Detecting and Analyzing Privacy Leaks in Tweets

Paolo Cappellari[1], Soon Chun[1] and Christopher Costello[2]
[1]*College of Staten Island, City University of New York, New York, U.S.A.*
[2]*Macaulay Honors College, City University of New York, New York, U.S.A.*

Keywords: Privacy, Text Analytics, Machine Learning, Social Media.

Abstract: Social network platforms are changing the way people interact not just with each other but also with companies and institutions. In sharing information on these platforms, users often underestimate potential consequences, especially when such information discloses personal information. For such reason, actionable privacy awareness and protection mechanisms are becoming of paramount importance. In this paper we propose an approach to assess the privacy content of the social posts with the goal of: protecting the users from inadvertently disclosing sensitive information, and rising awareness about privacy in online behavior. We adopt a machine learning approach based on a crowd-sourced definition of privacy that can assess whether messages are disclosing sensitive information. Our approach can automatically detect messages carrying sensitive information, so to warn users before sharing a post, and provides a set of analysis to rise users awareness about online behavior related to privacy disclosure.

1 INTRODUCTION

Over the past decade, the rise of social media outlets such as blogs and social networks has provided people with the means to disclose information about themselves publicly. Cybernauts directly or indirectly share information about themselves or others, e.g. friends (Krishnamurthy and Wills, 2009; Mao et al., 2011; Malandrino et al., 2013). People are aware that their online behaviors are, to some degree, monitored in exchange for the fruition of the services. Nevertheless, users tend to neglect the potential implication such information can have on their life, especially when interacting on social networks (Wang et al., 2011). Awareness of the potential implications of online behavior is becoming of paramount importance. People need to realize that repercussions of online behavior are no longer confined to the social aspect of life, but can propagate into the professional one. As an example, in a recent CareerBuilder's annual social media recruitment survey, [1] it has been observed that 70 percent of recruiters use social media to research and consider job candidates; 49 percent of which admitted to having rejected candidates or the content they have shared.

[1]http://press.careerbuilder.com/2017-06-15-Number-of-Employers-Using-Social-Media-to-Screen-Candidates-at-All-Time-High-Finds-Latest-CareerBuilder-Study

Information sharing can be achieved via social posts, chat messages, emails, blogs, etc. Generally, repercussions of online behavior are correlated with the disclosure of sensitive (or private) information about the individual user or others. Privacy is a major challenge in the modern web-oriented society. In this paper, we present an approach to analyze and assess "the amount" of privacy content disclosed by a user when sharing information online. Our approach allows to analyze social media user's activity over a period of time, as well as to assess in real-time whether a piece of information about to be shared contains sensitive information or not. Ultimately, our approach provides the mean to assess information leakage both historically and momentarily: it can be used by users or organizations to detect and assess information leakage before it occurs, or to analyze already occurred leakage in order to educate users.

One major issue in detecting sensitive information is that the definition of what is private (from what is not) varies between individuals: people have different, subjective, views on what constitutes a private content. It is also true, however, that people belonging to the same community tend to share the similar definition of what constitutes sensitive information. In our approach we use a "societal definition" of sensitive information. The societal definition we use in this work has been built by asking groups of people to an-

Cappellari, P., Chun, S. and Costello, C.
Detecting and Analyzing Privacy Leaks in Tweets.
DOI: 10.5220/0006845602650275
In *Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA 2018)*, pages 265-275
ISBN: 978-989-758-318-6

notate social posts into two basic categories: private vs non-private. Each post has been classified by multiple people, whose votes have been democratically coalesced in a final annotation.

When it comes to social media users privacy protection, there is a surprising lack of protection tools and controls related to what Richthammer et al. (Richthammer et al., 2014) described as "semantically unspecified" data (i.e.: social media posts). On one hand, tools such as the browser add-on NoTrace help users against collection of personally-identifiable information, web tracking and other privacy threats. On the other hand, social media platforms provide users with some form of control to define which audience has access posted contents. However, social media sites largely leave post contents to the discretion of their users. According to Sapuppo (Sapuppo, 2012), 14 to 17 percent of social media users are unconcerned with the privacy value of what they share, while only 10 to 17 percent are "fundamentalists" who are extremely reserved with what information they share. The majority of users are "pragmatists" who are somewhat concerned with the information they disclose, but even they are liable to unwittingly disclose some private information.

In the academic setting, a number of efforts are trying to tackle the problem of information leakage. For example, in (Acquisti and Gross, 2006) authors focus on analyzing a number of privacy issues related to the user of social networks, but do not address the problem of prompting the user before an information leakage occurs. The authors of (Mao et al., 2011) can automatically classify social posts in set of privacy categories, however such categories are fixed and have been defined arbitrarily and upfront. The work in (Malandrino and Scarano, 2013) prompt users when sensitive information is about to be disclosed, however it limits its scope to well structured data in emails, name or social status. In (Cappellari et al., 2017) authors propose an initial study to generic privacy detection and assessment, however the approach seems to be in a small scale and geared towards the analysis of past information leakage occurrences. Overall, a comprehensive approach helping users assessing and overcoming information disclosure issues is missing.

With this work, we present a machine learning based approach to automatically detect messages carrying private information. Our main goal is to detect messages carrying sensitive information, so to alert users before the messages are shared with others, and to provide online behavior analysis to rise users awareness regarding online privacy behavior. Our contributions are as follows:

- A semi-automated system to generate annotated datasets, to be used directly as training set for the machine learning model;
- A data pre-processing processor that reduces the text to analyze to its core and essential features;
- A model to detect and assess messages carrying sensitive information;
- A performance comparison of several machine learning models;
- A privacy analytics platform to analyze individual and population level privacy leakage behaviors in social media message sharing.

Our contributions also include that we share a dataset of 6000 tweets annotated with the privacy-related labels with the research community, to enable follow-up research of this study for verification and replications, as well as enhance the machine learning models to detect the privacy-related messages.

The paper is organized as follows: in Sec. 2 we present the related research; in Sec. 3 we describe how we select the data that will be used to build our privacy classification model; in Sec. 4 we present our approach to the crowd-sourced annotation of social posts, to create our societal definition of privacy; in Sec. 5 we present our privacy detection approach; in Sec. 6 we discuss our finding; finally, in Sec. 7 we draw our conclusions.

2 RELATED RESEARCH

The approaches in (Islam et al., 2014; Liu and Terzi, 2010) both focus on associating a user with a privacy score, that indicates her/his tendency to disclose sensitive information. In (Islam et al., 2014) the authors analyze the Twitter users privacy disclosure habits by building a machine learning model that associates each Twitter user with a privacy score. The privacy score is then used to analyze how the user's privacy behavior is influenced by other users in the same network. Similarly to our work, authors creates a machine learning model by using tweets annotated by Amazon Mechanical Turks (AMT) workers. Differently from ours, however, AMT workers are presented with a fixed set of possible privacy categories, therefore limiting the ability of the model to recognize privacy in general. Also, the authors focus on scoring the user privacy leakage behavior by analyzing their timelines, and by correlating the score of the user with the score other users the former has befriended or mention in posts (so to determine a privacy leakage influence pattern). Our work, instead,

focuses on assessing each message individually: each message can be classified in real-time before it is disclosed. Also, by collecting series of messages, analytics can be built to provide insights on the day, time, location, and content of privacy leaks, as well as on the likelihood that a user (or group of users) will disclose sensitive information.

In (Liu and Terzi, 2010), authors propose a framework to associate users with a privacy risk score. The score is determined by analyzing the user's messages, and can be used to: alert the user when a posted message exceed a set privacy risk score, or to let the user know where s/he stands compared to the rest of the community. The approach mostly focuses on devising a mechanism to formalize a mathematical function to calculate the user privacy risk score. Many factors are considered in the mathematical function that calculates the user's score. However, the approach does not account for the societal factor, which is the first major difference with respect to our approach. The second major difference is the attention to the evaluation of the user score after the privacy leak has occurred, similarly to (Islam et al., 2014), rather than on the assessment and prediction of the generic piece of text the user is about to post, like in ours.

In (Mao et al., 2011), authors analyze information leak associated with a number of fixed categories of interest: vacation, drug, and health condition. The first category is concerned with users disclosing plans and/or locations of where they will (or not) be and when. The second, is concerned with people posting messages under the influence. The third, looks into social posts disclosing medical conditions, personal or not. Authors focus on Twitter posts, associating tweets to the mentioned categories by relying on a set of keywords representative of each category. The limit of this approach, and the difference with respect to ours, lies in the small number of categories of information disclosure authors look into, and in the arbitrarily, fixed, and subjective, set of keywords that are associated with each category. Our approach, on the other hand, relies on crowd wisdom to know what should be considered as sensitive information, rather than on a fixed set of keywords. As a consequence, we are looking at a larger spectrum of sensitive information disclosure possibilities, and to a more democratic way of classifying messages which is more aligned with the privacy perception of the society.

In (Sleeper et al., 2013; Wang et al., 2011), authors study what messages users from Twitter and Facebook tend to regret. Authors surveyed a number of users from both platforms to classify regretted posts into categories, and analyze the effort and time users spend in making amends for their posts, when possible. Both works analyze the aftermaths of information disclosure, and focus on educating users on the use of social media and on the implication of underestimating information sharing. Differently, our work is more focused in providing users with insights about privacy leakage, as well as in supporting users with actionable mechanisms that can prevent these (regret) situations from happening altogether.

Hummingbird (Cristofaro et al., 2012) is a Twitter-like service providing users with a high degree of control over their privacy. The service offers a fine grained privacy control, including: the ability to define access control lists (ACL) for each tweet; and the protection of server-side user behavior identification. With this approach, users are limited to a specific service, and have to proactively address the privacy issue by taking actions before using the service itself, such as defining ACL. With our approach, users are free to use any service, do not have to configure any tool, and can assess the amount of information leakage in a message before sharing it.

The work in (Kongsgård et al., 2016) focuses on sensitive information leakage detection for corporate documents. Authors employ machine learning techniques to automatically classify a document as sensitive vs non-sensitive. A curated training set of documents has to be provided to create the classification model: an administrator has to craft, select, and annotate which documents should be considered as private vs not private. This solution can provide great degree of customization, which is ideal of a corporation need. On the other hand, it is impractical on a large scale, which is on what our approach focuses, where an administrator cannot possibly prepare the dataset(s) manually.

3 DATA SELECTION

Selecting the right data is a crucial task for both creating the training dataset for our classification model, and for validating our approach. Due to the lack of available privacy related datasets, we had to devise a mechanism to collect and create our own privacy dataset. Our data source is Twitter, where we have been careful to select only tweets that users have marked as fully public (no restrictions). We have used Twitter as our data source due to its public and openness policy. We have collected millions tweets from the live sample Twitter stream, over multiple periods of time during Fall 2017 & Spring 2018. Part of this dataset, after annotation, has been used to train the machine learning model, and to run privacy leaks analysis on historical data. Note that our application also use the

the live Twitter stream to provide privacy leakage information in real-time.

To guarantee an approach as application agnostic as possible, we decided not to consider any information beyond the tweet text itself and its geo-reference (as latitude and longitude). In fact, all tweets' metadata, user profile information, etc., and all application specific lingo, such as the hashtags for Twitter, have not been considered in our approach. In addition, in order to build a higher quality training set, we focused on the English language, and on tweets of reasonable length, which are more likely to provide a meaningful content. In summary, we have retained tweets with the following characteristics:

- are in English;
- have at least 10 words (beyond stop-words);
- have no hashtags;
- have no URL;
- are not retweets.

The focus on the English language is motivated by the availability of libraries to process the English language and by the fact that adding additional languages would not have improved the generality of our approach. In fact, having to support multiple languages would require more work, but would not add meaningful contribution to the methodology itself. Therefore, we decided to focus on the English language only, thus discarding tweets in any other language.

Filtering on tweets that have at least 10 words, in addition to stop-words, maximizes our chances of collecting tweets with a well formed, meaningful, sentence. Tweets, and social posts in general, can be rather short in nature, which poses a major challenge: they tend to provide little-to-none "surrounding contextual information" about the message it is being shared. The lack of contextual information makes it harder to assess the content of the messages for privacy leaks. Therefore, we decided to discard all those messages that are too short, and would not only provide little value, but could potentially generate undesired "noise" in our privacy assessment approach, resulting in improper classification.

In tweets, hashtags are metadata information that users embed in their messages. Hashtags are a simple, yet effective, way for users to annotate their messages with a theme or topic, so it is easier to search and follow social trends for such topic. While hashtags can provide a very valuable information on the topic and context regarding the message, and can be used to assess privacy leaks, we have decided to ignore them. This way, we do not rely on any lingo or feature of a specific application, Twitter in this case, thus developing an application agnostic approach that can be used with any application and in any context.

Filtering out URLs maximizes the chances that each message is self-contained, thus does not relying on information located somewhere else on the web. Such linked information should be considered as fully part of the message, and included in the privacy leak analysis, ideally However, URLs can link to heterogeneous resources, such as images or videos which, while potentially of paramount relevance, pose a completely different challenges from a sensitive information disclosure analysis point of view, which are beyond the scope of this work.

Finally, we skip retweets because from a privacy leak point of view they are duplicate information that do not provide additional value. As a result of our data selection criterion, we are able to generate a generic privacy dataset that allow us to build an application independent privacy leak classification model.

4 DATA ANNOTATION

Our privacy assessment approach rely on a machine learning model. In order to use such models we have to craft a so called *training set* of annotated data in order to build the classification model.

Since we want our model to assess each message in isolation, the training set is composed of a set of messages, where each message is annotated with either the following tags: private, or non-private. The first tag denotes a message that is disclosing sensitive information; the second one states the opposite, that is no privacy is leaked in the message.

From our collected data we selected just more than 6000 messages, satisfying the criterion described in Sec. 3. These messages have then been manually annotated as *private* or *non-private* by Amazon Mechanical Turk (AMT) workers. In order to minimize the risk of an arbitrary annotation of messages (which would introduce bias in the model), each message was annotated by 5 different workers, where the final annotation for such message is decided by a majority vote: if at least 60% of the workers have tagged the message as private, then message is deemed as private; otherwise, the message is deemed as not-private.

When annotating a message, a worker is presented with 3 mutually exclusive options to choose from:

a) `not-private`, the message does not contain sensitive information;

b) `somewhat-private`, the message discloses sensitive information, to some extent;

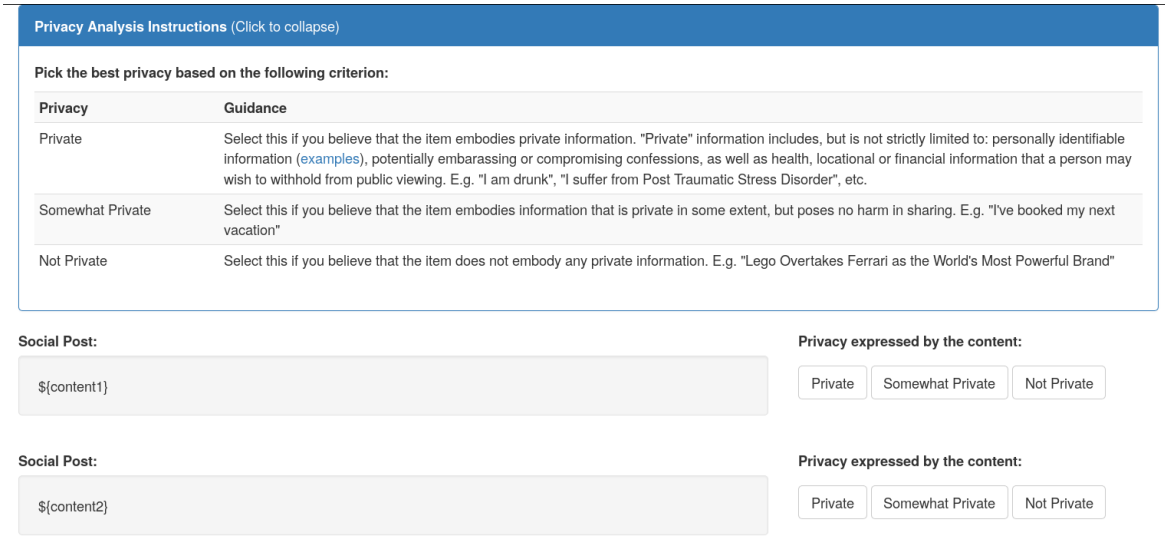


Figure 1: HTML custom template for posts annotation for Amazon Mechanical Turks.

c) `private`, the message discloses sensitive information, definitively.

The first and third options, `not-private`, and `private`, are intended for messages that without any doubt are either not disclosing or disclosing sensitive information, respectively; the second option, `somewhat-private`, is meant to capture that gray area where users, in their subjective definition of privacy, are not sure themselves on whether the message is or is not disclosing sensitive information. At annotation election time, the `somewhat-private` tag is equipped to `private`, meaning that both counts as a `private`. In fact, whether the user is ready or not to admit it to herself, if she considers a piece of information to be "somewhat private," then that message is carrying private information, indeed, thus should be distinguished from the non-private ones.

We have developed an AMT custom HTML template, and a program to automatically convert collected messages between ours and AMT format. In addition to speeding up the annotation process, this also allows us to easily and semi-automatically extend our training set, if/when necessary. The HTML template is designed to presents AMT workers with 12 messages at the time on a single web page. Fig. 1 illustrate the top part of our custom template, including 2 of the possible 12 messages. We chose 12 because we believe that it is the right trade-off that guarantees a level of consistency among the answers of a single task (for a single worker), while also keeping the task itself reasonably short, thus not tedious for the worker. The template is also supplemented with the definitions of our 3 possible annotation tags, each supported with an example.

To maximize the quality of the annotations, we have restricted the set of workers that are allowed to work on tasks. Tasks are made available to workers that satisfy the following criteria:

- are AMT masters, and have a history of task approval rate above 90%;
- are English speakers from these countries: USA, UK, Ireland, Canada, Australia, New Zealand.

Messages to be annotated have been posted on AMT in batches of incremental sizes. The first batch was a test batch of about 100 messages, so that we could verify the initial results, detect and correct any potential issue with the annotation process. Subsequently, we incremented the batch size to 300; and finally to 5600. To further guarantee the quality of the annotation results, we injected the 6000 messages to tag with an additional set of 140 for which the annotation was already known. This set of 140 pre-annotated tweets has been constructed in the same manner as the primary dataset, however it has been manually inspected for quality assurance. Tasks for AMT workers would include some of these 140 verified messages, so we could determine whether their annotations were sound.

The turnaround time for individual tasks, from publishing to results availability, has been less than a week. Overall, including our incremental batch tests, it took about one month to have all 6000 tweets annotated. The resulting annotated privacy dataset can be used in many different contexts to improve people understanding of privacy in the modern information sharing oriented society.

5 PRIVACY LEAKAGE MODELING

We tackle the sensitive information disclosure assessment problem by adopting a supervised machine learning approach. Messages are classified individually according to our two categories of interest discussed in earlier sections: private, and non-private. Several classification models exit, each with pros and cons. In choosing which model to use, we followed a pragmatic approach: we tested the most popular models for text classification and compared their performance. Then, we selected the model showing the best performance to develop our privacy assessment application as detailed below. Regardless of the model, a number of data pre-processing steps had to be put in place in order to maximize the quality of the result.

5.1 Model Training Set

Results from AMT annotation process were reasonably balanced. Out of the 6000 messages, about 60% were labeled as `not-private`, and the remaining 40% as `private`. A training a model with such a dataset would generate little-to-no bias in the classifier, thus providing a balanced prediction model. Therefore, we did not sample the dataset to reduce the data to a 50-50 perfectly balanced training set between private and not-private tweets.

5.2 Text Preparation

Social media users often use a less formal version of a language. In Twitter, for instance, messages are contracted to a more succinct form because of the platform nature (micro-blogging). In doing so, users often resort to lingo, abbreviations, and other sort of "broken" versions of a language.

To create a model that is resilient to these variations, we pre-process collected data so that the messages are reduced to a more uniform and basic form. This text pre-processing is applied to both the training data and the actual input. In this pre-processing, messages undergo the following transformations:

- the text is converted to lowercase;
- common and stop words are removed;
- letters are converted to basic ASCII alphabet (e.g. no accents, etc.);
- contractions are collapsed into one word, e.g. "I'm" is reduced to "Im";
- words are stemmed to remove derived or inflected variations;
- the less statistically relevant terms are removed.

These data transformations have been implemented in a series of cascading scripts developed in R. Lower-casing, removal of numbers, punctuation, extra white-space, and stop-words is achieved via common standard libraries; ASCII conversion via the *stringi* package. Stemming is applied to reduce inflected variation of words to their common root. In doing so, we are able to reconcile multiple words to a single semantic, thus increasing the reach of our vocabulary. Finally, less relevant terms are removed to further improve correct classification as follows: the words are filtered into a document-term matrix (DTM) where the text of each message counts as a document, and the term frequency is measured by term frequency-inverse document frequency (TF-IDF): the terms that appear in less than 0.1% of the documents are removed altogether.

5.3 Model Training and Assessment

The 6000 labeled Tweets are randomly sampled to create two partitions: a training set partition, containing 80% of the messages, to be used to train the machine learning model; and a validation set containing the remaining 20% of messages (i.e. 1200 labeled tweets) to be used to evaluate the quality of the model.

A variety of classification models have been developed over the past number of years. Each has pros and cons, and performs differently depending on the contextual settings. To decide which classification model to adopt in our work, we followed a pragmatic approach: instead of tampering with state-of-the-art algorithms, we have tested the most popular models and selected the best performing one for our case. The models we include in our evaluation are the following: Support Vector Machine (SVM), the Generalized Linear Model (GLM), the Maximum Entropy (MAXENT), the Supervised Latent Dirichlet Allocation (SLDA), as well as Bagging (BAGGING), Boosting (BOOSTING), Decision Tree (TREE), and Random Forest (RF) models. The models have been trained and tested against the same training set. The results are illustrated in Table 1

The best performing model resulted to be SVM, with an accuracy of 70%, roughly. Therefore, we created our privacy model using SVM, which is used in our applications detailed later.

5.4 Privacy Topics

Besides detecting whether a message discloses sensitive information, it is also of interest understating what kind of information is being disclosed, or at least

Table 1: Classification Models Accuracy, Precision and Recall Performance.

Model	SVM	GLM	MAXENT	SLDA	BOOSTING	BAGGING	RF	TREE
Accuracy	68.5%	66.4%	61.0%	66.7%	53.4%	58.6%	67.6%	48.6%
Precision	68.2%	64.1%	60.5%	65.5%	51.1%	54.7%	69.6%	48.6%
Recall	66.1%	70.4%	57.5%	67.1%	96.4%	87.0%	59.5%	100%

on which topic. Topic modeling (Blei, 2012) and topic identification (Stein and zu Eissen, 2007) are two research areas that try to cluster documents in categories of topics, and to associate a label to each identified category, respectively.

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a popular statistical model to extract a set of topics (categories) from an unstructured set of documents. Briefly, the intuition behind LDA is that a set of documents only covers a limited set of topics, that is the documents can be classified in a few categories, and that only the same small set of words is used frequently in each topic.

We use LDA to analyze the topics of privacy disclosure across the users' timelines, and geographical areas. Currently, we do not try to identify the topic: we limit our attention to find the most occurring words on each topic, delegating the labeling of the discovered topics to users.

6 APPLICATIONS FOR PRIVACY ANALYTICS

Analyzing the set of messages annotated from AMT workers we have observed that almost half of the messages discloses private information. This fact alone proves that sensitive information is being disclosed rather frequently, which reinforce the motivation for our work. Clearly, users should be provided with supporting mechanisms to catch these information leaks automatically, giving users both a protection against possible "share regrets" episodes, as well as an educational tool.

We have developed a set of applications that can assess users privacy leak behaviors by classifying tweet messages as disclosing sensitive information or not. Specifically, we developed a web application composed of multiple services: a text privacy assessment service, a user disclosure behavior analysis, a geo-area disclosure analysis, and user disclosure topic analysis. All our applications, including the privacy dataset are available on our web-server[2]. The first is a tool that users can use to assess their message for sensitive information before posting it. The other services are detailed below and validate the quality of our approach. At the core, the application is developed in R, running on a Shiny server. Eventually, we would like to release this service as a browser plug-in so that users can have immediate privacy assessment without having to copy-and-paste in third-party web pages before posting their messages.

6.1 Assessing a Post for Privacy Leak

Before sharing a post, the user can use our application to assess a message for privacy content. The application is a simple web-page with an input text field and a submit button. Fig. 2 illustrates an example for a sample post. By submitting the message, the user is prompted with a report deeming the messages as private, or not, and with a degree of confidence regarding the classification. Under the hood, the application performs the data pre-processing as described in Sec. 4 before passing the text to our privacy model. The confidence value provide with the annotation, informs the user how accurate the classification is. Simplifying, the user can interpret this information as the amount of sensitive information the post would disclose, if published.

6.2 User Disclosure Behavior

The *User Behavior* service analyzes a user timeline to assess the behavior of such user. The application scans through the user's messages, classifies each message as either private or non-private, and return an aggregate view on how much, and when, the user was disclosing sensitive information.

More in detail, the application expects a user Twitter handle as input; then, it fetches in real-time all the messages posted by the user, retrieving as many tweets as allowed by the Twitter API (currently, about 1500). Retrieved tweets are then passed to our classification model, where each tweet is assessed individually for sensitive information content. The contribution of each message is aggregated to provide a user profile that reveals: how much of the user tweets contains sensitive information, as a percentage of all tweets, and when such tweets are shared on the platform. The application presents a breakdown of which day of the week and which hour of the day the user has shared tweets with sensitive information. Fig. 3 illustrates an sample run of this application for the Twitter handle: @realDonaldTrump.

[2]http://isi.csi.cuny.edu/privacy

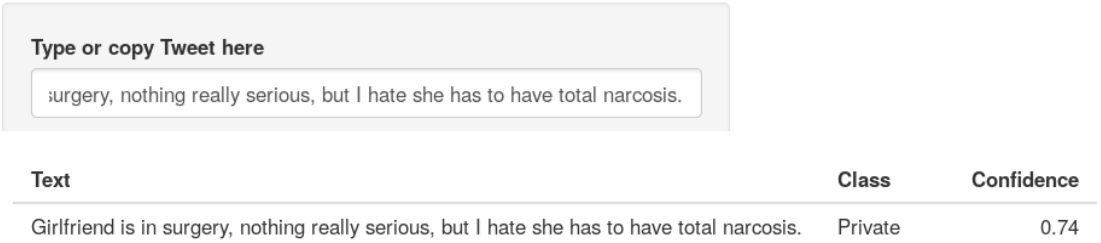


Figure 2: Privacy assessment evaluation for an individual post.

At the top of the illustration, there is a pie chart summarizing the tendency of the user to disclosure (or not) sensitive information. Right below it, there is a stacked-bar chart providing information on how much sensitive information is disclosed in which day of the week. Finally, the application also presents the breakdown by the time of the day, using an horizontal stacked-bar chart. For the specific user we can observe a 51-49 ratio, roughly, in terms of amount of messages carrying sensitive information disclosed. Peak days are Tuesday through Friday, with peak hours occurring at the start and end of the day.

6.3 Geo-area Disclosure Analysis

The *Geo-area Service* collects and assess all messages within a defined geographic area. The geo-area is defined as a "box" of four points indicating the North-East, North-West, South-West, and South-East corners. Each box's corner is a pair latitude and longitude. Tweets are analyzed by sub-area, where a sub-area can be a country, a state, a county, or a neighborhood. Fig. 4 illustrates two examples of the geographical area analysis, on the world (Fig. 4a), and on New York City, USA (Fig. 4b).

In the application page, the slider at the top allows to browse through the hours of the day, so to see the amount of sensitive information disclosure varies in countries, on the left, or neighborhoods, on the right during the day. The lightest the blue color in an area in the figure, the higher the percentage of messages classified as carrying sensitive information (for such area). During our experiments, we have observed that sometimes there are areas for which (almost) 100% of the messages are disclosing sensitive information. This, while possible, is more likely caused by the an unrepresentative sample of messages, which is due to the limitations of the free Twitter API. The latter, in fact, only returns a small percentage of all the tweets posted in a selected area which, by chance, all happen to be labeled as private by our classification model.

6.4 Topic Analysis of Privacy Disclosure

With the *Topic Analysis service* we want to understand which topics a person is more likely to share sensitive information on. The application pulls the user timeline (currently limited to the last 1500 tweets) and tries to assess the topic for each message. A major challenge with Twitter messages is that they contain little amount of text, thus making it very hard for statistical methods to reliably classify the topic. Furthermore, while methods such as LDA allow to model the topics, identifying (labeling) the topics themselves is an additional challenge. Therefore, in the application we limit our attention to discover the topics, leaving the user with the task of understanding and labeling them. Fig. 5 illustrate a use case of this analysis.

In the figure (top-right) we can see a number of topics, one per column, where the most representative (key)words are listed within each topic. An observer can (arbitrarily) identify the first topic as about "fake news", the second about the "presidential motto", the third about jobs, and so on. At the top of each keyword list we leave an editable input field, initially populated with the word "Topic-#" (with # a number), so that the user can provide a label for such topic, if s/he wishes. For future work, it is our intention to investigate this topic labeling challenge, possibly in the same fashion we crowd-sourced the private vs non-private annotation of tweets, so to develop a AI approach to topic label identification.

7 CONCLUSION

We have presented an approach to automatically detect sensitive information disclosure in social posts. The approach relies on supervised machine learning algorithms to assess in real-time whether a text message carries sensitive information so that the user can be alerted before the message is shared, therefore protecting the user from sharing regrets episodes. In addition to individual message assessment, the appro-

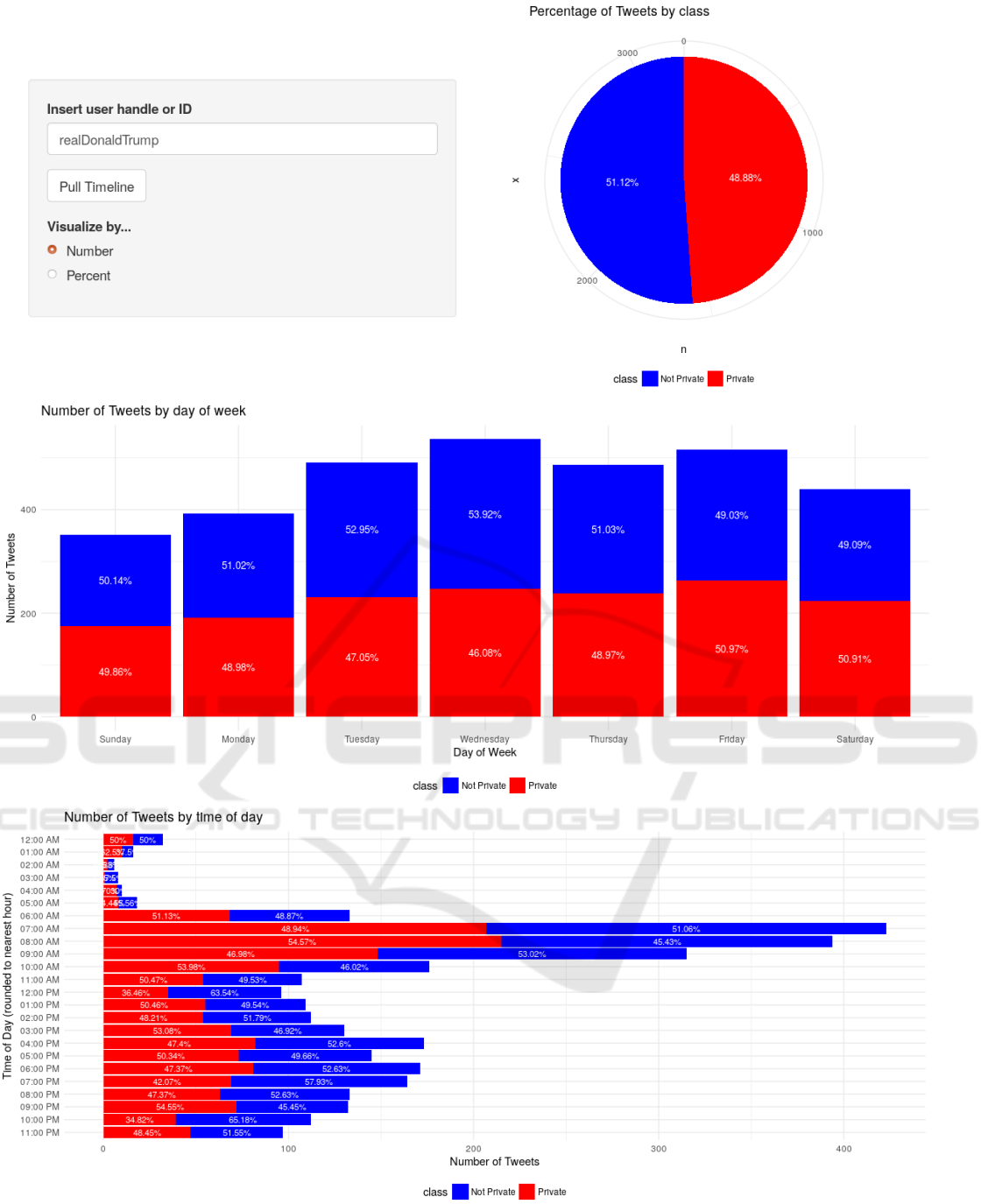
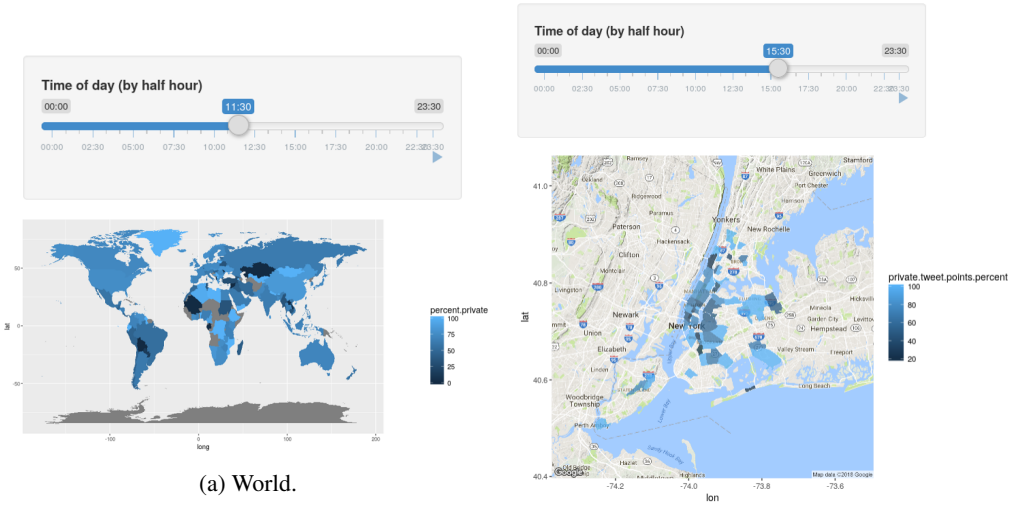


Figure 3: User disclosure behavior evaluation.

ach can also mine through a user timeline or a geographical area to summarize the online behavior with respect to privacy leaks, where summary information shows the online habits of the user(s), detailing when and where sensitive information is mostly disclosed. Finally, the approach allows the user to analyze the topics to which her social posts belong to, so to understand what kind of sensitive information is disclosed.

As part of out contributions, we have tested multiple machine learning models, we have provided a semi-automatic procedure to build the training set required by the supervised learning model, and we have published the set of privacy annotated data.

For the future work, we would like to tackle the following problems. First, we want to explore methods to improve the privacy classification for very



(a) World.

(b) New York City.

Figure 4: Snapshots for the world and the New York City area.

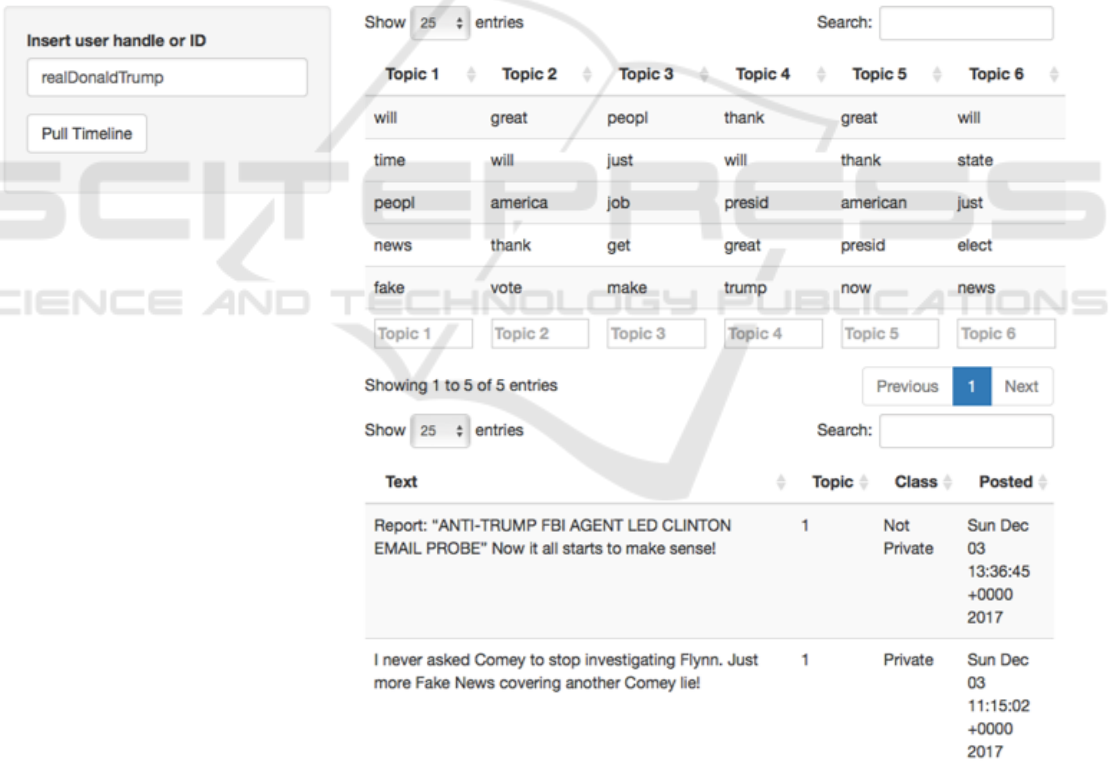


Figure 5: Topic disclosure analysis for an individual user.

short text: this is a challenging problem because the less the text, the less reliable the supervised model is. Second, we want to automate the identification and labeling of the topics of privacy disclosure, so to further refine the support we can provide to the users. Finally, we want to build an open platform for the continuous refinement of the privacy annotation mechanism, were to publish social posts and collect people's annotations, to create an automatic self-improving system for privacy classification.

REFERENCES

Acquisti, A. and Gross, R. (2006). Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Proceedings of the 6th International Conference on Privacy Enhancing Technologies*, PET'06, pages 36–58, Berlin, Heidelberg. Springer-Verlag.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Cappellari, P., Chun, S. A., and Perelman, M. (2017). A tool for automatic assessment and awareness of privacy disclosure. In *Proceedings of the 18th Annual International Conference on Digital Government Research, DG.O 2017, Staten Island, NY, USA, June 7-9, 2017*, pages 586–587.

Cristofaro, E. D., Soriente, C., Tsudik, G., and Williams, A. (2012). Tweeting with hummingbird: Privacy in large-scale micro-blogging osns. *IEEE Data Eng. Bull.*, 35(4):93–100.

Islam, A. C., Walsh, J., and Greenstadt, R. (2014). Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society, WPES 2014, Scottsdale, AZ, USA, November 3, 2014*, pages 35–46.

Kongsgård, K. W., Nordbotten, N. A., Mancini, F., and Engelstad, P. E. (2016). Data loss prevention based on text classification in controlled environments. In *Information Systems Security - 12th International Conference, ICISS 2016, Jaipur, India, December 16-20, 2016, Proceedings*, pages 131–150.

Krishnamurthy, B. and Wills, C. (2009). Privacy diffusion on the web: A longitudinal perspective. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 541–550, New York, NY, USA. ACM.

Liu, K. and Terzi, E. (2010). A framework for computing the privacy scores of users in online social networks. *ACM Trans. Knowl. Discov. Data*, 5(1):6:1–6:30.

Malandrino, D., Petta, A., Scarano, V., Serra, L., Spinelli, R., and Krishnamurthy, B. (2013). Privacy awareness about information leakage: Who knows what about me? In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*, WPES '13, pages 279–284, New York, NY, USA. ACM.

Malandrino, D. and Scarano, V. (2013). Privacy leakage on the web: Diffusion and countermeasures. *Computer Networks*, 57(14):2833 – 2855.

Mao, H., Shuai, X., and Kapadia, A. (2011). Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society, WPES 2011, Chicago, IL, USA, October 17, 2011*, pages 1–12.

Richthammer, C., Netter, M., Riesner, M., Sänger, J., and Pernul, G. (2014). Taxonomy of social network data types. *EURASIP J. Information Security*, 2014:11.

Sapuppo, A. (2012). Privacy analysis in mobile social networks: the influential factors for disclosure of personal data. *IJWMC*, 5(4):315–326.

Sleeper, M., Cranshaw, J., Kelley, P. G., Ur, B., Acquisti, A., Cranor, L. F., and Sadeh, N. M. (2013). "i read my twitter the next morning and was astonished": a conversational perspective on twitter regrets. In *2013 ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '13, Paris, France, April 27 - May 2, 2013*, pages 3277–3286.

Stein, B. and zu Eissen, S. M. (2007). Topic identification. *Journal of Knstliche Intelligenz*, 21(3):16–22.

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., and Cranor, L. F. (2011). "i regretted the minute I pressed share": a qualitative study of regrets on facebook. In *Symposium On Usable Privacy and Security, SOUPS '11, Pittsburgh, PA, USA - July 20 - 22, 2011*, page 10.