number of URLs, user handles, and hashtags also have information gain and are included as supplemental features. Since there is a word limit on the timelines being analyzed, these features are considered normalized.

7. RESULTS

The first set of AMT annotations show that 10.37% of Twitter users frequently reveal personal information (privacy score-3), 21.11% reveal some private information (privacy score-2), 68.52% tend not to reveal much private information by tweeting (privacy score-1). Twitter users need to be aware that the number of people revealing private information is a significant portion of all users and make conscious decisions when thinking of posting any text with private content.

We obtain 95.45% accuracy in a two class task (users with scores of 1 and 3), and 69.63% accuracy in a three class task (users with scores of 1, 2, and 3) after performing 10-fold-cross-validation by using AdaBoost with Naive Bayes and standardizing the features on the dataset obtained from AMT annotators. These results show that the extracted features represent privacy from a general standpoint instead of focusing on single privacy categories. This differentiates our work from previous efforts and makes our approach applicable to a broader range of privacy concerns. Using the Brown clusters and converting the text to a format that is more natural language processing friendly was a key element of our success at distinguishing private and non-private tweets. Without these transformations, accuracy drops to 58.93% in a two class task (users with scores of 1 and 3), and 38.10% accuracy in a three class task (users with scores of 1, 2, and 3) after performing 10-fold-cross-validation by using AdaBoost with Naive Bayes, and standardizing the features on the same dataset without preprocessing the text.

7.1 Twitter Database User Scores

We trained a classifier from our dataset that reached 69.63% accuracy in a 3-class supervised experiment. The timelines in this dataset are not present in our Twitter database. We used this classifier to predict the scores of 1,982 Twitter users that had at least 500 words of tweets in their timelines. The Twitter database experiment's results show that 18.62% of Twitter users frequently reveal personal information, 30.52% reveal some private information, 50.86% tend not to reveal much private information by tweeting. We created a privacy map of these 1982 users in Figure-4, where each node represents a user, each edge represents a following relationship, and the node colors represent privacy score where light yellow is a score of 1, orange is a score of 2 and red is a score of 3.

7.2 Correlation between User's Privacy Score and User's Friends' Privacy Score

The privacy scores of users, and the average of privacy scores of people they follow is positively correlated. This means that the higher a user's privacy score, the higher her friends' privacy scores are and vice versa. Spearman's Rho was calculated to measure the direction and strength of relationship between users' and their friends' privacy scores. We used the privacy scores of 45 users, who had at least 30 friends with sufficient amount of tweets, and these friends' privacy scores in Spearman's Rho calculation. The resulting R value is 0.41, and two-tailed P value is 0.005, which shows that there is a statistically significant positive correlation between the two variables.

We chose Spearman's correlation over Pearson's correlation

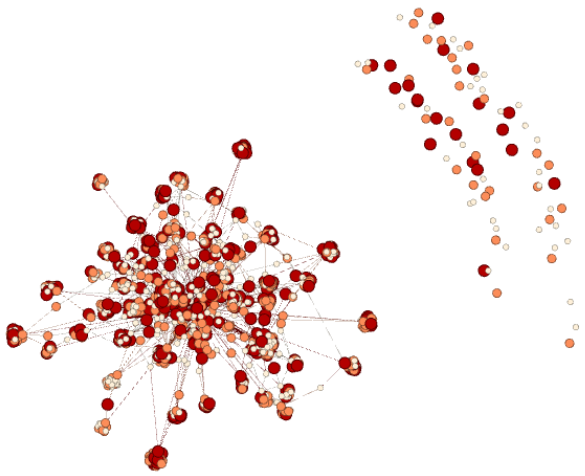

Figure 4: Twitter Privacy Map

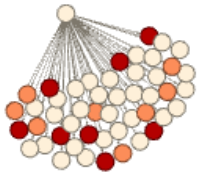

Figure 5: User with privacy score-1

because Spearman's correlation does not make any assumptions about the distribution of the values, and the calculations are based on ranks, not the actual values. Pearson correlation assumes that both of the two variables are sampled from populations that follow a Gaussian distribution. There has been no study showing that Twitter privacy scores follow a Gaussian distribution and our sample size is not large enough to support or neglect such an argument. Three random users with privacy scores 1,2, and 3 and their friends' scores, are illustrated in Figure-5, 6, and 7. The correlation between the user's privacy score and her friends' privacy scores are shown by the main node's color of light yellow, orange or red being more dominant than the dataset's average distribution.

7.3 Correlation between User's Privacy Score and Mentioned Users' Privacy Score

There is a positive correlation between a user's privacy score and the privacy scores of users she mentions in tweets. Spearman's Rho calculation on 45 users that mentioned at least 30 other users with calculated privacy scores returned an R value of 0.37 and a two-tailed P value of 0.01, which shows

Figure 6: User with privacy score-2

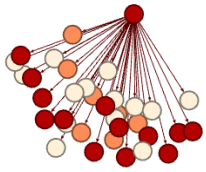

Figure 7: User with privacy score-3

that the positive correlation between two variables is statistically significant. This correlation is weaker than the correlation between a user's privacy score and the privacy scores of her friends. This indicates that users prefer to follow other users that have similar privacy revealing habits and users tend to mention users with similar private information revealing habits. Nevertheless, a user's friends' average privacy score is a stronger indicator of a user's own privacy score than the average privacy score of people a user mentions in tweets.

7.4 Correlation between User's Privacy Score and Number of Followers

Number of followers for each user that had a calculated privacy score was obtained. There was no statistically significant correlation between a user's privacy score and her number of followers. Both Spearman's Rho and Pearson's correlation coefficient were close to 0.

For example, at the time of gathering data from Twitter, *rogerfederer*, who is a professional tennis player ranked world no. 4 had around 1,500,000 followers and a privacy score of 1, whereas *mark_wahlberg* who is an American actor also had around 1,500,000 followers and a privacy score of 3. We can conclude that there is no correlation between how much private information you reveal and how many followers you have.

7.5 Inter-Annotator Agreement

Cohen's Kappa coefficient [11] was calculated to measure the inter-annotator agreement in a 95% confidence interval. Cohen's kappa coefficient is a statistical measure of inter-annotator agreement for categorical items which takes into account the agreement occurring by chance. Cohen's Kappa is a measurement of concordance that can be applied to data that is not normally distributed or binary data such as true/false, but is best suited to an ordinal scale, such as our 3 point privacy score scale. Kappa statistics is generally thought to be a more robust measure than simple percent agreement calculation since it excludes the agreement expected from random chance.

Cohen's Kappa can be calculated in two ways, namely weighted kappa coefficient and unweighted kappa coefficient. Weighted Kappa coefficient [12] is recommended when the score categories are more than two and not binary. Since our predictions and annotations have 3 categories, we used the weighted Kappa, which takes into consideration the distance between the annotated categories.

Landis and Koch [17] characterized Kappa coefficient values less than 0 as indicating no agreement, 0 to 0.20 as slight agreement, 0.21 to 0.40 as fair agreement, 0.41 to 0.60 as moderate agreement, 0.61 to 0.80 as substantial agreement, and 0.81 to 1 as almost perfect agreement.

There is a *fair* agreement between the annotations of the first set and second set of AMT annotators. The agreement between the first set of AMT annotators and our classifier is *fair*. There is also a *fair* agreement between the annotations of the second set of AMT annotators and our supervised machine learning predictions. These three results suggest that the variance of privacy annotations between humans is in the same range as the variance between human annotators and supervised machine learning predictions. Determining if a given tweet is private or not is subjective to an extent for AMT workers even though we provide detailed annotation directions. Seeing that privacy detective's results fall under the same level of subjectivity makes it more reliable in addition to the accuracy obtained from supervised experiments.

8. LIMITATIONS

The ground truth in our training set is provided by AMT workers and not the original writers of the tweets. Even though we tried to provide turkers a detailed explanation of how to annotate tweets and choose privacy categories, the original author of the tweet might have a different intension in writing the tweet. We wanted to obtain a man on the street view of privacy, therefore this limitation did not harm our approach.

The length of timelines and the number of tweets have an effect on how much private or sensitive information is released. A personal profile can be formed by investigating the writings of a person. The more text that is present the more accurate the profile will be. We do not have a clear understanding of the quantified effect of writing length on the amount of personal information leakage. There are numerous components in text that are representative of private information or neutral data. Each component's effect need to be factored out in order to investigate the effect of text length. In order to keep the length factor stable, we limited our study to 500 words of randomly selected tweets from a Twitter user's timeline.

Most tweets in a user's timeline could be benign and a few could be very private. Our sample of 500 words might only capture the neutral tweets from this user. Not including such exceptions in our analysis is not affecting the privacy score calculations adversely. We are interested in users' habits rather than the outliers in their timelines.

9. DISCUSSION

Entity recognition requires proper English sentences to detect sentences and the entities within. Tweets by nature do not resemble proper English sentences and therefore render natural language processing tasks quite challenging. We believe that improving named entity recognition accuracy on tweets will boost our private information detection performance.

Table-4 shows the information gain ranks of features. Not-private sentiment count is the most important feature followed by 13 topics and the rest of the non-topic related features. The information gain ratios which are close to 1%

Feature	Rank
Not-Private Sentiment Count	1
13 Topics	14
Private Sentiment Count	15
122 Topics	137
Privacy Dictionary Matches	138
Percentage Entity Count	139
Organization Entity Count	140
Name Entity Count	141
Time Entity Count	142
Quote Count	143
Retweet Count	144
Handle Count	145
Hashtag Count	146
URL Count	147
Money Entity Count	148
Location Entity Count	149
Date Entity Count	150
65 Other Topics	215

Table 4: Information Gain

for all of the 215 features show that all features contribute proportionately and they are all important.

There are many topics that contribute to correct classification. Creating a topic model with correct number of topics and precise LDA parameters is crucial for accurate analysis. Topic discovery is more effective on a larger dataset, which covers a greater range of topics and words. As we collect more tweets through the Twitter API, we periodically update our topic models to include recent topics. 13 topics that had the highest information gain ranks among 200 topics in our feature set are shown in Table-5.

Topic	Top 20 terms
People	url mammal person girl family bad dogs boy age front man cats location dog hot lucky loves color baby cat
Sports	order refresh year draft round eagles games pick rank game history trade fantasy number player nfl team calls top season
Fiction	letters url fiction lekker met hate pack win pur funny weer rico unit moet nar kick reaction net arv heel
Fun	url check great love free awesome site store food today photos tips party order time songs peek design weekend clothes
Emotions	people bad i'm admit hate love strange make annoy play it's makes time funny feel friends true angry matter good
Location	url i'm philadelphia park city mayor location york philly design box bank photo search ave citizens center opening reveals day
Discussion	url change follow education pregnancyloss computer propulsion cycle lbs item secret money security gas save built boxing vin personal jobs
Curse	fuck bad fucking female person i'm people inappropriate shit ass laugh appeal funny man holy fun real hell hate talking
News	url school news sports high video upper fox lines group great washington wtf today temple darby location blurred weather back
Time	hours number days application time years minutes order ago back url unit late day started top running today shows left
Personal	people life things make love time good rank i'm hard emotion don't happen stay find person feel it's forget change
Religious	god love jesus life bless lord give respect man world good heart christ people day job family sex hope peace
Family	bad people event family person inappropriate call problems man make time feel kids admit makes world making age good thing

Table 5: Topics with high information gain

66.66% of wrong predictions are a miss by one in privacy score and the remaining 33.33% of wrong predictions are a miss by two. Many of the wrong classifications lie on classifier boundaries. For example, one timeline was misclassified as a privacy score of 1, and it actually had 30% private tweets and needed to be classified as a privacy score of 2. We believe such cases can be eliminated by improving the quality of extracted features.

9.1 Future Work

A dataset made up of tweets is a challenging one for text analytics compared to formal writing. Our methods will be more effective on regular writings of people. We would like to test this hypothesis in the future once we are provided a dataset with formal writing and ground truth on private information.

We plan to quantify the relationship between text length and the amount of personal information leakage as we obtain more annotated data. We want to apply our methods to other social media now that we have further experience with privacy annotation to test our ability to detect private content in similar but differently formatted data.

The text analysis software LIWC has dictionaries relevant to privacy. In future work, we would like to incorporate and study the effects of other related LIWC dictionaries on our supervised classifiers.

In future work, we plan to investigate whether a person has a private information disclosure resulting from a direct interaction or an indirect side effect from a person they have not previously had a relationship.

10. CONCLUSION

Some topics are more likely to include private information since topic ratio features are helping us detect private information. Entity recognition by itself is not enough to show if private information is being revealed, but added to topic features which define the context of the entity, it greatly increases the detection rate of private information. Keyword based private information detection helps us detect private information to some extent since privacy dictionary matches feature improves the accuracy by 4%, but it is too limited to be generalized for all privacy concerns.

We will incrementally improve our approach with future work and aim to provide an assistive tool that can be more than a research platform for privacy and security researchers. For example, a user can use privacy detective to have a sense of friends privacy scores to build friends lists accordingly.

Online privacy behavior is socially constructed and this knowledge can be used to effectively design privacy enhancing technologies and target educational interventions.

11. ACKNOWLEDGMENTS

This material is based on work supported by the National Science Foundation under grant 1253418. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not neccessarily reflect those of the National Science Foundation.

12. REFERENCES

[1] https://opennlp.apache.org.

[2] http://alias-i.com/lingpipe. October 2008.

[3] S. Aksoy and R. M. Haralick. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters*, 22(5):563–582, 2001.

[4] D. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4), 2012.

[5] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[6] J. Bollen, H. Mao, and A. Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*, 2011.

[7] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479, 1992.

[8] R. Chow, I. Oberst, and J. Staddon. Sanitization's slippery slope: the design and study of a text revision assistant. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, page 13. ACM, 2009.

[9] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

[10] N. A. Christakis and J. H. Fowler. The collective dynamics of smoking in a large social network. *New England journal of medicine*, 358(21):2249–2258, 2008.

[11] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

[12] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[13] E. D. Cristofaro, C. Soriente, G. Tsudik, and A. Williams. Hummingbird: Privacy at the time of twitter. In *IEEE Symposium on Security and Privacy*, pages 285–299. IEEE Computer Society, 2012.

[14] Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *ICML*, volume 96, pages 148–156, 1996.

[15] A. J. Gill, A. Vasalou, C. Papoutsi, and A. N. Joinson. Privacy dictionary: a linguistic taxonomy of privacy for content analysis. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 3227–3236. ACM, 2011.

[16] M. Hart, P. Manadhata, and R. Johnson. Text classification for data loss prevention. In *Privacy Enhancing Technologies*, pages 18–37. Springer, 2011.

[17] J. R. Landis, G. G. Koch, et al. The measurement of observer agreement for categorical data. *biometrics*, 33(1):159–174, 1977.

[18] J. H. Lau, N. Collier, and T. Baldwin. On-line trend analysis with topic models: #twitter trends detection topic model online. In *COLING*, pages 1519–1534, 2012.

[19] 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)*, 5(1):6, 2010.

[20] H. Mao, X. Shuai, and A. Kapadia. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, pages 1–12. ACM, 2011.

[21] A. K. McCallum. Mallet: A machine learning for language toolkit. 2002.

[22] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390, 2013.

[23] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 2001.

[24] J. C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. *Advances in Kernel Methods Support Vector Learning*, 208(MSR-TR-98-14):1–21, 1998.

[25] A. Ritter, S. Clark, O. Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[26] 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*, pages 3277–3286. ACM, 2013.

[27] K. Thomas, C. Grier, and D. M. Nicol. unfriendly: Multi-party privacy risks in social networks. In M. J. Atallah and N. J. Hopper, editors, *Privacy Enhancing Technologies*, volume 6205 of *Lecture Notes in Computer Science*, pages 236–252. Springer, 2010.

[28] A. Vasalou, A. J. Gill, F. Mazanderani, C. Papoutsi, and A. Joinson. Privacy dictionary: A new resource for the automated content analysis of privacy. *Journal of the American Society for Information Science and Technology*, 62(11):2095–2105, 2011.

[29] 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*, SOUPS '11, pages 10:1–10:16, New York, NY, USA, 2011. ACM.

[30] Z. Xue, D. Yin, B. D. Davison, and B. Davison. Normalizing microtext. In *Analyzing Microtext*, 2011.

APPENDIX

Please select the most appropriate category/categories for each tweet.	
CATEGORY	**DESCRIPTION**
Location	**Giving out location information**
Self	Location information about the writer of the tweet is revealed.
Someone else	Location information about someone else other than the writer of the tweet is revealed.
Medical	**Revealing information about someone's medical condition.**
Self	Medical information about the writer of the tweet is revealed.
Someone else	Medical information about someone else other than the writer of the tweet is revealed.
Drug/Alcohol	**Giving information about alcohol/drug use or revealing information under the influence.**
Self	Drug/Alcohol related information about the writer of the tweet is revealed.
Someone else	Drug/Alcohol related information about someone else other than the writer of the tweet is revealed.
Emotion	**Highly emotional content, frustration, hot states, etc.**
Self	Emotion information about the writer of the tweet is revealed.
Someone else	The writer of the tweet reveals emotion information of someone else.
Personal Attacks	**Critical statements directed at a person, general statements rather than specific.**
Self	The writer of the tweet is showing signs of personal attack.
Someone else	The writer of the tweet reveals that someone else is showing signs of personal attack.
Stereotyping	**Ethnic, racial, etc stereotypical references about a group**
Self	The writer of the tweet is stereotyping.
Someone else	The writer of the tweet reveals that someone else is stereotyping.
Family/Association detail	**Revealing information about family members, or revealing their associations, e.g. ex-partner,mother-in-law, step brother**
Self	The writer reveals family and/or other association details about himself/herself.
Someone else	The writer reveals someone else's family and/or other association details.
Personal details	**e.g., relationship status, sexual orientation, job/occupation, embarrassing or inappropriate content, reveal/explain too much**
Self	Personal details of the writer of the tweet is revealed.
Someone else	The writer of the tweet reveals personal details of someone else other than himself/herself.
Personally Identifiable Information	**Personally identifiable information(e.g., SSN, credit card number, home address, birthdate)**
Self	The writer of the tweet reveals personally identifiable information about himself/herself.
Someone else	The writer of the tweet reveals personally identifiable information about someone else other than himself/herself.
Neutral/Objective	**Neutral or objective tweets that reveal no private or sensitive information.**
Self	The neutral/objective tweet is about the writer of the tweet.
Someone else	The neutral/objective tweet is about someone else other than the writer of the tweet.

Table 6: Tweet Privacy Categories