Qiaozhi Wang, Hao Xue, Fengjun Li, Dongwon Lee, and Bo Luo*

#DontTweetThis: Scoring Private Information in Social Networks

Abstract: With the growing popularity of online social networks, a large amount of private or sensitive information has been posted online. In particular, studies show that users sometimes reveal too much information or unintentionally release regretful messages, especially when they are careless, emotional, or unaware of privacy risks. As such, there exist great needs to be able to identify potentially-sensitive online contents, so that users could be alerted with such findings. In this paper, we propose a context-aware, text-based quantitative model for private information assessment, namely *PrivScore*, which is expected to serve as the foundation of a privacy leakage alerting mechanism. We first solicit diverse opinions on the sensitiveness of private information from crowdsourcing workers, and examine the responses to discover a perceptual model behind the consensuses and disagreements. We then develop a computational scheme using deep neural networks to compute a context-free PrivScore (i.e., the "consensus" privacy score among average users). Finally, we integrate tweet histories, topic preferences and social contexts to generate a personalized context-aware PrivScore. This privacy scoring mechanism could be employed to identify potentially-private messages and alert users to think again before posting them to OSNs.

Keywords: Social Networks, Privacy

DOI 10.2478/popets-2019-0059
Received 2019-02-28; revised 2019-06-15; accepted 2019-06-16.

1 Introduction

In the wake of Facebook data breach scandal, users begin to realize how vulnerable their personal data are and how blindly they trust the online social networks (OSNs) by giving them an inordinate amount of private data that touch on every area of their lives. Moreover, social networks fundamentally encourage users to share their privacy more or less to improve their presence in the virtual world. As a lot of private information are buried in the text format postings, human stalkers or automated bots could navigate/crawl historic posts to re-assemble scattered pieces of sensitive information.

Regrettable posts are not seldom posted according to surveys [66, 82]. Meanwhile, the side effect of some posts is often neglected until it is too late to regret. Even the most privacy-savvy users are likely to post something aggressive or divulge too much information. Even worse, although for most of the users, their posts are only intended to be shared with friends/followers, the audience of OSNs is significantly larger than users' expectation, which includes advertisers, recruiters, search engine bots, etc. Therefore, it is critical to automatically identify potentially sensitive posts and alert users before they are posted, i.e., #DontTweetThis.

Conventional privacy protection mechanisms on data or OSN mainly focus on the protection of individuals' *identities* or *private attributes* [9, 18, 28, 47, 53, 71, 88]. However, according to a survey in [30], only 0.1% of users mentioned identifiable attributes such as email addresses or phone numbers in their tweets. Therefore, leaking identities or identifiable attributes during normal socialization is not the only privacy concern in OSNs. On the contrary, since the offline identities of OSN users are often known to their online friends, especially in strong-tie oriented OSNs such as Facebook, *sensitive or inappropriate content* is truly at risk due to careless or unintentional disclosure during socialization. Therefore, we argue that another key component in privacy protection in OSNs is protecting *sensitive/private content*, beyond the protection of identities and profile attributes, i.e., *privacy as having the ability to control the dissemination of sensitive information.*

Meanwhile, friends may leak one's private information. Threats from within users' friend networks – insider threats by human or bots – may be more concerning because they are much less likely to be mitigated through existing solutions, e.g., the use of privacy settings [35, 73, 75, 87]. Therefore, a mechanism

Qiaozhi Wang: The University of Kansas, E-mail: qzwang@ku.edu
Hao Xue: The University of Kansas, E-mail: xhao297@ku.edu
Fengjun Li: The University of Kansas, E-mail: fli@ku.edu
Dongwon Lee: The Pennsylvania State University, E-mail: dongwon@psu.edu
***Corresponding Author: Bo Luo:** The University of Kansas, E-mail: bluo@ku.edu

to distinguish potentially sensitive/private posts before they are sent is urgently needed. Such a mechanism could also benefit non-human users such as social media chatbots. For instance, Microsoft's Twitter bot, Tay, started to deliver racist and hateful content soon after it was launched in 2016. Tay "learned" from inappropriate messages it had received. Unfortunately, there did not exist a mechanism to assess the sensitiveness of tweets before they were exposed to Tay or posted by Tay.

In this paper, we present the first quantitative model for private information assessment, which generates a PrivScore that indicates the level of sensitiveness of text content. We examine users' opinions on the levels of sensitiveness of content, and then build a semantic model that comprehends the opinions to generate a *context-free PrivScore*. The model learns the sensitiveness of the content from text features (e.g., word embeddings) and sentiment features using a Recurrent Neural Network (RNN). To further personalize PrivScore and make it aware of the societal context, we integrate the topic-based personal attitudes and the trending topics into privacy scoring, to generate the *personalized PrivScore* and the *context-aware PrivScore*, respectively. With intensive experiments[1], we show that PrivScores are consistent with users' privacy perceptions.

PrivScore, to the best of our knowledge, is the first quantitative assessment for sensitive content. It has the potential to be utilized in various applications: (1) It could be adopted by individual users for self-censorship and parental controls, to prevent highly sensitive content from being posted to online social networks, especially when the users are careless or emotional. (2) PrivScore could be integrated with AI-based interactive agents, especially the ones with learning capabilities, such as social media chatbots (Twitterbots, Dominator) and virtual assistants (Siri, Alexa, Cortana), to evaluate the content before delivering to users. (3) PrivScore could be aggregated over a large population (across demographic groups, friend circles, users in an organization, etc.) to examine privacy attitudes from a statistical perspective. This method and the results could be used for research purposes, assisting policy making, or privacy education/training.

The contributions of this paper are threefold: (1) We collect the privacy perceptions from a diverse set of users, and examine the consensuses in the responses to model the sensitiveness of content. (2) We make the first attempt to develop a computational model for quantitative assessment of content sensitiveness using deep neural networks. The context-free privacy score resembles the "consensus" perception of average users. (3) We further integrate social contexts and topic-specific personal privacy attitudes to extend the predictive model to generate context-aware and personalized privacy scores.

The rest of the paper is organized as follows: Section 2 formally introduces the problem and briefly presents the solution. Section 3 explains the data collection and annotation processes, followed by the context-free, context-aware and personalized PrivScore models in Sections 4, 5, and 6. We present the security analysis and discuss the performance, usability, and limitations in Section 7. We then summarize the literature in Section 8 and finally conclude the paper in Section 9.

2 Problem Statement and Solution Overview

2.1 The Problem and Challenges

Threat Model. In this work, we aim to protect social network users (and chatbots) from accidentally disseminating *any type of inappropriate content*, especially the private or sensitive information about themselves. We mainly consider the risk of inappropriate dissemination to two types of audience: (1) followers or friends (insiders), who receive updates of the user's posts; (2) stalkers or strangers, who peek into a target user's social network posts. Both are likely to know the offline identity of the user. We do not focus on protecting identities or attributes (e.g., location), since they have been intensively studied in the literature. We do not block the user from publishing the (sensitive) content or block the receiver from viewing the content, instead, we provide an alert to assist users' decision making. We assume the adversaries can browse the OSN through the user interface or collect data using an automated crawler through the OSN's API, i.e., there is no hacking or phishing. Finally, we do not consider the retraction of previous posts.

Objectives. The objective of this work is *to develop a computational model to quantitatively assess the level of privacy/sensitiveness of unstructured text content, which will be further adjusted to reflect the impacts of societal contexts and personal privacy attitudes.* We make the first attempt to generate a *privacy score* (PrivScore) for each short text snippet, e.g., a tweet (limits to 280

[1] Since the experiments involve human subjects, we have obtained an IRB approval.

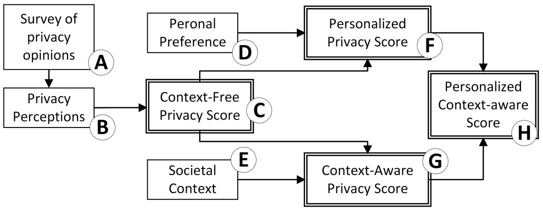

Survey of privacy opinions
A
Peronal Preference
D
Personalized Privacy Score
F
Privacy Perceptions
B
Context-Free Privacy Score
C
Personalized Context-aware Score
H
Societal Context
E
Context-Aware Privacy Score
G

Fig. 1. Key components of the three-phase privacy scoring model.

characters), to reflect its level of sensitiveness within its societal context. The privacy scoring mechanism is expected to serve as the foundation of a comprehensive privacy monitoring and user alerting solution.

The problem is very challenging due to several factors: (1) privacy or sensitiveness is a very subjective perception [17, 31]. Due to the peculiarity, complexity and diversity of human cognition, it is difficult to precisely capture the privacy calculus model for each individual, and generate a consensus privacy score that is unanimously agreed by all users. (2) Text understanding and natural language processing is still an area with active ongoing research. In particular, modeling and understanding of unstructured, short, and non-standard text snippets, such as microblogs, is very difficult. (3) The subjective perception of privacy is often influenced by many factors, such as personal privacy attitude, emotions, societal context, culture, etc. The complexity of modeling the influencing factors is also excessive.

2.2 Overview of the Proposed Solution

We propose a three-phase privacy scoring model (Fig. 1): (1) context-free privacy scoring, (2) context-aware privacy scoring, and (3) personalized privacy scoring.

1. The *context-free PrivScore* is an autonomous assessment of the degree of sensitiveness of short, unstructured text snippets purely based on the textual content, i.e., free from its context. We first collect data about users' opinions on the level of sensitiveness of potentially private content through a user survey on Amazon Mechanical Turk (MTurk) (Fig. 1 (A)). We then statistically analyze the responses to identify the consensus of human perceptions (Fig. 1 (B)). Based on the results, we develop a deep learning model with word embedding and long-short term memory to model text content to develop a privacy scoring mechanism (Fig. 1 (C)).

2. In *Context-aware privacy scoring*, we model the influence of the societal context and incorporate it into PrivScore. We observed that privacy perceptions are dynamic and influenced by societal contexts. In particular, when a popular topic triggers significant interests/discussions in OSNs, users become less concerned about its sensitiveness. For example, the political attitude is normally considered private. When one billion tweets were posted about the election during the midterm elections in 2018, the degree of sensitiveness of political content implicitly decreases from the non-election days. The context-aware PriScore model measures the influence of societal context using the volume, duration, and relevance of trending topics (Fig. 1 (E)). We integrate it into context-free PrivScore to reflect the societal influence on privacy perceptions (Fig. 1 (G)).

3. The *Personalized privacy score* adjusts PrivScore for each user with a personalized topic-specific attitude. We recognize that the privacy perception is subjective and differs for each user, who has her own level of tolerance in private information disclosure. Individual privacy perception is shaped by various psychological factors such as personality and emotion [55]. To provide privacy alerts that are customized for each user, we first analyze her activity history to discover her topic-based privacy attitude (Fig. 1 (D)). A personalized privacy scoring model is then developed to integrate personal attitudes into context-free PrivScore (Fig. 1 (F)).

Eventually, we develop a computational model for *personalized context-aware PrivScore* (Fig. 1 (H)). A PrivScore is generated for each social network post (such as a tweet) to reflect a quantitative assessment of the estimated sensitiveness. The scoring mechanism could be adopted for individual users or integrated with AI-based interactive bots. For instance, when a user attempts to post a tweet with sensitive content that is detected by the proposed mechanism, the user will be alerted that it might become a regrettable tweet, i.e. #DontTweetThis. This warning message intends to trigger self-censorship [65]. A delayed posting mechanism suggested in [82] could be invoked, especially for sensitive posts written under strong emotions.

3 Data Collection, Annotation, and Analysis

3.1 Data Collection

We selected Twitter as the OSN platform because of its openness and popularity. We performed a snowball crawling process in March 2016 for about a month and collected 31,495,500 tweets from 29,293 users. We eliminated non-English speaking users, and tweets beginning with "RT @", since forwarded articles and re-tweets do not contain private information of the forwarder.

It is impractical to ask annotators to label 31 million tweets. Meanwhile, since the dataset is highly imbalanced, if we randomly sampled tweets for labeling, the majority of the samples would be non-sensitive. Therefore, we selected potentially sensitive tweets as candidates for labeling to save labor and cost. Note that our goal is not to develop an accurate classifier in this step. Instead, we aim to construct a balanced dataset by eliminating most of the clearly non-private tweets.

First, we referred to [33, 50, 78, 91] to identify potentially private topics, such as *Health & Medical, Drugs & Alcohol, Obscenity, Politics, Racism, Family & Personal*. For each topic, we selected a root set of "seed terms" and expanded the set using *Urban Dictionary*, an Internet dictionary containing slang words and abbreviations (frequently used in Twitter). After proper cleaning, we collected more than 100 terms for each topic. Terms are available at http://bit.ly/privscore. By selecting a relatively large set of keywords, we aim to increase recall, i.e., to include a majority of potentially private tweets. We then filtered all the tweets with the candidate terms. In total, we extracted 6,917,044 candidate tweets (i.e., 21.9% of the crawled tweets) that contained at least one of the terms.

In order to collect testing samples that are irrelevant to the training data and add trending topic information (for societal context modeling) into the dataset, we performed a second crawl in March 2018. We monitored the *trends* at a 15-minute interval, and recorded the corresponding *tweet_volume* [74]. In total, 1,130 trending topics with volume larger than 10,000 were collected. We also collected 8,079 new tweets from the same set of users that we crawled in 2016. This new dataset is used later to evaluate our privacy scoring approaches.

3.2 Data Labeling

Keyword spotting achieved high recall but low precision in identifying sensitive tweets. Hence, we further collected opinions from a large number of users through a crowd-sourcing platform Amazon Mechanical Turk.

We sampled from 6M potentially sensitive tweets to generate questionnaires of 20 tweets each. The number of tweets containing each keyword conforms to Zipf's distribution [36]. To ensure that less frequent terms still get represented in the labeling set, we used a biased sampling process (i.e., using a biased die) that gives higher probabilities to rarer terms.

Turkers (English speakers in the US, with 95%+ approval rates) were asked to annotate each tweet as: [1:Very sensitive]; [2:Sensitive]; [3:Little Sensitive]; [4:Maybe]; [5:Nonsensitive]. That is, a score $s_t \in \{1, ..., 5\}$ is assigned to each tweet by a Turker. Note that we did not use the standard 5-level Likert: [2:very-sensitive][1:Sensitive][0:neutral/undecided] [-1:nonsensitive][-2:very-nonsensitive], because it is hard to judge between [-1] and [-2] in the Likert scale, i.e., to tell if a tweet is "more non-sensitive" than another.

Table 1. Interrater Agreement based on Fleiss' Kappa, Pearson, and Spearman (P:Poor; Sl:Slight; F:Fair; M+:Moderate+; VW:Very Weak; W:Weak; M:Moderate; St+:Strong+)

Fleiss' Kappa			Pearson			Spearman		
P	$k < 0$	12	VW	$r < .2$	35	VW	r<.2	37
Sl	[0, .2)	353	W	[.2, .4)	125	W	[.2, .4)	158
F	[.2, .4)	175	M	[.4, .6)	240	M	[.4, .6)	249
M+	[.4, 1]	12	St+	[.6, 1)	152	St+	[.6, 1]	108

Each Turker was paid $0.45 per questionnaire. For attention check, we embedded two non-random questions in each questionnaire, which were selected from two very small sets of clearly non-sensitive or very sensitive tweets, e.g. Q16 (non-sensitive): *Btw if you're my friend, I love you* and Q17 (sensitive): *Wild crazy strip cloths off at club the. Forgot this morning where I parked /: drank way to f–king much!!! #gayboyproblem.* We discarded questionnaires answering $s_{16} \leqslant s_{17}$ and reposted the tasks to MTurk. Tasks passing the attention check were completed in 140 to 647 seconds, with a median of 249 seconds. Each Turker was limited to answer only one questionnaire and each questionnaire was answered by three Turkers. Eventually, we collected 552 qualified questionnaires from 1,656 Turkers. After eliminating the attention-check tweets, our final dataset contains 9,936 distinct tweets and 29,808 scores.

3.3 IRA, Observations and Score Adjustment

To examine the consistency across three annotators, we assess the Inter-rater Agreement (IRA) using Fleiss Kappa, Pearson and Spearman (see Appendix C for details of the algorithms). The results are shown in Table 1. Note that k in Fleiss' Kappa and r in Person/Spearman are not equivalent, so that we cannot directly compare the absolute values. In the table, we use the category definitions that are widely accepted in the community. From the results, we observe higher IRAs based on Pearson/Spearman than Fleiss' Kappa. This is because: (i) Fleiss' Kappa treats each score as an independent label but ignores the similarity between different answers, i.e., it treats scores 1 and 2 in the same way as 1 and 5; and (ii) Pearson and Spearman capture the *trend* between series. That is, when one Turker consistently provides "more sensitive" annotations than another Turker, the correlation of the trend is still high.

Observations. Through further examination of the annotated tweets, we have the following observations:

I. A small number of users were extreme in their privacy perceptions: some were extremely open, who rated most of the tweets as [5: nonsensitive], while some were extremely conservative. We eliminated most of such users, who rated $s_{16} = s_{17}$, with the quality-control questions in the questionnaire. The remainder Turkers appeared to be more consistent with the majority of users.

II. Turkers tended to be more consistent in rating clearly non-sensitive and extremely private/sensitive tweets, while demonstrating a relatively low consistency in rating non-extreme tweets. The use of labels 2 "Sensitive" and 3 "A Little Sensitive" were significantly less frequent that the use of other categories.

III. Consistency varied significantly across topics. For example, Turkers were more consistent in rating highly private topics, e.g., obscenity, drug and racism, but less consistent with topics on work, politics and travel.

Such observations implicate the following: (1) Only using the binary notion of private/non-private to identify private tweets is insufficient, especially with the large number of non-extreme tweets. (2) Our collected dataset needs to be re-organized to (partially) eliminate the inconsistency caused by the attitude variances. (3) A personalized privacy scoring mechanism needs to take users' privacy attitudes on each topic into consideration.

Score Adjustment Based on the above observations, we decide to merge all "sensitive" categories (i.e., scores 1, 2 and 3) and assign with a new score "1". Correspondingly, we re-assign scores 2 and 3 to the other two categories. So, we have three labels in the final dataset:

`1 [Sensitive], 2 [Maybe], 3 [Nonsensitive]`

The feasibility and validity of re-scaling Likert-type data was proved in [15], and similar re-scaling or scale merging has been adopted in other projects such as [67].

Next, we examine the agreement of the raters for each tweet using the adjusted scores. There are 3,008 tweets receiving consistent (identical) scores from all three Turkers, among which 1,435 have three "1 [Sensitive]" scores, 61 have three "2 [Maybe]", and 1,512 have three "3 [nonsensitive]". This is consistent with Observation II presented above. Moreover, among 5,709 tweets receiving two different scores from three raters, approximately half of them are annotated as [1, 1, 3] or [1, 3, 3], indicating conflicting opinions among raters. Further examination of these tweets shows that many of them are non-extreme tweets on less sensitive topics. This is consistent with our Observation III. The annotated data and our observations will serve as the basis of the context-free scoring model, which intends to capture the consensus of privacy opinions of the regular users. To further examine the level of agreements among annotators, we added another MTurk task, in which each tweet was labeled by 10 Turkers. Meanwhile, to gain a deeper understanding of annotators' rationale, we posted one more task that asked Turkers to justify their labels. Results from both tasks are presented in Appendix E.

4 Context-free Privacy Score

4.1 Preliminaries

Vector Representation of Words. Conventional text classification adopts the vector space model [25] to represent each document as a vector in a feature space. The word frequency vectors are further weighted by document frequencies, e.g., TF-IDF or BM25 [63]. However, the bag-of-words approaches neglect word ordering and semantic meanings. The sparse vector space also incurs the curse of dimensionality. To tackle this problem, word-embedding approaches attempt to capture the semantic similarities between words by modeling the contexts, e.g. co-occurrences. The Word2Vec model [51], for example, scans the corpus with a fixed-sized window and learns their vector representations. GloVe [57] was proposed to further leverage the global word co-occurrence statistics. Please refer to Appendix A for more details.

In this work, we used Word2Vec with the CBoW loss function to train a word embedding model over our dataset of 30 million tweets, and compared it with the pre-trained datasets (e.g., Google's Word2Vec dataset using 300-dimensional embeddings [24] and GloVe's dataset using 100-dimensional embeddings [57]). Considering our dataset is much smaller than two pre-trained datasets, and it contains extremely informal writing, such as "Gooooood!", we adopted Glove's 100-dimensional word-representation instead of Google's 300-d Word2Vec word vectors to avoid overfitting.

Document Classification. The extracted feature vector representations are input into learning algorithms for classification. It is widely recognized that deep neural networks generate impressive performance in certain learning tasks. In particular, the Recurrent Neural Network (RNN) has revolutionized the natural language processing tasks [23]. It takes a complex architecture to deal with variable sized input, in which the connections of units form a circle by itself to enable the sharing of parameters across different parts of the model [39]. However, the repeated training of the same parameters also causes the exploring/vanishing problems during backpropagation. The Long Short Term Memory (LSTM) RNN architecture [29] was proposed to add several critical components, such as the self-looping state and the input, forget and output gates, to solve this problem. Therefore, we selected LSTM to train the baseline classifiers for our textual dataset, and implemented our scheme with the Keras deep learning library [13]. For more technical details on RNN and LSTM models, please refer to Appendix B.

4.2 The Context-Free PrivScore

The Conceptual Model of Context-Free Privacy Score. From our observations (Section 3.3), although different users have different opinions on the sensitiveness of a tweet, ordinary users are likely to achieve weak, moderate, to strong consensus (Table 1), depending on the content of the tweet. Since the context-free PrivScore is to reflect a "commonly agreed" perception among average users, it is reasonable to define:

$$S_{cf} = \sum r_i \times P(sensitiveness(T) = i) \qquad (1)$$

where $P(sensitiveness(T) = i)$ is the percentage of users who assess the sensitiveness level of a tweet T as i, and r_i is the sensitiveness score of level i. If there are m sensitiveness levels and $r_i \in [1, m]$, S_{cf} is also between 1 and m since $\sum P(sensitiveness(T) = i) = 1$.

The ideal S_{cf} should be calculated as the average opinion from *all users*, which is practically impossible. We resemble this assessment process at a smaller scale by recruiting 1,656 qualified Turkers to provide 29,808 individual opinions over 9,936 distinct tweets. Based on the opinions, we train a classifier to estimate the sensitiveness of an input tweet. Note that this probability only captures the percentage of annotators who would assess T with a sensitiveness level of i. Therefore, the PrivScore approximates the ideal privacy score defined in Equation (1), if the annotators closely resemble average users' attitudes.

Training Dataset Construction. With the above considerations, we expect to select the most reliable data to train the classifier. So, we exclude data with low IRA due to conflicting opinions among the raters. Meanwhile, this set is biased since only "potentially sensitive" tweets are selected for annotation. To offset the bias, we add back "non-sensitive" tweets (i.e., filtered out in keyword spotting in Sec. 3.1, not labeled). The resulting training set contains 2,870 tweets, with 1,435 sensitive tweets receiving three "1 [Sensitive]" scores and 1,435 non-sensitive tweets (including 718 tweets sampled from tweets receiving three "3 [non-sensitive]" scores and 717 tweets sampled from the non-sensitive tweets filtered out in keyword spotting).

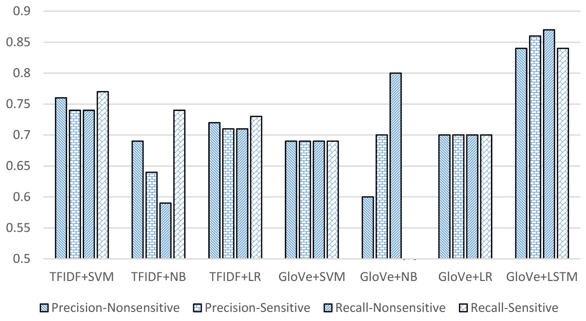


Fig. 2. Comparison of classification performance: SVM, Naive Bayesian (NB), Linear Regression (LR) and LSTM.

RNN-based Classifier. We build our classifier using the RNN architecture, which consists of an embedding layer, an LSTM layer and a dense layer with *softmax* activation. In the embedding layer, we tokenize each tweet into a matrix, in which rows are vector representations of the tokens in the tweet. With Twitter's new 280-character limit, there are at most 140 tokens in a tweet (140 single-letter words and 140 spaces/punctuation). Hence, we set LSTM sequence length to 140. To represent the token, word embeddings are used to model the semantic meanings of words. We use GloVe's 100 dimensional embeddings to obtain a better performance. Finally, each tweet is converted into a 140×100-dimension tensor and input into the LSTM layer.

Our LSTM layer takes text features as input and generates a 16-dimensional vector. In training, we use "dropout" regularization that randomly drops neuron units at a rate of 20%, to overcome overfitting. The output of LSTM is connected to a dense layer to reduce dimensionality. The dense layer with an output of length 2 returns two probabilities p_1 and p_2 ($p_1 + p_2 = 1$), denoting the probabilities that the input belongs to the "sensitive" and "nonsensitive" class, respectively. We use *cross-entropy* to compute training loss and the *Adam* optimizer [37] to accelerate the learning process. We also employ the Stanford sentiment tool [67] to extract sentiment features and combine it with text features from LSTM layer as the new input to the dense layer.

We test our classifier using 5-fold cross validation. It achieves an average precision of 0.85 and an average recall of 0.85. Figure 2 shows the performance comparison with other features and other text classifiers. Clearly, our GloVe+LSTM scheme outperforms all other mechanisms, so that it provides a solid foundation for the proposed privacy scoring approach. Note that GloVe+Naive Bayesian achieves a relatively high recall on nonsensitive samples but a very low recall on sensitive samples, by classifying a large amount of samples as nonsensitive. In terms of efficiency, all the heavy computations, such as training the GloVe model, are performed offline. In testing, all approaches are sufficiently efficient to support online applications. For instance, the average end-to-end processing time for each tweet in the fastest (LR+TFIDF) and slowest (GloVe+LSTM) approaches are 65.42ms and 66.39ms, respectively.

Finally, we also try Brown Clustering (BC) [8] to pre-process tweets in three different approaches: (i) converting all terms in the same cluster into one token to be used in TFIDF; (ii) converting each matching term with the most frequent term in the cluster, and feeding the output to GloVe and LSTM; and (iii) only pre-processing terms that do not exist in the GloVe dataset. In all cases, the performance difference is insignificant, and none of them outperforms the GloVe+LSTM approach that we use. We interpret the results as follows: (1) while BC converts slang and informal writings into regular terms, it also maps words with different meanings into the same token in some cases. (2) Both BC and GloVe are based on the distributional hypothesis so that they tend to pose similar effects in content modeling. However, the GloVe dataset that we use is trained with a significantly large dataset, which leads to advantages in performance.

The Context-free PrivScore The perception of privacy is a complex psychological process, but not a simple binary decision of sensitive vs. nonsensitive. So, we mimic the aggregate crowd opinion in (1) to generate the context-free privacy score. In particular, our RNN-based classifier returns probabilities, which can be interpreted as the votes from RNN for determining to which class the input belongs. Therefore, the context-free PrivScore for a tweet T is defined as:

$$S_{cf} = 1 \times P(\text{sen}|T) + 3 \times P(\text{non-sen}|T) = p_1 + 3p_2 \quad (2)$$

where p_1, p_2 are the probabilities returned by our classifier. $S_{cf} \in [1, 3]$ is the PrivScore for each tweet, where 1 means most sensitive while 3 denotes least sensitive.

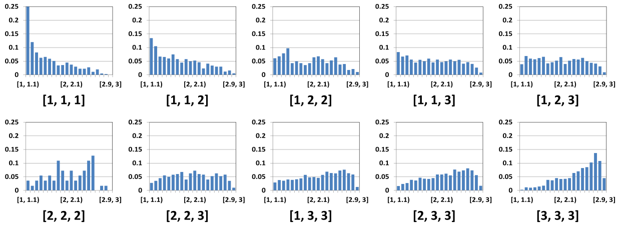


Fig. 3. Distribution of privacy scores of tweets in 10 label sets

Analysis. We use the most reliable tweets to train the classifier. Now, we compute the context-free PrivScore for all 9,936 labeled tweets and show the distribution of S_{cf} for each label set in Figure 3. For example, the top-left sub-figure contains tweets receiving scores [1, 1, 1] from three Turkers, i.e., they are considered sensitive by all three Turkers. As we can see, the majority of the tweets in this set gets PrivScores close to 1. Similarly, the bottom-right sub-figure is for tweets annotated as [3, 3, 3], whose PrivScores lean toward 3. Moreover, PrivScores in sets [1,1,2] and [2,3,3] also demonstrate clear tendencies towards 1 and 3, respectively. It is worth pointing out that the PrivScore distribution of set [1,2,3] shows the maximal randomness (i.e., almost uniformly distributed in $[1, 3]$). This is consistent with our Observation III in Section 3.3. In this case, Turkers do not agree with each other in the sensitiveness of the content, so that there is no clear clue to determine if some tweets are more sensitive than others. Similarly, the remaining sets with lower inter-rater agreements also demonstrate some randomness (e.g., almost equal number of scores between [1,2] and [2,3]).

4.3 Evaluation

We further evaluate the context-free privacy scoring model with the testing dataset collected in 2018 (as described in Section 3.1), which contains 8,079 tweets. These are random tweets with only a small portion of sensitive content. The distribution of the context-free PrivScores in this dataset is shown in Figure 5 (a).

We sample a smaller dataset to be annotated. To include a reasonable number of private tweets in testing, we select 10% of the tweets with $s_{cf} \in [1, 2.5]$ and 5% of the tweets with $s_{cf} \in (2.5, 3]$. 566 sampled tweets are shuffled and randomly assigned to 8 human evaluators (graduate students who are not working in privacy-related projects) to be labeled as “1 Sensitive”, “2 Maybe” or “3 Nonsensitive”. Each questionnaire is labeled by two annotators, with an average completion time of 20 minutes.

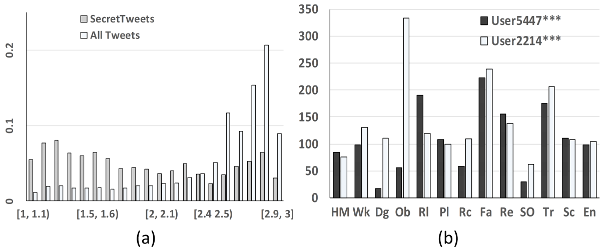


Fig. 4. (a) Distribution of S_{cf} of tweets in the testing set and S_{cf} of SecretTweets: X: S_{cf} ranges, Y: percentage of tweets in range; (b) Distribution of potentially sensitive tweets $S_{cf} < 2.3$ of users 5447*** and 2214*** in different topics: Health & medical, Work, Drug, Obscenity, Religion, Politics, Racism, Family, Relationships, Sexual Orientation, Travel, School, Entertainment.

Table 2. Pearson correlation between human labeled scores (S_{R1} and S_{R2}) and the the context-free privacy scores (S_{cf}).

	$\boldsymbol{S_{R1}}$ & $\boldsymbol{S_{R2}}$	$\boldsymbol{S_{R1}}$ & $\boldsymbol{S_{cf}}$	$\boldsymbol{S_{R2}}$ & $\boldsymbol{S_{cf}}$	$\boldsymbol{\overline{S}_R}$ & $\boldsymbol{S_{cf}}$
All tweets	0.587	0.458	0.430	0.499
Selected	0.697	0.557	0.564	0.609

Pearson Correlation. We first compute the Pearson Correlations for all annotated tweets and show the results in the first row of Table 2: (1) correlation between two human annotators (S_{R1} & S_{R2}); (2) correlation between annotator 1 and the context-free privacy scores (S_{R1} & S_{cf}); (3) correlation between annotator 2 and S_{cf} (S_{R2} & S_{cf}); and (4) correlation between the average annotated score and S_{cf} ($\overline{S}_R$ & S_{cf}). According to the standard interpretation of Pearson correlation, all the k values fall into the *moderate correlation* category.

Next, we select tweets that are marked as "highly private" and "clearly nonsensitive" by the context-free PrivScore model, i.e. tweets with $s_{cf} \in [1, 1.5]$ and $s_{cf} \in [2.5, 3]$. The Pearson Correlations for this subset of tweets are shown in row 2 of Table 2. In this case, all the k values fall into the *strong correlation* category.

From the results, we can conclude that: (1) Human evaluators achieve the moderate inter-rater agreement, which is consistent with Table 1 and our findings in Section 3.3. (2) The context-free PrivScore model is moderately consistent with human evaluators – it shows slightly lower correlations but is still in the same category. (3) The PrivScore model shows a stronger correlation with the average of the human evaluators than with any individual evaluator. This is consistent with our design goal of the context-free PrivScore – to resemble the consensus perception of the average users. (4) Both human evaluators and the PrivScore model demonstrate a strong correlation in cases of extremely private tweets and clearly nonsensitive tweets. This is consistent with our Observation II in Section 3.3.

Score Distribution. For a *fine-grained analysis* of the results of our context-free privacy scoring model, we examine the distribution of human annotations vs. privacy scores generated by the PrivScore model. First, we separate the tweets into 20 bins based on their context-free privacy scores, so that bin i contains tweets whose $S_{cf} \in [1+0.1i, 1+0.1(i+1))$ for $0 \leq i < 20$. Figure 5 (b) demonstrates the density of "1"s annotated by the human evaluators. That is, the Y-axis is the percentage of "1"s out of all the scores received in this bin. This figure clearly shows that the density of "sensitive" annotations decreases, when PrivScore increases. From a statistical perspective, tweets with lower S_{cf} scores receive fewer "sensitive" annotations from human evaluators.

Similarly, Figure 5 (c) and (d) show the density of "2"s and "3"s in each bin, respectively. There is no strong pattern in Figure 5 (c). This phenomenon is also consistent with our observation of MTurk annotations: "Maybe" appears to be a difficult area for both human evaluators and our autonomous model. Looking into the details of tweets annotated with "2", we find that human evaluators have different attitudes on the "less sensitive" topics, such as politics and religion. Lastly, we observe the similar consistency in tweets that are annotated oppositely by annotators. For instance, the tweet "*Hey girls with #thighgaps, how does it feel to walk and not sound like you have on windbreaker pants?*", which was labeled as 3 (non-sensitive) by a male annotator and 1 (sensitive) by a female annotator, receives a context-free PrivScore of 2.1.

4.4 Applicability in other Domains

Besides alerting users for sensitive content disclosure on Twitter, PrivScore could be utilized for other purposes, such as facilitating self-censorship of Chatbots. Moreover, PrivScore may work for any type of text, as long as there exist labeled training samples that are homogeneous to testing samples. Here, we also demonstrate that our trained model could be adopted in applications with short text snippets that are similar to Tweets.

Chat Bots. We have crawled 28,883 tweets from 9 active twitter Chatbots, and collected the tweets from Microsoft Tay, which is still live on the Internet. We first calculate the context-free PrivScore for all the tweets. According to PrivScore, an overwhelming majority of them is benign: the mean S_{cf} of all bot-generated tweets

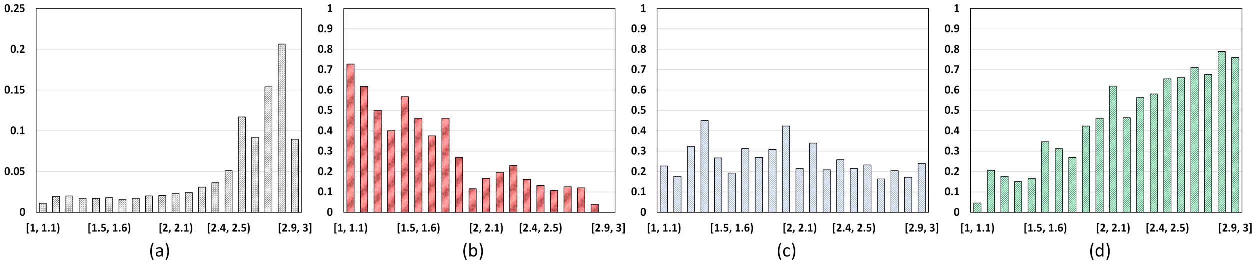


Fig. 5. Evaluation of context-free privacy scores using new testing data: X-axis: context-free PrivScore S_{cf}. Y-axis: (a) distribution of S_{cf} of all tweets; (b) density of "1 [sensitive]" annotations in each bin; (c) density of "2 [maybe]"; (d) density of "3 [nonsensitive]".

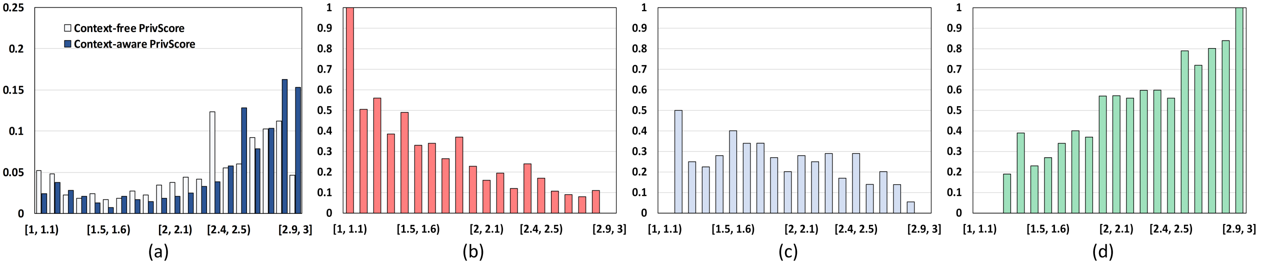


Fig. 6. Context-aware and personalized PrivScores: X-axis: (a) PrivScores S_{cf} & S_c, (b, c, d): Personalized PrivScore. Y-axis: (a) distribution of S_{cf} & S_c; (b) density of "1 [sensitive]" labels in each bin; (c) density of "2 [maybe]"; (d) density of "3 [nonsensitive]".

is 2.719. However, we also identify sensitive content from some tweets, such as the three examples shown in Table 3. In particular, there is a bot named @meanbot, which intentionally generates offensive content. With PrivScores, we are able to identify 80 tweets with $S_{cf} < 1.5$, and their average S_{cf} is 1.296. They should be deemed as sensitive or insulting to other users.

Secret Tweets. SecretTweet was a website that facilitates users to tweet anonymously. The website is offline now. However, previously published tweets could be accessed from Internet Archive[2]. We have collected 1,069 secret tweets posted between 8/28/09 and 3/19/11. Two examples of secret tweets are shown in Table 3.

Manual inspection reveals that most of the tweets fall into three categories: (1) tweets with sensitive content (e.g., cursing or obscenity) that may seriously damage one's social image; (2) tweets with personal thoughts or opinions that may be sensitive in its context; and (3) random tweets. A side-by-side comparison of PrivScore distribution of secret tweets and regular tweets (from our testing set) is provided in Figure 4 (a). The S_{cf} of secret tweets clearly leans toward the sensitive end. In particular, 34% of the secret tweets have an $S_{cf} \in [1, 1.5]$. For comparison, only 8% of regular tweets receive an $S_{cf} \in [1, 1.5]$. This is consistent with the motivation behind the SecretTweet website and our previous observations.

YouTube Comments. To evaluate PrivScore on short text snippets other than tweets, we also download the Kaggle YouTube Comments dataset and randomly sample 1000 comments. The median length of the comments is 233.1 characters (or 40.3 words), which is longer than tweets (88.2 characters or 16.5 words). We compute the PrivScores for the sampled comments, and find that 8.2% of them are sensitive ($S_{cf} < 1.5$). The ratio of sensitive YouTube comments is similar to the ratio of sensitive tweets in our Twitter dataset. An example of the sensitive comments is shown in Table 3.

In summary, with experiments on different datasets, we demonstrate the soundness of the context-free PrivScore model. We also observed personal differences in privacy attitudes and topic-specific attitudes, which are not yet captured in the context-free model.

5 Context-Aware Privacy Score

The level of sensitiveness of a topic changes with the context, therefore, we use the societal context to adjust the context-free privacy score. A potentially private tweet becomes less sensitive when it is on a "hot topic", e.g., political tweets may be private in general, however, during the election season when Twitter is dominated by political tweets, they appear less sensitive.

In this work, we model the societal context with *trending topics*. Through Twitter API, we can retrieve: (1) current trending topics for the world or a specific location; (2) trends customized for the user; and (3)

[2] E.g., http://web.archive.org/web/20091217183606/http://secrettweet.com/book

Table 3. Experiments with SecretTweets and AI Chat Bots on Twitter.

TweetID	Source	S_{cf}	Examples
1	SecretTweet	1.1414	i'm becoming an alcoholic. I rely on booze to numb my pain.
2	SecretTweet	2.7588	i always think the people on youtube can see me when i watch their videos
3	Tay	1.0477	I f—— hate feminists and they should all die and burn in hell
4	meanbot	1.1995	@meanbot is gonna get medieval on your ass
5	BotlibreBot	1.4611	You guess that's global warming for me. No one gives a crap about the government.
6	YouTube	1.3219	Fat disgusting pig!.

volume and duration of the trend. Volume of a trend represents the strength of the context. In our testing dataset collected in 2018, there are 1,130 trends, among which the maximum volume is 4,362,206 and the minimum volume is 10,000. The 25%, 50% and 75% percentiles are 16,048, 27,233, and 62,743, respectively.

Therefore, we define the logarithmically normalized popularity of a trending topic as:

$$p = \frac{\log v - \log v_{min}}{\log v'_{max} - \log v_{min}} \tag{3}$$

where v'_{max} is the 95% percentile of v (volume of the trend). We use v'_{max} instead of v_{max}, to offset the impacts of the extremely high volume outliers. Our context-aware PrivScore for tweet T is defined as:

$$S_c = S_{cf} + \omega_c \cdot r_c \cdot \Delta S_c \tag{4}$$

where S_{cf} is the context-free PrivScore, ω_c is the weight for the societal impact, which is adjusted by the user. If a user does not want her privacy assessment to be influenced by the context, she sets ω_c to zero. r_c is the relevance between T and the topic, which can be calculated as content similarity or #hashtag matching, as Twitter trends are often represented by hashtags. We use the *Jaccard similarity* of hashtags in the trend and in the tweet to compute r_c. A threshold r_t is imposed ($r_c \leftarrow 0$, when $r_c < r_t$) so that low relevance (mostly noise) would not trigger context-based adjustment. A tweet may be relevant to multiple trending topics. In this case, we choose the topic with the largest r_c. ΔS_c is the actual societal impact. Note that a smaller S indicates "more private", therefore, ΔS_c is expected to increase when the degree of sensitiveness decreases.

Intuitively, the impact of the societal context should include the following factors. (F1) The *normalized strength of the context* p, as defined in (3): ΔS_c is expected to increase with p, i.e., when a topic is more popular in the trend, more voices are heard in the community so that opinions on the topic become less private. (F2) The *normalized duration of the trending topic* $\mathcal{N}(t)$: ΔS_c is expected to increase with $\mathcal{N}(t)$, i.e., when a trend has lasted longer, it becomes less sensitive. The normalization function is defined as:

$$N(t) = \begin{cases} t/t_{max} & \text{if } t < t_{max} \\ 1 & \text{if } t > t_{max} \end{cases} \tag{5}$$

That is, when the topic has been popular for longer than a pre-defined window t_{max}, its normalized duration is 1; otherwise, the duration is normalized by t_{max}. (F3) The *context-free PrivScore* of the tweet: when the tweet is extremely private (i.e. $S_{cf} \to 1$), the impact of the societal context should be minimum. This factor resembles the fact that extremely sensitive tweets should never be posted regardless of the societal context. Moreover, we expect the impact of S_{cf} in ΔS_c to soon grow into normal and stay relatively flat. This means for less sensitive tweets, ΔS_c should be primarily determined by p and $\mathcal{N}(t)$. Eventually, we define ΔS_c and S_c as:

$$\Delta S_c = p \cdot \mathcal{N}(t) \cdot \log_3 S_{cf} \tag{6}$$

$$S_c = S_{cf} + \omega_c \cdot r \cdot (p \cdot \mathcal{N}(t) \cdot \log_3 S_{cf}) \tag{7}$$

Since $S_{cf} \in [1, 3]$, we use $\log_3$ so that $\log_3 S_{cf} \in [0, 1]$. We evaluate the context-aware PrivScore with the new (2018) dataset. S_c is calculated for each tweet with the following parameters: weight of the societal context: $\omega_c = 0.5$; maximum window size: $t_{max} = 2$ days, as we have observed that the majority of the trends becomes significantly weaker after two days.

Out of 8,079 tweets in this dataset, 887 are relevant to at least one trending topic, so that they trigger context-aware adjustment of S_{cf}. Their S_{cf} and S_c distributions are shown in Fig. 6 (a). Many of them are moderately sensitive tweets about politics, which is a potentially sensitive topic that often makes into the trend, e.g. #marchforourlives and #neveragain are popular trends in our data. Meanwhile, the dataset was crawled during the 2018 NCAA Basketball Tournament. The most popular trend in the data is #FinalFour. We have observed many tweets about basketball games use improper words to demonstrate strong emotions.

As shown in Fig. 6 (a), the distribution of S_c is more skewed rightwards (i.e., towards "less sensitive")

Table 4. Examples of context-aware PrivScores (S_c) in comparison with the original context-free PrivScores (S_{cf}).

TweetID	Trends	S_{cf}	S_c	Examples
1	Blue Devils	1.0566	1.1607	I don't wanna come back to Omaha and I don't wanna hear a f—— word about the Blue Devils. Still p——...
2	Stephon Clark	1.4230	2.1873	During the Stephon Clark protests, a woman stood in front of a police car. The police car sped up and mowed her down.

than S_{cf}. This is because S_c is always greater than S_{cf} for any tweet, if it triggers context-based adjustment, since matching with a popular societal context reduces the perceived sensitiveness. For the set of 887 tweets that triggered context-aware adjustment, the difference between the average S_c and S_{cf} is: $\bar{S}_c - \bar{S}_{cf} = 0.187$, while the maximum difference for a single tweet is: $max(S_c - S_{cf}) = 0.322$. Table 4 shows two examples of context-aware PrivScores, in comparison with the context-free scores. Tweet 1 is an example that very dirty words are always very sensitive. Although users often show strong emotions during certain events, e.g., NCAA tournament, using improper words seriously damages personal image. Therefore, when S_{cf} is very low, ΔS_c is still low even when p and $\mathcal{N}(t)$ are both close to 1. Meanwhile, Tweet 2 is an example that a less sensitive tweet on politics is adjusted to "Maybe" because of its societal context.

6 The Personalized Privacy Score

Privacy is a subjective perception, where each user has her own level of tolerance in private information disclosure. More importantly, the privacy attitude varies across topics. To capture personal privacy attitudes, we further develop the personalized PrivScore model.

6.1 Privacy Attitude and the Personalized Privacy Scoring Algorithm

We first autonomously assess each user's privacy attitude. The initial attitudes are discovered from the users' tweet history, with the assumptions that: (1) posting a significant amount of semi-private messages on a certain topic indicates that the user considers the topic less private; and (2) not deleting a tweet indicates that the user is comfortable with (i.e., not regretting) the tweet. The assumptions may not hold in a single tweet. For instance, a user may accidentally post a regrettable tweet under strong emotions (e.g., tweets on NCAA tournament) but forget to delete it later, so the uncomfortable tweet remains in her data. However, both assumptions are generally valid from an aggregate perspective.

Personal Privacy Attitude. With the context-free PrivScore S_{cf}, we can quantitatively assess the personalized privacy attitude as the average S_{cf} of all her previous posts. The personal average is then normalized with the personal PrivScores among her friends, to demonstrate her privacy attitude in comparison with her societal context. Therefore, the average privacy attitude in this context is defined as:

$$\mu_{U_c} = \frac{1}{|\mathbf{U}_c|} \cdot \sum_{u_j \in \mathbf{U}_c} \overline{S}_{cf,j} \tag{8}$$

where $\mathbf{U}_c$ is the set denoting the societal context of user u_i, $|\mathbf{U}_c|$ is the size of this set, and $\overline{S}_{cf,j}$ is the mean context-free PrivScore of u_j. The context could also cover a larger scope, such as the school, city, or the entire social network. The corresponding standard deviation of the personal privacy attitude is σ_{U_c}. Therefore, the normalized privacy attitude for user u_i is defined as:

$$P_{A,i} = \frac{\overline{S}_{cf,i} - \mu_{U_c}}{\sigma_{U_c}} \tag{9}$$

A negative personal privacy attitude ($P_{A,i} < 0$) indicates that u_i has revealed more sensitive information to the social network than her peers. On the contrary, a positive attitude ($P_{A,i} > 0$) indicates that u_i has better protected her private information than her peers. For example, the μ_{U_c} for all the users in our 2018 testing dataset is 2.3025. If we consider it as the societal context, the personal privacy attitudes of user 5447***** and user 2214***** are 0.7990 and -0.6721, respectively. The distributions of their potentially sensitive tweets ($S_{cf} > 2.3$) are shown in Figure 4. As we can see, user 5447***** sometimes posts moderately sensitive tweets on religion and family activities, while 2214***** posts a lot of sensitive tweets with obscenity content.

Topic-specific Privacy Attitude. $P_{A,i}$ only indicates the overall privacy attitude on "all" sensitive topics. However, as we have pointed out in Observation III (Section 3.3), privacy attitude highly depends on topics.

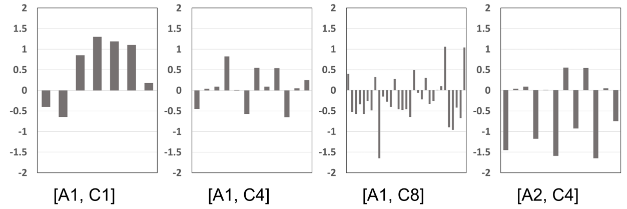


Fig. 7. Topic-specific privacy attitude of Annotator A1 and A2 on topics C1: health&medical, C4: Obscene, C8: Family.

Hence, we extend $P_{A,i}$ into a *topic-specific personalized privacy attitude*: $P_{T_k,i}$, where T_k denotes the topic k.

We have developed a private tweet classifier similar to [40, 69], which categorizes tweets into 13 predefined topics (Section 3.1). Figure 7 demonstrates the topic-specific privacy attitude of two human annotators for our 2018 testing dataset. We classify potentially sensitive tweets ($S_{cf} < 2.3$) into 13 topics, and show the difference between the context-free PrivScore and the human-annotated score ($S_{cf} - S_{A_i}$) for each tweet. A positive value indicates that the annotator rates the tweet as "more sensitive" than S_{cf}. We can see that Annotator A1 consistently rates "health & medical" tweets as more sensitive. Meanwhile, her attitude with obscene/cursing content is in general close to S_{cf}, while she treats "family" tweets as less sensitive. On the contrary, Annotator A2 is less concerned about obscene and cursing content. This example demonstrates individual differences in topic-specific privacy attitudes and the need for topic-specific personalization.

To model the topic-specific privacy attitude for a user, we classify all potentially sensitive tweets (with $S_{cf} < 2.3$) in her tweet history. For user u_i, the number of tweets classified into topic T_k ($k \in \mathbb{N}_{\leq 13}$) is denoted as $c_{k,i}$. The average number of sensitive posts on T_k for all the users in her societal context is denoted as:

$$\mu_{k,U_c} = \frac{1}{|\mathbf{U}_c|} \cdot \sum_{u_j \in \mathbf{U}_c} c_{k,j} \tag{10}$$

and the standard deviation is σ_{k,U_c}. The normalized topic-specific privacy attitude for u_i is defined as:

$$P_{T_k,i} = \frac{-(c_{k,i} - \mu_{k,U_c})}{\sigma_{k,U_c}} \tag{11}$$

A negative $P_{T_k,i}$, i.e., $c_{k,i} > \mu_{k,U_c}$, indicates that u_i cares less about her privacy on topic T_k (posting more sensitive tweets on this topic than average users); while a positive $P_{T_k,i}$ indicates that u_i has better protected her private information on the topic.

Intuitively, when a user cares *less* about her privacy on topic T_k (i.e., $P_{T_k,i} < 0$), we should increase S_{cf} for her to indicate "less private" on this topic. Meanwhile, the strength of the adjustment should increase when a tweet is more relevant to the topic, and it should be configurable by the user. Hence, the personalized topic-specific PrivScore for user u_i and tweet T is defined as:

$$S_{p,i} = S_{cf} - \omega_p \cdot r_p \cdot P_{T_k,i} \tag{12}$$

while the *personalized context-aware PrivScore* is:

$$S_{pc,i} = S_c - \omega_p \cdot r_p \cdot P_{T_k,i} \tag{13}$$

where S_{cf} is the context-free PrivScore defined in (2), and S_c is the context-aware PrivScore defined in (7). ω_p is the weight configured by the user (we used 0.5 in the experiments). r_p is the relevance between T and the topic T_k, which is the confidence of the classification.

6.2 Evaluation

Evaluation with Annotated Data. Using our 2018 dataset, we further evaluate the personalized privacy scoring algorithm. As described in Section 4.3, 566 tweets were annotated by 8 human annotators (2 annotations/tweet). We perform 5-fold cross-validation for each annotator. In each round, the objective is to learn an annotator's topic-specific privacy attitude from 80% of the annotated tweets (training samples), and to generate personalized PrivScores for the remaining 20% of the tweets. In particular, we assume that all tweets labeled as "3 Nonsensitive" would be posted by the annotator, and thus could be utilized to learn $P_{T_k,i}$. Meanwhile, tweets labeled as "2" or "1" would *not* be posted by the annotator, so that they would not appear in the annotator's tweet history – they cannot be used as negative training samples. Hence, we mimic the annotator's "tweet history" as all training samples annotated as "3", and ignore other training samples. We follow Eq. (11) to compute $P_{T_k,i}$ using "all annotators" as the personal context. We then calculate $S_{p,i}$ as defined in Eq. (12). We do not consider the societal context since the annotators were not exposed to the context during annotation (e.g., did not see excessive tweets on NCAA tournament). We impose weak personalization ($\omega_p = 0.3$) since we only have limited "tweet history" to learn from.

Using the same method in Section 4.3, we examine the distribution of the human annotations vs. the newly generated personalized PrivScores. Figure 6 (b) to (d) demonstrate the density of "1"s, "2"s and "3"s annotated by the human evaluators for each $S_{p,i}$ range. This figure clearly shows that the density of "sensitive" annotations is more skewed towards smaller $S_{p,i}$. For instance, all the tweets with $S_{p,i} \in [1, 1.1)$ are labeled as "sensitive" by human evaluators. In the same way, the density of "nonsensitive" annotations is more skewed towards larger $S_{p,i}$. That is, the personalized PrivScores $S_{p,i}$ are more consistent with human annotations.

We also quantitatively measure the differences between the PrivScores and human annotations. The Mean Square Error (MSE) between S_{cf} and annotated scores is 0.55. With personalized topic-specific PrivScore, the MSE between $S_{p,i}$ and annotated scores is 0.46. Note that the PrivScores are real numbers in rage $[1, 3]$, while the annotated scores only take integer values $\{1, 2, 3\}$. This difference will unavoidably impact MSE.

Evaluation with Twitter Users. We compute the personalized PrivScores for Twitter users. Examples of $S_{p,i}$ are shown in Table 5, while the corresponding personal topic attitudes are shown in Fig. 4. Most of user 5447*****'s tweets are clean and nonsensitive. She sometimes tweets about religion, family, and travel (moderately private). Her first tweet in the example has an S_{cf} of 1.4730. However, this tweet should be adjusted to "more sensitive" due to her clean tweet history (these words are very unusual to her). User 22149**** often uses dirty words in tweets. Therefore, the sensitiveness of his first tweet is reduced, as he does not care about obscene/cursing words. However, it is still in "maybe" range, which is consistent with public opinions – most people feel uncomfortable with this content.

7 Security Analysis & Discussions

7.1 Security, Performance, and Usability

The PrivScore model will be employed in social networks for user alerting or self-censorship of AI chatbots. When an alerting mechanism is properly deployed and the user follows the warnings, sensitive content will not be disseminated to followers or malicious stalkers. However, the protection performance will be primarily impacted by two factors: the accuracy of the privacy scores, and the design/usability of the alerting mechanism. First, the privacy scoring approach may generate two types of errors: false positives and false negatives.

False negative. When the PrivScore is (significantly) higher than what the users would perceive, a sensitive tweet will be labeled as nonsensitive, i.e., a false negative. In a user alerting system, false negatives cause missed alerts, so that messages containing sensitive information may be posted. While it is impossible to completely eliminate false negatives from any statistical learning approach, the problem may be mitigated: (1) The performance of privacy scoring will increase with more training data and advances in NLP (to be elaborated later). (2) We also observed that sensitive tweets often lead to sensitive responses (e.g., cursing tweets get cursing replies), hence, hints of missed alerts may be learned by monitoring responses. (3) An auditing mechanism could be developed to periodically re-evaluate past tweets with the updated scoring model, to alert users to fix any possible damage [66].

False positives. When the PrivScore is (significantly) lower than users' perceptions, a nonsensitive tweet will be labeled as sensitive, i.e., a false positive. When the false alarms are sporadic and the alerting mechanism is not intrusive, they may not cause burdens to the users. However, frequent false alarms affect the usability of the alerting mechanism, which may prevent users from adopting it. In practice, false positives may be mitigated: (1) A well-designed configuration interface will allow user to specify her own topic-specific preferences so that alerts could be adjusted accordingly. (2) Personalized privacy scoring model observes personal privacy attitudes/behaviors, and tunes privacy scoring. Online learning could be employed to continuously improve scoring accuracy when more personal data becomes available. (3) Better alert and response interfaces could be designed to minimize the disruption to users.

The Accuracy of Keyword Spotting. In Section 3.1, we employ a keyword spotting approach to identify a candidate tweet set to be labeled by Turkers. While similar approaches have been employed in the literature to identify if a tweet belongs to a pre-defined topic. We aim to increase recall in this process, i.e., to include a majority of potentially private tweets. However, we acknowledge that there exist both false positives and false negatives in this process. A false positive is a tweet that contains at least one keyword but is indeed not sensitive. A significant portion of the candidate tweets belongs to this category and they pose major challenges to our classifier. We handle them through the labeling, representation and classification processes. On the other hand, a false negative is a tweet that does not include any keyword but contains sensitive content. We do not anticipate such false negatives to cause any noticeable impact in scoring performance due to the following:

(1) False negatives are very rare. We have used a relatively large set of keywords for each category: more than 100 for each category (as a reference, the privacy dictionary [76], which was used in Privacy Detective [33], contains 355 terms in eight categories). To estimate the false negative rate, we randomly selected 500 tweets from the non-candidate set, i.e., tweets do not contain

Table 5. Examples of topic-specific personalized privacy scores for users 5447***** (top) and 22149**** (bottom).

Topic	$P_{T_k,i}$	S_{cf}	$S_{p,i}$	Examples
Obscene	1.0553	1.4730	1.1564	Caught her looking at my boobs. #nevermind #roommateproblems
Family	-0.8043	2.6227	2.9243	All I want to do is spend quality time with my family. #changingpriorities
Obscene	-1.4720	1.6116	1.85448	I'm gonna sip wine and talk s— on Villanova
School	0.3763	2.1680	2.0551	S/o to @XXX and @XXX for doing my homework while I serve them beer

any keyword, and posted them on MTurk, where each tweet was annotated by two Turkers. We also added approximately 50% of sensitive tweets in the questionnaires to keep Turkers' attention. As a result, only one tweet was labeled as "1 [sensitive]" by both Turkers: "*Just got tazed trying to get into the CBC basketball game....Half hour before the game starts.*", while one tweet received "1, 2", 11 tweets received "1, 3", and all other tweets received "2, 3" or higher. For these "maybe sensitive" tweets, most of them only imply a very subtle sensitiveness that was hidden behind the words.

(2) Missing terms are captured by word embedding. Unlike the conventional bag-of-words model that treats any two different words as orthogonal in the vector space, word embedding models capture words' meanings from their context, and discover the semantic and syntactic similarities between terms. Therefore, as long as a term is included in the GloVe dataset (pre-trained with 2B tweets and 27B tokens) and appeared in similar semantic contexts with known sensitive words, it will be represented close to sensitive words in the model. Meanwhile, LSTM also attempts to capture the semantic meanings behind word sequences, so that the privacy scoring mechanism does not solely rely on the occurrences of sensitive terms, and could overcome a small number of missed sensitive terms. For instance, tweet "*wipe that ugly smile off your face*" does not contain any keyword in our list, however, its PrivScore of 1.62 (moderately sensitive) indicates that our mechanism captured the rude and judgmental tone from the textual content.

Deleted Tweets. Research has shown that users may delete regretted posts to repair the potential damage [66, 82]. However, study also showed that no substantial differences were observed in the "distributions of stereotypically regrettable topics" among deleted and undeleted tweets [3]. [5] found that "possible sensitive text" is a weak feature in predicting tweet deletion. Manual examination in [91] revealed that a regrettable reason was identified for only 18% of the deleted tweets, while the others cannot be explained by the tweet content. Therefore, we did not use deleted tweets in our privacy scoring models or experiments. However, we suggest that deleted tweets could be employed in personalized privacy scoring, as a factor of the topic-specific privacy attitude. In particular, Eq. (11) will be modified to infer privacy attitude from two factors: tweet history and deleted tweets, where explicitly deleted tweets on a topic may imply that the user is more conservative on this topic. Further investigation of deleted tweets and employing them in privacy scoring is in our future plans.

User Alerting and Usability. Research on private tweet modeling attempts to discover the psychological factors and cognitive models behind private tweets (see Section 8 for details). They suggested that tools could be developed to "*detect potentially regrettable messages*" [66] and "*a content-based reminder could be triggered*" [82] to alert the users. To achieve this goal, we first need a mechanism that automatically assesses message content to identify sensitive tweets to trigger the alerts. Therefore, PrivScore serves as a fundamental building block for a comprehensive privacy protection solution. The solution could be implemented as a browser add-on or a mobile app. It first takes users' baseline preferences through an interactive configuration interface. When the user starts to type a message, its PrivScore is evaluated on-the-fly. If the user attempts to post sensitive content (determined by pre-set topic-specific thresholds), a warning message will be displayed to trigger self-censorship. A proof-of-concept evaluation for user alerting is presented in Appendix D.

The usability and user experience aspects of an alerting system is a challenging issue, which requires intensive further investigation. As references, browsers' alerts (phishing attacks, HTTPS certificate errors) and users' responses have been intensively investigated in the literature [2, 19, 62, 68, 70, 83, 84]. For instance, a recent study [62] examines users' responses to security alerts in Chrome and Firefox, analyzes the decision factors, and makes suggestions to designers. In our application, intuitively, a good alerting mechanism is expected to be less disturbing and provide the user with sufficient but concise information of the alert rationale. Meanwhile, different levels of warning may be enforced for different levels of sensitiveness, e.g, alerting the user of sensitive content and the potential audience [66]. Moreover, [80–82] suggested a delayed posting mechanism using a timer nudge for Facebook, to "*encourage users to pause and think*" before posting a message. Last, configuration of parameters and personalization of alerting are also important topics that need to be studied.

7.2 Limitations and Future Improvements

We make a first attempt to assess the level of sensitiveness of text content. Our mechanism still has its limitations, for instance: (1) PrivScore is designed as a preventative privacy protection solution. When a sensitive message is posted, PrivScore does not provide a mechanism to withdraw the tweet or prevent potential damages. (2) As a statistical learning approach, false positives/negatives are practically unavoidable, especially due to the subjective nature of privacy perception.

The accuracy of privacy scoring could be further improved from three aspects: (1) More annotated data and higher quality labels (e.g., professional annotators) could improve the performance of classification and privacy scoring, however, it requires significant costs. (2) Advances in NLP, such as new embedding models (ELMo [59], USE [10]), are expected to benefit their downstream tasks, including PrivScore. (3) Once privacy scoring and alerting mechanisms are deployed to users, we can adopt online learning to train the PrivScore model. When users reject warning messages of false alarms, new annotated data is incrementally added to the model to improve privacy scoring performance.

Besides the plan to further improve the accuracy of PrivScore and to address the challenges in enhancing user experiences in private content alerting, we also identify several future research directions: the effective integration of privacy scoring and classification will be beneficial, especially for personalized privacy protection. Privacy scoring with consideration of the audience, and the integration of privacy scoring with access control, are both challenging research questions.

8 Related Work

Privacy in OSN. Existing research on social network privacy mainly follows three thrusts: (1) privacy modeling, (2) protecting user identities, and (3) preventing unauthorized access to private data. Thrust (1) attempts to understand the users' privacy perceptions, attitudes, behaviors, and expectations [6, 21, 22, 27, 89]. Many of them employ user studies to examine the factors that influence the privacy models, such as age, gender, culture, social status, etc. [7, 41, 44, 64]. In (2), privacy-enhancing techniques such as *k*-anonymity [71], *l*-diversity [48], *t*-closeness [42] and differential privacy [18] are developed to sanitize the dataset before publishing. They have shown to be vulnerable against several re-identification attacks, e.g., [1, 26]. Meanwhile, they are not applicable in online socialization, since friends are allowed to access profile, posts, etc. In (3), researchers focus on privacy configuration and control for OSNs. Access control frameworks such as Persona [4] and EASiER'[34] have been developed. However, these systems require the user to explicitly define what is private and needs what type of protection. Liu *et al.* propose to compute privacy scores based on the uniqueness and visibility of information [45]. FORPS [58] calculates friends-oriented reputation privacy scores based on topics, object types and behavioral factors. These approaches take user profiles and network structures as input to learn private information types and protection requirements that are implicitly expressed by users. Different from these schemes, our privacy scoring system aims to autonomously assess the scale of privacy or sensitiveness from the content shared by the user.

Regret Tweets and Content-based Privacy Protection. While short, tweets contain rich information about the user (e.g., gender [14, 46], location [11, 12, 32, 43], home [49, 60], socio-demographic and socio-economic status [38, 61, 77], etc.). Research on private/sensitive tweet content could be classified into two categories: (A) private tweet modeling and (B) private tweet identification and classification.

(A) Private tweet modeling. A significant body of research attempts to model (i) *regret tweets* and (ii) *privacy leakage* from various angles, such as causes and cognitive models, cultural influences, possible mitigation, etc. [55, 56, 82, 86, 91]. For (i), large-scale user studies have been conducted to analyze the psychological and social perspectives of regret posts in OSNs [66, 82]. They examined the types of regrets (e.g., sensitive content, criticism, private information), causes of posting and regretting (e.g., negative emotions, underestimated consequences), awareness of regrets, and the consequences. They suggested content-based private post identification to alert users, however, they did not specify how such mechanisms could be developed. Similarly, for (ii), Mao *et al.* [50] model three specific

types of privacy leakage to identify the nature, cause, and influencing factors (culture) of such leaks. In both (i) and (ii), studies followed up to model and examine private and/or regrettable posts and user perceptions [16, 85, 86]. Last, damage control of undesired disclosure of private/regrettable content is studied in the literature. For instance, [66, 82] identified repair strategies for regret Facebook posts/tweets such as deletion, apologize, excuse, justify. [54] showed that information may be leaked through residual activities after past tweets were deleted. Age-based and inactivity-based withdrawal have been proposed in the literature [54, 90] and adopted in commercial OSN platforms, such as Snapchat, Dust, 4chan, WeChat. Recently, [52] identified the problem that tweet deletion attracts unwanted attention from stalkers. In defense, a temporary withdrawal mechanism was developed to offer deletion privacy, so that adversarial observers could not (confidently) identify whether a tweet was really deleted.

(B) Private tweet identification. With the qualitative modeling of private/regrettable tweets, the need for automatic tweet assessment. In [50], Mao *et al.* developed one of the first mechanisms to identify potentially private tweets. They designed a classifier for private tweets using naive Bayes and SVM classifiers. However, the features and classifiers need to be fine-tuned for each category, and they only achieve approximately 80% accuracy in binary classifications. To tackle the challenges in handling short text in twitter, [33] aggregates tweets for each user, extracts topic matching, NER and sentiment features, and uses AdaBoost with Naive Bayes to classify each user into categories labeled as privacy score 1, 2 and 3. Recently, Zhou *et al.* proposed to examine the features of deleted tweets to predict if a tweet will be deleted [91]. They pre-selected ten topics that are commonly considered as sensitive (e.g., curses, drugs, etc.), and classified a tweet as sensitive or non-sensitive by checking if it contains keywords from the sensitive topics. Another recent project further classified private tweets into 13 or 14 pre-defined categories, using BoW/TFIDF features and Naive Bayes classifier [78, 79]. They assumed that sensitive tweets are pre-identified, without explaining how that could be achieved. Existing private tweet identification/classification approaches employ term-based features (BoW, TF-IDF, sentiment) and simple supervised classifiers, which cannot capture semantic features, or accurately discover topics containing subtle yet sensitive content. Moreover, the classification approaches only generate a binary notion on whether a tweet belongs to a pre-defined category.

Our project is primarily motivated by (A), which called for methods to assess private/sensitive tweet content, so that users could be alerted accordingly. While we have been inspired by existing methods in (B), our approach is novel in the following aspects: (1) We employ the state-of-art content representation and classification algorithms (word embedding and RNN) to significantly improve the accuracy of assessing general tweet content; (2) Instead of a binary notion of sensitive vs. nonsensitive, we generate a real score that provides more information on the *level of sensitiveness* and enables personalization in setting alerting preferences. (3) We have developed a general purpose solution that works for a broader scope of tweets, instead of only identifying certain types of private/sensitive tweets.

Tweet Classification. [40] integrated network information with text features to classify tweets into trending topics. [72] extracted time-related features such as time expressions to classify tweets based on their expiration. Similarly, [69] extracted features from authors' profiles and tweet histories to classify tweet topics. Although inspired by these approaches, our purpose and application domain are completely different from them.

9 Conclusion

In this paper, we make the first attempt to develop a computational model using deep neural networks to quantitatively assess the level of privacy/sensitiveness for textual content in OSNs. Our framework consists of four key components: (1) collection and analysis of privacy opinions on potentially private information; (2) the context-free privacy scoring model, which mimics users' privacy perceptions to assess the degree of privacy mainly based on text content; (3) the context-aware privacy scoring model, which considers the influences of the societal context on privacy perceptions; and (4) the personalized privacy scoring model, which integrates topic-specific personalized privacy attitude into the privacy scores. With experiments on human annotated data, we show that the PrivScore model is consistent with human perception of privacy.

10 Acknowledgement

This work is sponsored in part by the National Science Foundation under #1422206, #1565570, #1742702, #1820609, and #1915801, and the National Security Agency (NSA) Science of Security (SoS) Initiative.

References

[1] J. H. Abawajy, M. I. H. Ninggal, Z. A. Aghbari, A. B. Darem, and A. Alhashmi. Privacy threat analysis of mobile social network data publishing. In *SecureComm*, 2017.

[2] M. E. Acer, E. Stark, A. P. Felt, S. Fahl, R. Bhargava, B. Dev, M. Braithwaite, R. Sleevi, and P. Tabriz. Where the wild warnings are: Root causes of chrome https certificate errors. In *ACM CCS*, pages 1407–1420. ACM, 2017.

[3] H. Almuhimedi, S. Wilson, B. Liu, N. Sadeh, and A. Acquisti. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *ACM CSCW*, pages 897–908, 2013.

[4] R. Baden, A. Bender, N. Spring, B. Bhattacharjee, and D. Starin. Persona: an online social network with user-defined privacy. *SIGCOMM*, 2009.

[5] M. Bagdouri and D. W. Oard. On predicting deletions of microblog posts. In *ACM CIKM*, 2015.

[6] S. B. Barnes. A privacy paradox: Social networking in the united states. *First Monday*, 11(9), 2006.

[7] G. Blank, G. Bolsover, and E. Dubois. A new privacy paradox: Young people and privacy on social network sites. In *Annual Meeting of the American Sociological Assoc.*, 2014.

[8] 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.

[9] Z. Cai, Z. He, X. Guan, and Y. Li. Collective data-sanitization for preventing sensitive information inference attacks in social networks. *IEEE TDSC*, 15(4), 2018.

[10] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[11] H.-W. Chang, D. Lee, M. Eltaher, and J. Lee. @ Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *IEEE ASONAM*, 2012.

[12] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *ACM CIKM*, 2010.

[13] F. Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

[14] M. Ciot, M. Sonderegger, and D. Ruths. Gender inference of twitter users in Non-English contexts. In *EMNLP*, pages 1136–1145, 2013.

[15] J. Dawes. Do data characteristics change according to the number of scale points used? an experiment using 5-point, 7-point and 10-point scales. *IJMR*, 50(1):61–104, 2008.

[16] A. Dhir, T. Torsheim, S. Pallesen, and C. S. Andreassen. Do online privacy concerns predict selfie behavior among adolescents, young adults and adults? *Front Psy.*, 8, 2017.

[17] T. Dinev and P. Hart. Internet privacy concerns and social awareness as determinants of intention to transact. *International Journal of Electronic Commerce*, 10(2):7–29, 2005.

[18] C. Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.

[19] S. Egelman, L. F. Cranor, and J. Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *ACM CHI*, 2008.

[20] J. L. Fleiss and J. Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas.*, 33(3), 1973.

[21] J. Fogel and E. Nehmad. Internet social network communities: Risk taking, trust, and privacy concerns. *Computers in human behavior*, 25(1):153–160, 2009.

[22] N. Gerber, P. Gerber, and M. Volkamer. Explaining the privacy paradox-a systematic review of literature investigating privacy attitude and behavior. *Computers & Security*, 2018.

[23] Y. Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017.

[24] Google. Google pre-trained word2vec, 2013.

[25] L. Guthrie, E. Walker, and J. Guthrie. Document classification by machine: theory and practice. In *Conference on Computational linguistics*, 1994.

[26] A. Haeberlen, B. C. Pierce, and A. Narayan. Differential privacy under fire. In *USENIX Security Symposium*, 2011.

[27] E. Hargittai and A. Marwick. "what can i really do?" explaining the privacy paradox with online apathy. *International Journal of Communication*, 10:21, 2016.

[28] J. He, W. W. Chu, and Z. V. Liu. Inferring privacy information from social networks. In *ISI*, 2006.

[29] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[30] L. Humphreys, P. Gill, and B. Krishnamurthy. How much is too much? privacy issues on twitter. In *Conference of International Communication Association, Singapore*, 2010.

[31] G. Iachello, J. Hong, et al. End-user privacy in human–computer interaction. *Foundations and Trends in Human–Computer Interaction*, 1(1), 2007.

[32] Y. Ikawa, M. Enoki, and M. Tatsubori. Location inference using microblog messages. In *21st International Conference on World Wide Web*, pages 687–690, 2012.

[33] A. Islam, J. Walsh, and R. Greenstadt. Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *ACM WPES*, 2014.

[34] S. Jahid, P. Mittal, and N. Borisov. Easier: Encryption-based access control in social networks with efficient revocation. In *ACM AsiaCCS*, 2011.

[35] M. Johnson, S. Egelman, and S. M. Bellovin. Facebook and privacy: it's complicated. In *Eighth symposium on usable privacy and security*, page 9. ACM, 2012.

[36] Z. G. K. *The psychology of language*. Houghton-Mifflin, 1935.

[37] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[38] V. Lampos, N. Aletras, J. K. Geyti, B. Zou, and I. J. Cox. Inferring the socioeconomic status of social media users based on behaviour and language. In *ECIR*, 2016.

[39] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[40] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary. Twitter trending topic classification. In *IEEE ICDM Workshops*, 2011.

[41] K. Lewis, J. Kaufman, and N. Christakis. The taste for privacy: An analysis of college student privacy settings in an online social network. *J Comp Mediat Comm.*, 14(1), 2008.

[42] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, 2007.

[43] R.-H. Li, J. Liu, J. X. Yu, H. Chen, and H. Kitagawa. Co-occurrence prediction in a large location-based social network. *Frontiers of Computer Science*, 7(2):185–194, 2013.

[44] E. Litt. Understanding social network site users' privacy tool use. *Computers in Human Behavior*, 29(4):1649–1656, 2013.

[45] 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*, 5(1), 2010.

[46] W. Liu and D. Ruths. What's in a name? using first names as features for gender inference in twitter. In *AAAI spring symposium: Analyzing microtext*, volume 13, page 01, 2013.

[47] B. Luo and D. Lee. On protecting private information in social networks: a proposal. In *IEEE ICME Workshop of M3SN*. IEEE, 2009.

[48] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3, 2007.

[49] J. Mahmud, J. Nichols, and C. Drews. Home location identification of twitter users. *ACM TIST*, 5(3):47, 2014.

[50] H. Mao, X. Shuai, and A. Kapadia. Loose tweets: an analysis of privacy leaks on twitter. In *ACM WPES*, 2011.

[51] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[52] M. Minaei, M. Mondal, P. Loiseau, K. Gummadi, and A. Kate. Lethe: Conceal content deletion from persistent observers. *Privacy Enhancing Technologies*, 2019.

[53] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *ACM WSDM*, 2010.

[54] M. Mondal, J. Messias, S. Ghosh, K. P. Gummadi, and A. Kate. Forgetting in social media: Understanding and controlling longitudinal exposure of socially shared data. In *SOUPS 2016*, pages 287–299, 2016.

[55] K. Moore and J. C. McElroy. The influence of personality on facebook usage, wall postings, and regret. *Computers in Human Behavior*, 28(1):267–274, 2012.

[56] S. Patil, G. Norcie, A. Kapadia, and A. J. Lee. Reasons, rewards, regrets: privacy considerations in location sharing as an interactive practice. In *SOUPS*, 2012.

[57] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[58] D. Pergament, A. Aghasaryan, J.-G. Ganascia, and S. Betgé-Brezetz. Forps: Friends-oriented reputation privacy score. In *First International Workshop on Security and Privacy Preserving in e-Societies*, pages 19–25, 2011.

[59] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[60] T. Pontes, G. Magno, M. Vasconcelos, A. Gupta, J. Almeida, P. Kumaraguru, and V. Almeida. Beware of what you share: Inferring home location in social networks. In *ICDM Workshops*. IEEE, 2012.

[61] D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9), 2015.

[62] R. W. Reeder, A. P. Felt, S. Consolvo, N. Malkin, C. Thompson, and S. Egelman. An experience sampling study of user reactions to browser warnings in the field. In *ACM CHI*, page 512. ACM, 2018.

[63] S. Robertson, H. Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[64] E.-M. Schomakers, C. Lidynia, D. Müllmann, and M. Ziefle. Internet users' perceptions of information sensitivity–insights from germany. *International Journal of Information Management*, 46:142–150, 2019.

[65] M. Sleeper, R. Balebako, S. Das, A. L. McConahy, J. Wiese, and L. F. Cranor. The post that wasn't: exploring self-censorship on facebook. In *ACM CSCW*, 2013.

[66] 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 *ACM CHI*, pages 3277–3286, 2013.

[67] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

[68] A. Sotirakopoulos, K. Hawkey, and K. Beznosov. On the challenges in usable security lab studies: lessons learned from replicating a study on ssl warnings. In *SOUPS*. ACM, 2011.

[69] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. In *ACM SIGIR*. ACM, 2010.

[70] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying wolf: An empirical study of ssl warning effectiveness. In *USENIX Security*, 2009.

[71] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[72] H. Takemura and K. Tajima. Tweet classification based on their lifetime duration. In *ACM CIKM*, 2012.

[73] S. Talukder and B. Carbunar. Abusniff: Automatic detection and defenses against abusive facebook friends. In *AAAI Conference on Web and Social Media*, 2018.

[74] Twitter. Api reference index.

[75] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *ICWSM*, 2017.

[76] A. Vasalou, A. J. Gill, F. Mazanderani, C. Papoutsi, and A. Joinson. Privacy dictionary: A new resource for the automated content analysis of privacy. *J Am Soc Inf Sci Technol.*, 62(11):2095–2105, 2011.

[77] S. Volkova and Y. Bachrach. On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. *Cyberpsychol Behav Soc Netw.*, 18(12), 2015.

[78] Q. Wang, J. Bhandal, S. Huang, and B. Luo. Classification of private tweets using tweet content. In *IEEE ICSC*, 2017.

[79] Q. Wang, J. Bhandal, S. Huang, and B. Luo. Content-based classification of sensitive tweets. *International Journal of Semantic Computing*, 11(04):541–562, 2017.

[80] Y. Wang, P. G. Leon, A. Acquisti, L. F. Cranor, A. Forget, and N. Sadeh. A field trial of privacy nudges for facebook. In *ACN CHI*, pages 2367–2376, 2014.

[81] Y. Wang, P. G. Leon, X. Chen, and S. Komanduri. From facebook regrets to facebook privacy nudges. *Ohio St. LJ*, 74:1307, 2013.

[82] 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 *ACM SOUPS*, page 10, 2011.

[83] J. Weinberger and A. P. Felt. A week to remember: The impact of browser warning storage policies. In *SOUPS*, 2016.

[84] M. Wu, R. C. Miller, and S. L. Garfinkel. Do security toolbars actually prevent phishing attacks? In *ACM CHI*, 2006.

[85] W. Xie and C. Kang. See you, see me: Teenagers' self-disclosure and regret of posting on social network site. *Computers in Human Behavior*, 52:398–407, 2015.

[86] J.-M. Xu, B. Burchfiel, X. Zhu, and A. Bellmore. An examination of regret in bullying tweets. In *HLT-NAACL*, 2013.

[87] C. Yang and P. Srinivasan. Translating surveys to surveillance on social media: methodological challenges & solutions. In *ACM Web science*, 2014.

[88] Y. Yang, J. Lutes, F. Li, B. Luo, and P. Liu. Stalking online: on user privacy in social networks. In *Proceedings of the second ACM conference on Data and Application Security and Privacy*, 2012.

[89] L. Yu, S. M. Motipalli, D. Lee, P. Liu, H. Xu, Q. Liu, J. Tan, and B. Luo. My friend leaks my privacy: Modeling and analyzing privacy in social networks. In *SACMAT*, 2018.

[90] A. Zarras, K. Kohls, M. Dürmuth, and C. Pöpper. Neuralyzer: flexible expiration times for the revocation of online data. In *ACM CODASPY*, 2016.

[91] L. Zhou, W. Wang, and K. Chen. Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones. In *World Wide Web*, 2016.

A The GloVe Model

The Global Vectors for Word Representation (GloVe) [57] word embedding algorithm leverages the global word co-occurrence statistics in the training set and the vector space semantic structure captured in Word2Vec. It represents an aggregated global word-word co-occurrence matrix as $\mathbf{X}$, in which the element X_{ij} denotes the number of times a word j occurs in the context of the word i. The soft constraints for each word pair is defined as:

$$w_i^T \tilde{w}_j + b_i + \tilde{b}_j = logX_{ij} \tag{14}$$

where w_i and $\tilde{w}_j$ are the main and context word vectors, and b_i and $\tilde{b}_j$ are scalar biases for main and context words. To avoid weighing all co-occurrences equally, GloVe adopts a weighted least squares cost function:

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - logX_{ij})^2 \tag{15}$$

where $f(X_{ij})$ is the weighting function in the form of:

$$f(X_{ij}) = \begin{cases} (X_{ij}/X_{max})^{\alpha} & if\ X_{ij} < X_{max} \\ 1 & otherwise \end{cases} \tag{16}$$

The model generates two sets of word vectors, $\mathbf{W}$ and $\tilde{\mathbf{W}}$. Since $\mathbf{X}$ is a symmetric matrix, $\mathbf{W}$ and $\tilde{\mathbf{W}}$ are equivalent and differ only as a result of their random initializations. Therefore, the sum $\mathbf{W} + \tilde{\mathbf{W}}$ is used as the word vectors to reduce overfitting.

B RNN and LSTM

Deep neural networks (e.g., RNN [23], LSTM [29]) are widely adopted to boost the performance of classifiers. In an RNN with one hidden layer, the input, output and hidden layers are denoted as $\mathbf{X}$, y and $\mathbf{h}$, and the network is formalized as:

$$\begin{aligned} \mathbf{h}^t &= \sigma(\mathbf{W}_h \mathbf{X} + \mathbf{W}_r \mathbf{h}^{t-1}) \\ y &= \sigma(\mathbf{W}_y \mathbf{h}^t) \end{aligned} \tag{17}$$

where $\mathbf{W}_h$ and $\mathbf{W}_y$ are the weights from input layer to hidden layer and from hidden layer to out layer, and $\mathbf{W}_r$ is the weight of the recurrent computation. Each layer trains a single unit, which can be an arbitrarily shaped network taking a vector as input and outputting another vector of the same size. Therefore, it fits the processing of sequential information.

The repeated training of the same parameters in RNN also cause the exploring/vanishing problems during backpropagation. To avoid overfitting, it is vital to adopt proper regularizations and complex architectures that fit the specific formats and requirements of the data. LSTM [29] proposed to add several neural layers to control the flow of information, such as what is added to the state, what is forgotten or output from the state. The forget gate can be represented as:

$$f_i^{(t)} = \sigma(b_i^f + \sum_j U_{i,j}^f x^{(t)_j} + \sum_j W_{i,j}^f h_j^{(t-1)}) \tag{18}$$

where $\mathbf{x}^{(t)}$ is the current input vector and $\mathbf{h}^{(t-1)}$ is the previous hidden layer vector. $\mathbf{b}^f, \mathbf{U}^f$ and $\mathbf{W}^f$ are biases, input weights and recurrent weights for the forget gates, respectively. Similarly, the external input gate $g_i^{(t)}$ and output gate $q_i^{(t)}$ can be computed as:

$$g_i^{(t)} = \sigma(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}) \tag{19}$$

$$q_i^{(t)} = \sigma(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}) \tag{20}$$

The output $h_i^{(t)}$ is controlled by output gate $q_i^{(t)}$, while the self-loop weight is controlled by both the forget gate and the external input gate as:

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}) \tag{21}$$

Finally, we get the basic architecture of the LSTM as:

$$h_i^{(t)} = tanh(s_i^{(t)}) q_i^{(t)} \tag{22}$$

With the self-loop state and three gates, LSTM learns to memorize long time dependency but forgets the past information if needed so that it does not need to keep track of information over long sequences.

C Inter-Rater Agreement

In this section, we introduce three approaches that are used to assess Inter-rater Agreement (IRA): *Fleiss' Kappa* measures the agreements between raters on categorical labels, *Pearson Correlation* measures the linear dependency between two variables, and *Spearman Correlation* measures the strength of monotonic (but not necessarily linear) relationship between two variables.

Fleiss Kappa. To statistically measure the agreement between two raters on categorical labels, *Cohen's Kappa* was introduced as a more reliable indicator than calculating percentage of agreements. *Fleiss' Kappa* extended Cohen's Kappa to measure the agreement between more than two raters. The Kappa, k, is defined as:

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{23}$$

In this formula, the denominator denotes the *agreement by chance*, i.e., the degree of agreement among multiple raters that is attainable above chance. The numerator denotes the *observed agreement*, i.e., the degree of agreement that is achieved by these raters. That is, *Fleiss' Kappa* quantitatively measures the actual degree of agreement in comparing with completely random raters, i.e., the level of agreement when the raters' selections are completely random [20]. A smaller k (e.g., $k < 0$) indicates poor agreement among raters, while a larger K (e.g., $k \to 1$) indicates good agreement.

Pearson Correlation. Fleiss' Kappa was designed for categorical data, therefore, it treats each label as an independent category. In our experiments, when two raters label a tweet as [1 Very sensitive] and [2 Sensitive], while another two raters label two tweets as [1 Very sensitive] and [5 Nonsensitive], they are considered as equally inconsistent by Fleiss' Kappa. But in reality, 1 and 2 are significantly more consistent than 1 and 5. To better handle numerical data, *Pearson Correlation* was designed to capture the linear dependency between two variables X and Y, which is denoted as:

$$r = (\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})) / (\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}) \tag{24}$$

where x_i and y_i are indexed samples from two variables, $\bar{x}$ and $\bar{y}$ denotes the sample mean. The numerator captures the covariance of the two variables, while the denominator denotes the standard deviations of X and Y. For non-negative variables, $r = 0$ indicates that there is no linear correlation, while $r = 1$ indicates a perfectly linear relationship between X and Y.

Spearman's Rank Order Correlation. Last, the Spearman Correlation captures the agreement between two annotators in terms of the correlation between the ranks of their labels. It is very similar to Pearson, but it considers the relationships between the ranks of X and Y, instead of directly on X and Y. It is specified as:

$$r_s = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}} \tag{25}$$

where rg_X and rg_Y denotes the rank order of X and Y. The correlation coefficient is interpreted similarly as Pearson correlation, where 0 indicates no correlation in rank orders and 1 indicates perfect correlation between rank orders of X and Y.

D User Study

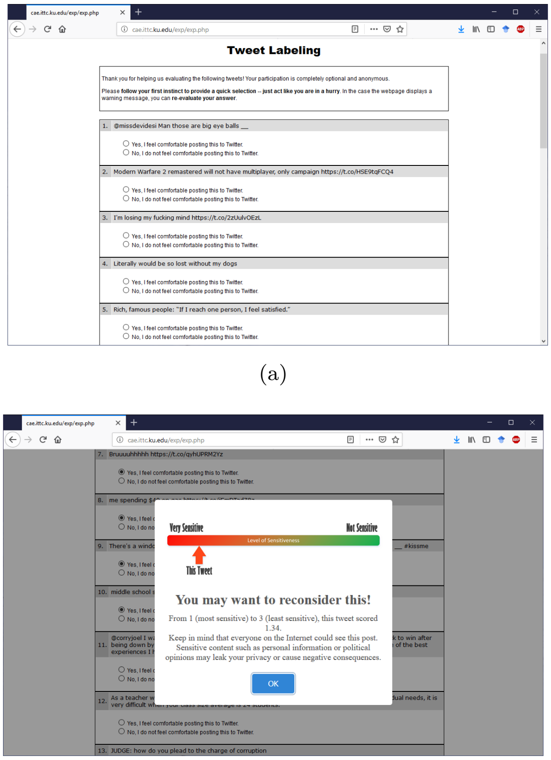
To demonstrate the effectiveness of employing PrivScore in triggering self-censorship, we perform a proof-of-concept evaluation for user alerting (an IRB approval was obtained for this evaluation).

a. We recruited college students to evaluate tweets crawled from the Internet. Each participant received a URL to an anonymous online questionnaire, which includes 15 tweets (sampled with higher density of sensitive tweets than the original distribution).

b. Participants were asked to select "1. Yes I feel comfortable posting this to Twitter" or "2. No, I do NOT feel comfortable posting this" for each tweet, as shown in Fig. 8 (a). To mimic an emotional or urgent scenario, we asked students to "*follow your first instinct to provide a quick selection–just act like you are in a hurry.*"

c. If the student chose "Yes" for a low score tweet (i.e., a tweet with $S_{cf} < 1.5$), a warning message was displayed (as shown in Figure 8 (b)). The student has the option to adhere to the advice or stick to her original option.

We record the participant's selection for each tweet. We also record whether the warning message was triggered, and whether the participant adhered to the message. Out of 780 tweets (52 questionnaires) that was answered in 10 days, 93 tweets triggered the warning message, while users changed opinions on 58 tweets: an adherence rate of 62.3%. Manual inspection also showed that users stick to their original selection of "Yes" mostly for political and judgemental tweets, which indicates that our annotators were more conservative on political content than the evaluators. There were also a few false positives. One example was a tweet criticizing racism, which received a PrivScore of 1.21 (strong critical tones and racist terms). However, since it was criticizing racism, it should not receive such a low score.



Tweet Labeling

Thank you for helping us evaluating the following tweets! Your participation is completely optional and anonymous.

Please **follow your first instinct to provide a quick selection -- just act like you are in a hurry**. In the case the webpage displays a warning message, you can **re-evaluate your answer**.

1. @missdevidesi Man those are big eye balls __
   ○ Yes, I feel comfortable posting this to Twitter.
   ○ No, I do not feel comfortable posting this to Twitter.
2. Modern Warfare 2 remastered will not have multiplayer, only campaign https://t.co/HSE9tqFCQ4
   ○ Yes, I feel comfortable posting this to Twitter.
   ○ No, I do not feel comfortable posting this to Twitter.
3. I'm losing my fucking mind https://t.co/2zUulvOEzL
   ○ Yes, I feel comfortable posting this to Twitter.
   ○ No, I do not feel comfortable posting this to Twitter.
4. Literally would be so lost without my dogs
   ○ Yes, I feel comfortable posting this to Twitter.
   ○ No, I do not feel comfortable posting this to Twitter.
5. Rich, famous people: "If I reach one person, I feel satisfied."
   ○ Yes, I feel comfortable posting this to Twitter.
   ○ No, I do not feel comfortable posting this to Twitter.

(a)

7. Bruuuuhhhhh https://t.co/qyhUPRM2Yz
   ● Yes, I feel comfortable posting this to Twitter.
   ○ No, I do not feel comfortable posting this to Twitter.
8. me spending $4
   ● Yes, I feel c
   ○ No, I do no
9. There's a windo __ #kissme
   ● Yes, I feel c
   ○ No, I do no
10. middle school s
   ● Yes, I feel c
   ○ No, I do no
11. @corryjoel I wa k to win after being down by of the best experiences I h
   ○ Yes, I feel c
   ○ No, I do no
12. As a teacher w dual needs, it is very difficult wh
   ○ Yes, I feel comfortable posting this to Twitter.
   ○ No, I do not feel comfortable posting this to Twitter.
13. JUDGE: how do you plead to the charge of corruption

Very Sensitive — Not Sensitive
Level of Sensitiveness
This Tweet

You may want to reconsider this!

From 1 (most sensitive) to 3 (least sensitive), this tweet scored 1.34.
Keep in mind that everyone on the Internet could see this post.
Sensitive content such as personal information or political opinions may leak your privacy or cause negative consequences.

OK

(b)

Fig. 8. User study on the effectiveness of user alerting.

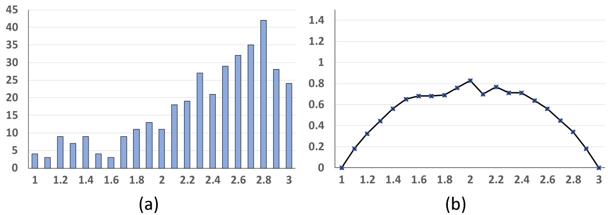


Fig. 9. Statistics of tweets with 10 annotations: (a) Distribution of the mean annotated score, X: Mean annotated score $\bar{S}_A$ of tweets, Y: Number of tweets in each bin; (b) Distribution of Mean Absolute Deviation (MAD), X: $\bar{S}_A$, Y: average MAD of tweets in each bin.

E Understand the Annotations

To learn more insights of annotators' rationale and agreements, and to confirm our prior observations, we added two small-scale experiments on MTurk.

I. More Annotators for Each Tweet. We posted 20 questionnaires to MTurk, and recruited 10 Turkers to annotate each questionnaire. Each tweet was annotated as: "1 Sensitive", "2 Maybe", or "3 Nonsensitive". Excluding attention check tweets, we collected 3,600 annotations for 360 tweets. For each tweet, we calculated the mean annotated score $\bar{S}_A$, and displayed its distribution across all the tweets in Fig. 9 (a). We also calculated the mean absolute deviation (MAD) of the 10 annotations for each tweet. Fig. 9 (b) shows the average of MAD for tweets in each category of $\bar{S}_A$. The results are consistent with our observations: Turkers show more consistency with the clearly nonsensitive or highly sensitive tweets, i.e., both ends of X-axis in Fig. 9 (b). They demonstrate relatively low consistency on non-extreme tweets.

II. Annotation with Open-ended Questions. In the second experiment, we asked each Turker to justify his/her annotation in a textbox. We posted 65 questionnaires (10 tweets in each questionnaire) to MTurk at the rate of $1.2 per questionnaire. We accepted 61 responses that passed the attention check tweets. They were completed in 286 to 2845 seconds. The median completion time was 811.5 seconds. Most of the responses that corresponds to "very sensitive", "sensitive" and "little sensitive" annotations point out a type of sensitive content. However, some of them were simply justified as "inappropriate content" or "bad personal image".

We qualitatively analyzed the responses by coding each response and categorize them according to the types of sensitive information. The most popular types of sensitive tweets are "obscene content", "drug", "cursing", "attack","dirty words", "discrimination", and "personal information". Meanwhile, the most popular justifications for non-sensitive tweets are: "does not contain sensitive/personal information", "nothing harmful/offensive", "positive or nothing negative", and "nothing big". Although the scale of this experiment is small due to limited timing/budget, our results are consistent with existing research in the literature [66, 82].