Privacy Estimation on Twitter: Modelling the Effect of Latent Topics on Privacy by Integrating XGBoost, Topic and Generalized Additive Models*

Arne Tillmann*, Lindrit Kqiku*†, Delphine Reinhardt*†, Christoph Weisser*, Benjamin Säfken‡, Thomas Kneib*
*Campus Institute Data Science (CIDAS), Georg-August-Universität Göttingen, Göttingen, Germany
†Insitute of Computer Science, Georg-August-Universität Göttingen, Göttingen, Germany
‡Clausthal University of Technology, Clausthal, Germany
Emails: {a.tillmann@stud., kqiku@cs., reinhardt@cs., c.weisser@stud., benjamin.saefken@, tkneib@}uni-goettingen.de

***Abstract*—Securing their users' privacy is a central duty of Online Social Networks (OSN), but the complex non-linear effects of social media content on privacy is not well understood. We propose a novel framework that integrates XGBoost, *Latent Dirichlet Allocation* (LDA) topic models and *Generalized Additive Models* (GAM) to perform statistical inference about the complex non-linear relationship between the topics and privacy of tweets. First, XGBoost is used to predict the privacy of tweets. Then, the predictions are improved by analyzing the classified tweets with LDA topic models. Finally, we model the non-linear relationship between topics and privacy with GAMs by using (penalized) splines. Instead of being limited to predictive modeling, our approach enables us to model the non-linear relationship between latent topics and the privacy of tweets.**

***Index Terms*—Privacy, Tweets, Topic Model, Latent Dirichlet Allocation, Generalized Additive Model**

I. Introduction

Internet users worldwide spend a significant amount of their daily activities on social media, leading to an average of 147 minutes, an increase by two minutes compared to last year [52]. On Twitter, users post at least about 500 million tweets per day [53]. Protecting the privacy of individuals is a duty of *Online Social Networks* (OSN) according to *General Data Protection Regulation* (GDPR), article 12 human rights, The Communications Decency Act (CDA), and The Children's Online Privacy Protection Act (COPPA). Current OSN privacy settings remain inefficient in such a task. To this end, several countermeasures have been proposed. Current approaches to secure privacy focus mostly on hiding the identity (e.g., hiding addresses to reduce the risk of stalking) [22], [29], [36], [39], [43], [59], [60]. Facebook, e.g., allows users to select a pre-defined group with which they share posts (e.g., all, friends), or to define the groups by manually selecting other users [35]. In contrast, Twitter, upon account creation, shares users' tweets by default with everyone unless a user decides to make the account private. In such a case, tweets are accessible only to user's followers. Recently, twitter circle option is introduced. Users can add up to 150 people and share the tweets with the circle only [65].

These methods are insufficient in matching users' desired privacy standards, resulting in regrettable posts [50], [55]. Moreover, the current mechanisms are time-consuming and complex for users [31] and often lead to sharing content with an unintended audience [23], [55]. Thus, users need to maintain their access control lists constantly. In practice, users, however, often reuse and rarely update such lists [56].

To address these issues, Reinhardt et al. [47] propose privacy suggestions that rely on the sensitivity of content and the current strength of the user's social relationships to suggest the appropriate audience for the post to a user. For such a privacy suggestion scheme, a user interface is conceptualized in [47] and further designed in a follow-up work [64]. Content about to be shared may greatly vary, from personal insights (incl. emails, phone numbers, and address) to location, time of day, and information about tweets' authors. In the case of Twitter, however, while only 0.1% of users mentioned identifiable attributes such as email addresses or phone numbers in their tweets according to [38], other types of disclosure of sensitive information were more common. For example, in their sample, 12.1% of tweets included a person's location, 20.1% included the time of certain activities and 22.7% an actual name. Such disclosures can lead to severe consequences, especially when coupled together. For instance, combining an identified person's name with location and time of day can be used for robbery. Therefore, we argue that another critical component in privacy protection in OSNs is supporting users to control better the dissemination of sensitive information beyond their identities and profile attributes.

Bridging the gap between the desired privacy in OSNs and the actual privacy settings as proposed in [47] is challenging, as privacy perceptions depend on cultural and individual traits [11], [40], [48]. Moreover, studies that explore the influential factors that characterize private content being shared remain limited. Thus, in this paper, we strive toward closing this gap. Our contribution is as follows:

- We use a set of tweets that users previously labeled

*The final publication is available at IEEE Xplore via https://doi.org/10.1109/SmartWorld-UIC-ATC-ScalCom-DigitalTwin-PriComp-Metaverse56740.2022.00365 private and public by Wang et al. [54] to train a classification model based on XGBoost classifier.

- We next crawled a set of tweets geo-located in Great Britain and used the previously trained classifiers on [54] to classify the crawled tweets' privacy into private or non-private tweets. Note that this is a necessary step for the following modelings due to the small size of the originally labeled data set [54].
- We fit LDA models on the classified tweets and use the topic probabilities to model the privacy classification with Generalized Additive Models (GAM).
- By using GAMs and splines, we further model the complex non-linear relationship between topics and privacy.

To this end, we find groups of topics that are more likely correlated with sensitive content. For instance, the topics related to *sexuality* positively impact content sensitivity. Moreover, we demonstrate that tweets posted during nighttime or that have a more negative sentiment are more likely to be sensitive. The latter remains consistent with similar studies that show that sentiment is a valuable parameter in estimating the sensitivity of textual content [8], [9], [54]. Furthermore, the negative users' expressions toward Covid and politics (precisely detected as "fuck covid & politics") is an interesting topic that is highly correlated with a tweet being sensitive. In addition to above findings, our framework, in general, can also be adopted to the investigation of sensitive topics in different constrains (e.g, different countries) and the role of structural variables (such as time or sensitivity of posted tweets), by estimating content sensitivity in one hand, as well as, LDA topic modeling and GAMs in the other.

The remainder of this paper is structured as follows. First, we detail related work in Sec. II and describe the data, i.e., the labeled tweets and privacy dictionaries by [54], and the scraped tweets in Sec. III. Then, in Sec. IV, we outline the used methods, i.e., the classifier XGBoost, LDA topic model, and GAM. Afterward, we present our results in Sec. V before we discuss the limitations and future directions in Sec. VI. Finally, in Sec. VII, we make concluding remarks and provide suggestions for further research.

II. Related Literature

We briefly discuss the literature on users' behaviours with regard to content sensitivity in Sec. II-A, followed by users' privacy perception of sensitive information in general (Sec. II-B). Lastly, we explore two different privacy-enhancing approaches, i.e., tweet sensitivity prediction (Sec. II-C) and deceptive tweets (Sec. II-D).

A. Behavioral Patterns in OSN and Sensitivity

Users adopt different measures to protect their privacy. For example, Twitter users can choose between having usernames that disclose their identities or can remain anonymous. Peddinti et al. [42] investigated the behavior of the so-called identifiable and anonymous users. They, in particular, explored the posts' topics that identifiable and anonymous users decide to share on Twitter. They uncovered a link between sensitive information and the tendency to have an anonymous user account. They also found that such users tweet more and tend to expose their activity to the general audience. Our work, however, focuses on tweet content and topics to investigate content sensitivity.

B. Information Sensitivity Perception

Several studies explored users' perceptions of information sensitivity [11], [40], [48]. Markos et al. [40] performed a cross-national and cross-cultural investigation of privacy-related behaviors in an empirical study. They investigated consumers' perceived sensitivity and willingness to provide information regarding country, age group, perceived privacy control, consumer data relationship, and type of information. They found that US users are more sensitive to the information collection and less willing to disclose information than Brazilians. They both, however, ranked similarly in the orderings of sensitivity for specific types of information.

In another study, Schomakers et al. [48] compared the perceived sensitivity of German internet users against the results of the US and Brazil from Markos et al. [40]. Their study yielded the same findings as Markos et al. [40]. While differences in the perception of sensitivity between Germany, Brazil, and the US were noticed, the ranked orderings of specific types of sensitive information are similar between all three countries. Almotairi et al. [11] extended the study to Saudi Arabia cohorts. They also found slight differences in the perception of information sensitivity compared to the US, Brazil, and German cohorts. Alemany et al. [61] further defined a sensitive ranking of information types from related works on types of regrets.

The literature shows that a certain consensus on the perceived types of sensitive information exists across nations and cultures. In this work, we investigate tweets from Great Britain. Our framework can be easily adapted to study information privacy of tweets in other countries.

C. Tweet Sensitivity Prediction

Sensitivity prediction of tweets using *Machine Learning* (ML) techniques falls into two main groups, namely, (1) topic-based supervised classification solutions and (2) content-based sensitivity.

Mao et al. [2] conducted an exploratory study to investigate whether pre-filtered revealing vacation and drunk tweets are sensitive or non-sensitive. They used naive Bayes and SVM ML classification models. They pre-filtered vacation tweets using keyword search on terms that may reveal the vacation plans like *vacation*, and *fly to*. They further analyzed the role of getting drunk and tweeting in privacy exposure by classifying drunk tweets as sensitive or non-sensitive and analyzing the topic of drunk sensitive tweets. They also designed a classifier to predict drunk-driving tweets. Moreover, they investigated the classification of tweets that contain disease information. However, their solution can be adopted for specific scenarios only (i.e, vacation and drunk) and not generic model solutions to classify any tweet as sensitive and non-sensitive despite their topics, as in our case.

Caliskan-Islam et al. [3] designed a tool to detect the privacy score of a user by calculating if a user's list (timeline) of tweets falls under one of three defined privacy scores, depending on the number of sensitive categories. They used topic matching extraction, a *Named Entity Recognition* (NER), and sentiment analysis together with AdaBoost and Naive Bayes ML classification models. During the annotation step, users were provided with a list of sensitive categories and had to decide whether a set of tweets were private or not and the sensitivity topic to which sensitive tweets belonged.

In two other studies, Wang et al. [6], [7] designed classifiers that predict one of 13 or 14 pre-defined sensitive topics of a tweet. We used the privacy dictionary based on Caliskan-Islam et al. [3] and extended by Wang et al. [54]. However, we do not focus on the classification of certain sensitive topics but rather on either sensitive or non-sensitive tweets content and on modeling the non-linear relationship between identified topics from those sensitive tweets.

Zhou et al. [4] aimed to predict whether a tweet is likely to be deleted or not based on previously deleted tweets. They extracted ten features based on a keyword search of a set of sensitive topics defined in another study [5]. They also generated user-based features from users' historical sharing and deleting patterns. They used conventional ML methods and found that Naive Bayes reaches the highest score. They used *Bag-of-Words* (BOW) and *Term Frequency-Inverse Document Frequency* (TF-IDF) to extract the features. The tweets are classified binary whether they belong to any of each category. A further limitation of their approach is the assumption that the private tweets are already pre-selected into one of the pre-defined categories.

In a similar research line, certain studies examined content sensitivity according to a binary categorization as either private or public content [8], [9], [54]. In a more recent study [54], use word embedding and recurrent neural network methods, specifically *Long Short Term Memory Network* (LSTM), to classify private tweets in either private or public categories. They further calculated a privacy score that includes content sensitivity classification, sentiment analysis, and user preferences. Mittal et al. [8] analyzed the role of users' sentiment in revealing sensitive content in tweets. They used *Bidirectional Encoder Representations from Transformers* (BERT) as a more recent transfer learning technique to classify content sensitivity and investigate the sentiment effect of sensitive and non-sensitive tweets. Sentiment analysis over the WASSA data set [10] was used to predict whether a tweet belongs to one of four basic emotion categories (i.e., *anger, joy, sadness, and fear*). Kqiku [9] et al. used the data set of [54] to fine-tune BERT in predicting content sensitivity and enhance further by coupling together sentiment estimation (i,e., *anger*, *disgust*, *joy*, *surprise*, *fear*, and *sad* emotions) in content sensitivity prediction. While we make use of classification models to train and evaluate classification models to recognize the sensitivity of tweets as in [54], we differ from other sensitivity prediction approaches by further exploring the topics of sensitive tweets using LDA and by modeling the non-linear relationship between topics and privacy using GAMs.

D. Deceptive Deletion of Tweets

Minaei et al. [62] found that even the act of deletion attracts unwanted attention from malicious parties. Therefore, they introduced a mechanism to withdraw tweets temporarily. They aimed to disguise malicious parties [62]. In another study, Minaei et al. [63] introduced another mechanism to protect withdrawn tweets from the attention of adversaries. They provided a deletion functionality to remove non-sensitive posts to confuse the adversaries in detecting sensitive posts.

Our aim, however, is to model the relationships between topics and sensitivity, along with exploring sensitive topics of tweets and tweets' sensitivity.

III. Dictionaries, Data, and Preprocessing

In this section, we first describe the dictionaries, and the labeled data set. The dictionaries and the labeled tweets by individuals are established by Wang et al. [54]. They extend the terms from other related works [3], [5], [50] for each sensitive topic by using the *Urban Dictionary website*. They ended up with more than 100 terms for 13 topics. They used the defined dictionary to filter a set of tweets to be annotated.

We next present our scraped unlabeled tweets and finally outline the data pre-processing steps.

A. Dictionaries

To create the dictionaries of potentially private topics, such as Health & Medical, Drugs & Alcohol, Obscenity, Politics, Racism, Family & Personal [54] use a root set of "seed terms," and expand the set using Urban Dictionary. The dictionaries are available from their website https://bit.ly/privscore. They consist of 13 defined privacy topics and range from 96 to 209 terms for a topic. For example, the first 20 words from the dictionary `entertainment` can be found in Tab. I. They [54] used the dictionary to reach a high number of potentially private tweets. Subsequently, they conducted a user study to ask how sensitive they consider a subset of those tweets, as explained more thoroughly in Sec. III-B.

music	cinema	silver screen	movie industry
rap	rock	melody	tune
classical	song	jazz	fusion
harmony	acoustic	folk	hard rock
instrumental	opera	cappella	heavy metal

TABLE I
First 20 words from the dictionary entertainment

B. Labeled Tweets

We employ the labeled tweet dataset from Wang et al. [54].

They [54]performed a snowball crawling process for about a month in March 2016 to collect a large number of tweets from different users and subsequently pre-filter them based on the dictionary above in Sec. III-A. Next, they conducted an IRB-approved user study to label selected tweets according to their perceived privacy sensitivity using Amazon Mechanical Turk. Each participant classified 20 randomly chosen tweets out of the original 6M tweets. Each tweet could be classified as [1:Very sensitive]; [2:Sensitive]; [3:Little Sensitive]; [4:Maybe]; [5:Non-sensitive]. Three different Turkers label the same tweets. The authors collected 552 qualified questionnaires from 1,656 Turkers. After post-filtering, they ended up with a data set of 3,008 labeled tweets in three categories (i.e., 1,436 *sensitive*, 61 *maybe*, and 1,512 *non-sensitive* tweets). The authors provided a version of the data set with an equal number of sensitive and non-sensitive tweets, i.e., 1,435 tweets for each category.

C. Unlabeled Scraped Tweets

We scrape unlabeled tweets using the official Twitter API from 2020-10-25 10:50:37 to 2020-12-26 18:57:36. We scrape them by using a geographical box that is large enough to encircle Great Britain (GB) and scrape all tweets with a geo-location inside that box. Subsequently, we filter out all tweets that do not have an English language stamp and tweets which do not have "GB" as country code. This process resulted in a total of 526,000 tweets. Each scraped tweet consists of the features presented in Tab. II.

1. created_at	2. id_str	3. user_id_str
4. full_text	5. hashtags	6. lang
7. place_full_name	8. country_code	9. coordinates
10. center_coord_X	11. center_coord_Y	

TABLE II
FEATURES OF THE UNLABELED SCRAPED TWEETS

D. Data Preprocessing for Classification

1) Preprocessing of the Training Data: As described in Sec. III-B, we reverse the labelling of the original data set ([0:Non-private], [1:Private]). We start the preprocessing by converting all letters to lowercase, removing the substring "@XXX" in all tweets, and removing the following punctuation marks: `! ( ) - [ ] ; : ′ " \, < > . / ? @ % & $ ^ * _ { }` ∼. Note that we remove the substring "@XXX" to avoid any undesired classification since "XXX' replaces the real identity of the user in the anonymized labeled data set introduced in Sec. III-B.

This process continues with tokenizing the text corpora using [14] (`nltk.tokenize.word_tokenize`), removing stop words (`nltk.corpus.stopwords`) and finally lemmatizing (`nltk.stem.lemmatizer`). Tokenizing is the process of breaking a stream of textual data into words, terms, sentences, symbols, or some other meaningful elements called tokens while lemmatizing in linguistics is the process of grouping together the inflected forms of a word so that they can be analyzed as a single item, identified by the word's lemma, or dictionary form. After we finish the data cleaning, we bring the tweets into a vectorized form. The vectorized form allows us to use numerical methods for text analysis. We use (`sklearn.sklearn.feature_extraction.text.CountVectorizer`) for the vectorization [45].

This process then gives us a vector representation for each tweet i of all I tweets over the dictionary of all W words that are included in the labeled tweets. In other words, $x_i^{(\omega)} \in \mathbb{N}$ indicates how often the word ω appears in the tweet i. The privacy label is denoted as y_i.

$$\text{labeled data} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{I}, \quad \mathbf{x}_i \in \mathbb{N}^W, \quad y_i \in \{0,1\}$$

The last step before the classifier training is to separate the labeled data into a training set and a test set to validate the resulting classifiers.

2) Preprocessing of the Unlabeled Scraped Tweets: The preprocessing of the unlabeled tweets is essentially very similar to the preprocessing of the labeled tweets introduced in Sec. III-D1. We first discard all the features from Tab. II that are not of our interest, i.e., features 3, 6-11. Then, we convert all tweets to lowercase, remove punctuation, tokenize, remove stop words and lemmatize with the same methods as above. We do not have to remove the substrings "@¡name¿", since the data here is not anonymized. To classify the unlabeled tweets, we need to bring them into a vectorized form as well but we need to keep in mind that we operate on the same dictionary W as the labeled tweets above. So we enlarge the dictionary used for the labeled tweets by adding the new words that appear and zeros to the columns of the vectors x_d from before to encompass these new words also in the labeled tweets.

$$\text{unlabeled data} = \{(\mathbf{x}_m)\}_{m=1}^{M}, \quad \mathbf{x}_m \in \mathbb{N}^W$$

E. Preprocessing for LDA

The preprocessing for the topic modeling with LDA is very similar to the preprocessing for the classification, as it also contains tokenization and removing stop words. We use the TTLocVis [33], which provides a broad range of methods to clean and analyze the contents of Twitter data. In particular, TTLocVis allows to apply LDA Topic Models on extremely sparse Twitter data. Tweets are pooled by hashtags [33]. The pooling procedure groups the tweets into larger text documents if they have one or more hashtags in common. We pool with respect to the hashtags as proposed by [41]. The LDA model implementation in TTLocVis is based on Gensim [46]. Note that for the description of the LDA model, we use the notation for individual tweets, even though we estimate the topics on pools.

IV. METHODS

We first train the model (Fig. 1Ⓐ) on the labelled dataset [54]. We then classify the tweets (Fig. 1Ⓑ), pool them, estimate the topics of the pools (Fig. 1Ⓒ), and finally estimate the GAMs on the topics of the tweets (Fig. 1Ⓓ), by assigning each tweet the topic of its respective pool. We also present the GAM plots for other variables, such as time and sentiment. We work with a balanced data set of 35,468 positively and 36,988 negatively classified tweets. Note that a tweet can be contained in multiple pools. We train the LDA topic model on the pools. However, we can increase the evaluation by dissolving the pools into their tweets, giving each tweet of a pool its respective topic distribution. This has two important consequences. First, we can now take the publication date of tweets into our analysis, and second, the predicted effects are much more precise since we do not have the problem of sparse data. Moreover, we can even consider the "size" of the pools. Larger pools result in more tweets and, therefore, a better prediction of the privacy of the content of an individual tweet. Note also that we do not need to consider the privacy label when pooling since we immediately dissolve the pools again. The pooling mainly improves the topic modeling.

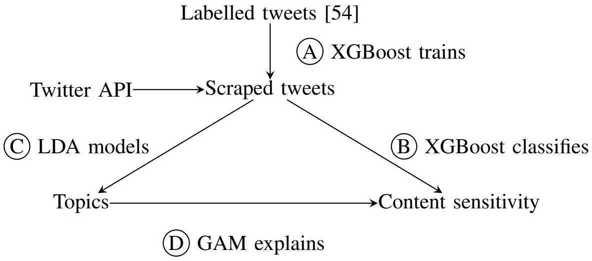


Fig. 1. Pipeline of our proposed framework.

A. Classifying Tweets with XGBoost

We use XGBoost to classify tweets as [1:Private], if the corresponding predicted privacy probability is above a certain threshold C_1 and as [2:Non-private] if it is below C_2 [18].

B. Topic modeling with LDA

We use the smoothed version of the topic modeling approach using LDA [15]. LDA is a generative probabilistic topic model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

C. Estimating Non-linear Effects with GAMs

GAMs are proposed in Hastie, and Tibshirani [28]. Although attractively simple, traditional linear regression often fails in practical situations because real-life effects are usually non-linear [27].

D. Sentiment Analysis

For the sentiment analysis, we use Vader [30]. Vader takes in a string and returns a score in each of four categories, as shown in Tab. III.

1.	negative
2.	neutral
3.	positive
4.	compound (computed by normalizing the scores above)

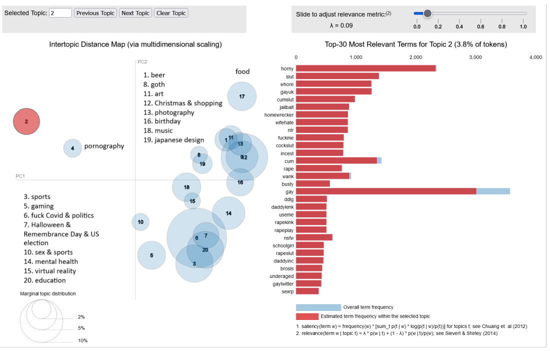
TABLE III
WE USE THE COMPOUND SCORE, RANGING FROM -1 (NEGATIVE SENTIMENT) TO 1 (POSITIVE SENTIMENT).

V. RESULTS

In this section, we first describe how to interpret the outcome of our introduced framework (LDA and GAM plots), before we discuss some of the obtained results thoroughly.

a) LDA Visualizations and GAM interpretations: Fig. 2 shows the annotated topics. The differently sized circles represent LDA topics, while their size reflects the number of total words from the dictionary that are being used by the specific topics (excluding words with an inner-topic prevalence of zero). The visual representation itself relies on a *Principal Component Analysis* (PCA) of the topic distributions, and the distances between the topic bubbles are calculated by the Jenson-Shannon distance measure, which compares the similarity of probability distributions [37]. The words shown on the right side of the plot are the most important words for a selected topic. These topic-defining words are ranked high when a small λ value is selected. A low λ gives more weight to a term's lift, defined as "the ratio of a term's probability within a topic to its marginal probability across the corpus" [49]. By contrast, a larger λ increases the weight of the topic-specific probability of a word in the ranking. Details on the metric can be found in [49].

We further explain the interpretation of GAM plots. For instance, Fig. 3 shows partial effect plots. They show the component effect of each of the smooth or linear terms in the model, which adds to the overall prediction. We show the data alongside the model predictions, the X-values along the bottom, and residuals on the plots. Partial residuals are the difference between the partial effect and the data after all other partial effects have been accounted for. The blue lines show the standard errors on our plots. These show the 95% confidence interval for the mean shape of the effect, which is marked by the red line.


Selected Topic: 2

Slide to adjust relevance metric:(2)
λ = 0.09

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 2. 20 identified topics from LDA visualization of individual tweets: Topic 2 (i.e., "sexuality & masturbation & rape") and 4 (i.e., "pornography") can be particularly addressed with privacy relevant information. Note that we did not consider some trending words within the crawling timeline like "autumn", and "US election" for GAM modelings.

The complete list of LDA visualizations and the GAM plots are also available to the reader (https://bit.ly/3tcOIfq). A deeper understanding of the topics can be obtained by adjusting the parameter λ in the LDA visualizations. Hereafter, we discuss some of the obtained results.

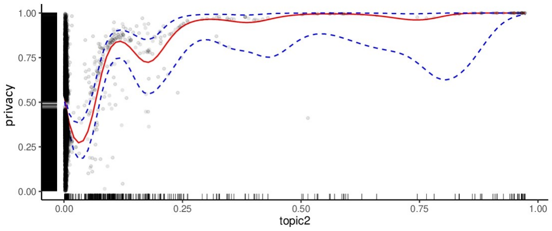


Fig. 3. Strong positive non-linear effect of the tweets from topic 2 (i.e., sexuality, masturbation, and rape x-axis) on privacy (y-axis) already at a small prevalence for this topic.

b) Obtained Results: The analysis of the LDA topic visualization in Fig. 2 shows content-sensitive topics, i.e., topics 2-7, 10, 14-15, and 17-19 ("sexuality & masturbation & rape, pornography, sports, gaming, fuck Covid & politics, Halloween & Remembrance Day & US election, sex & sports, mental health, virtual reality, food, music, Japanese design"). We further explore the above topics using GAMs.

Very Sensitive Topics. Topics 2 (i.e., "the sexuality & masturbation & rape topic") and 4 (i.e., "the pornography" topic) show a clear positive effect on the privacy of the tweets (see Fig. 3 and 4).

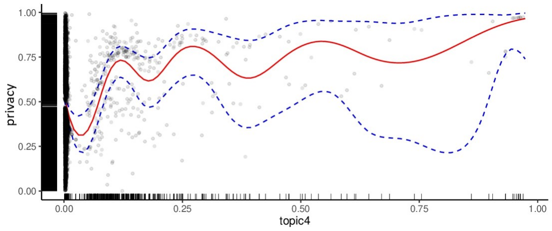


Fig. 4. Strong positive non-linear effect of the tweets from topic 4, (i.e., pornography x-axis) on privacy (y-axis).

Nighttime Tweets. In our findings, the publication date of a tweet is a valuable factor in determining content sensitivity, as shown in Fig. 5. The mean estimate for privacy is slightly higher for nighttime, although at a lower confidence interval, merely due to fewer tweets posted at night.

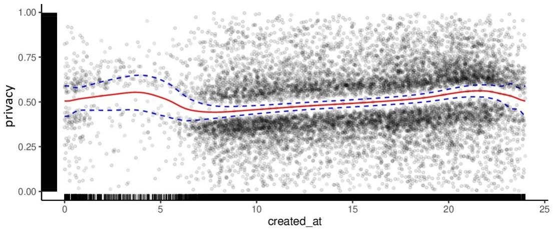


Fig. 5. Tweet sensitivity and time of the posted tweet.

Negative and Positive Sentiment. The sentiment of the individual tweets exhibits the effect that the more positive the sentiment gets, the less likely it is for a tweet to be classified as private and the more that sentiment is negative, the tweet is more likely to be labeled private (Fig. 6).

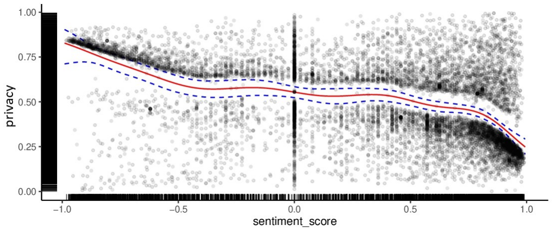


Fig. 6. Tweet sensitivity and sentiment negativity.

The same shape of both the mean estimate for the publication date and the sentiment can now be found in the three-dimensional plot with topic 2 (i.e., "sexuality & masturbation & rape") in Fig. 7 and sentiment in Fig. 8. However, the more dominant topic 2 becomes, the less impact the sentiment has on the privacy estimate, everything else being equal.

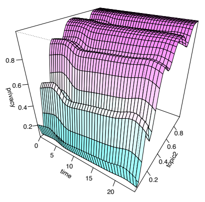


Fig. 7. Non-linear effects of time and topic 2 ("sexuality & masturbation & rape") on privacy (y-axis)

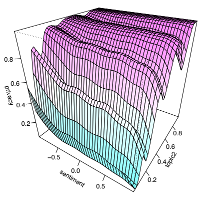


Fig. 8. Non-linear effects of sentiment and topic 2 ("sexuality & masturbation & rape") on privacy (y-axis)

Mental Health Topic. The 14^{th} topic regarded to mental health does not show a strong positive effect on the privacy score (see Fig. 9).

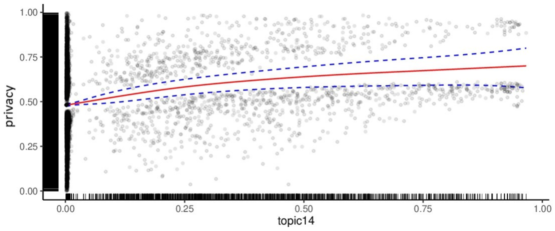


Fig. 9. Non-linear effect of topic 14 ("mental health") on privacy (y-axis)

In a three-dimensional analysis plotted with the sentiment score, it reaches a lower probability (0%-70%) for a tweet to be private given that it contains topic 14, in contrast to a tweet being private when only considering its sentiment (30%-80%). Regardless, a tweet is slightly more likely to be private if it contains topic 14, everything else being equal.

Trending Sensitive Topic. Topic 6, i.e., "fuck covid & politics," however, does show a very strong positive effect on privacy in this approach, as shown in Fig. 10.

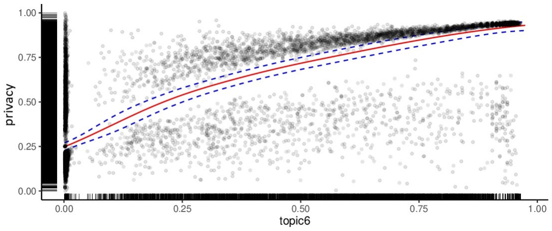


Fig. 10. Covid-related sensitive topic ("fuck covid & politics").

The same holds for topic 20 about education in Fig. 11, which also contains topics' terms, such as "antibullyingweek" and "race."

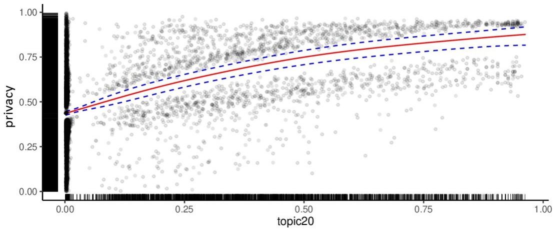


Fig. 11. Non-linear effect of topic 20 (education) on privacy (y-axis)

VI. Limitations and Future Work

a) Unlabeled data set: We crawled and subsequently predicted the sensitivity of a set of tweets indirectly, i.e., we trained it based on a labeled data set [54] by the users and used the trained models to predict the sensitivity of other crawled tweets without having their labels. However, ML models are probabilistic and thus prone to minor false predictions. Having more annotated user data would enrich our approach.

b) Cultural and individual differences: Our findings are limited to a single geo-location only. Our proposed framework can be further extended in various other locations. In a future work, in particular, the cultural differences in users' privacy perceptions remain to be investigated. Moreover, taking into account differences between individuals' privacy perceptions remain also for further considerations.

VII. Conclusion

We propose the integration of XGBoost, LDA topic model, and GAMs as a novel framework to model the complex non-linear relationship between the privacy of posts on social media based on their content and additional structural variables.

We show that using GAMs allows for a detailed and precise evaluation of non-linear effects of one or two variables at a time and enables us to integrate topics and structure in the model. We identify one private topic, "sexuality," and some candidates, such as "mental health" and "politics," in our sample. Moreover, we find that tweets that explicitly stated a frustration towards Covid and politics (precisely detected as "fuck covid & politics") strongly correlate to being of sensitive content. Interesting insights are also gained by a sentiment analysis of tweets and the time of creation, i.e, tweets of more negative sentiment or posted at nighttime are more likely to be of sensitive content. This suggests taking further structural variables into account.

Acknowledgment

This work is partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) referenced with #317687129 and Campus Institute Data Science (CIDAS).

References

[1] A. P. Raposo, H. J. Weber, D. E. Alvarez–Castillo, M. Kirchbach, Cent. Eur. J. Phys. 5, 253 (2007)

[2] Mao, Huina, Xin Shuai and Apu Kapadia. Loose Tweets: An Analysis of Privacy Leaks on Twitter. In Workshop on Privacy in the Electronic Society (WPES), 2011.

[3] Aylin Caliskan-Islam, Jonathan Walsh, and Rachel Greenstadt. Privacy Detective: Detecting Private Information and Collective Privacy Behavior in a Large Social Network. In Proc. of the Workshop on Privacy in the Electronic Society (WPES), 2014.

[4] Lu Zhou, Wenbo Wang, and Keke Chen. tweet Properly: Analyzing Deleted Tweets to Understand and Identify Regrettable Ones. In Proc. of the 25th International Conference on World Wide Web (WWW), 2016.

[5] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. I Regretted the Minute I Pressed Share: A Qualitative Study of Regrets on Facebook. In Proc. of the Seventh Symposium on Usable Privacy and Security (SOUPS), 2011.

[6] Q. Wang, J. Bhandal, S. Huang and B. Luo. Classification of Private Tweets Using Tweet Content. In IEEE 11th International Conference on Semantic Computing (ICSC), 2017.

[7] Q. Wang, J. Bhandal, S. Huang, and B. Luo. Content-Based Classification of Sensitive Tweets. International Journal of Semantic Computing, 2017.

[8] Mittal, Manasi and Asghar, Muhammad Rizwan and Tripathi, Arvind. Do My Emotions Influence What I Share? Analysing the Effects of Emotions on Privacy Leakage in Twitter. International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2020.

[9] Lindrit Kqiku, Marvin Kühn, Delphine Reinhardt. From Sentiment to Sensitivity: The Role of Emotions on Privacy Exposure in Twitter. In: Proc. of the 2nd ACM Workshop on Open Challenges in Online Social Networks (OASIS, HT workshop), 2022.

[10] S. M. Mohammad and F. Bravo-Marquez, WASSA-2017 Shared Task On Emotion Intensity. In arXiv preprint arXiv:1708.03700, 2017.

[11] Khaled Almotairi and Bilal Bataineh. Perception of Information Sensitivity for Internet Users in Saudi Arabia. *Acta Informatica Pragensia*, 9(2):184–199, 2020.

[12] Susan B. Barnes. A Privacy Paradox: Social Networking in the United States.. *First Monday*, 11(9), 2006.

[13] Daniel Berg. Bankruptcy Prediction by Generalized Additive Models. *Applied Stochastic Models in Business and Industry*, 23(2):129–143, 2007.

[14] Edward Loper Bird, Steven and Ewan Klein. Natural Language Processing With Python. 2009.

[15] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, 2003.

[16] Leo Breiman and Jerome H. Friedman. Estimating Optimal Transformations for Multiple Regression And Correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.

[17] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and Regression Trees*. Routledge, 2017.

[18] Tianqi Chen and Carlos Guestrin. XgBoost: A Scalable Tree Boosting System. In *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[19] Mark S Clements, Bruce K Armstrong, and Suresh H Moolgavkar. Lung Cancer Rate Predictions Using Generalized Additive Models. *Biostatistics*, 6(4):576–589, 2005.

[20] Dr. John Dawes. Do Data Characteristics Change According to the Number of Scale Points Used? An Experiment Using 5-Point, 7-Point and 10-Point Scales. *International Journal of Market Research*, 50(1):61–104, 2008.

[21] Koen W. De Bock, Kristof Coussement, and Dirk Van den Poel. Ensemble Classification Based on Generalized Additive Models. *Computational Statistics & Data Analysis*, 54(6):1535–1546, 2010.

[22] Cynthia Dwork. Differential Privacy: A Survey of Results. In Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li, editors, *Theory and Applications of Models of Computation*, 2008.

[23] Lujun Fang and Kristen LeFevre. Privacy Wizards for Social Networking Sites. 2010.

[24] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Special Invited Paper. Additive Logistic Regression: A Statistical View of Boosting. *The Annals of Statistics*, 28(2):337–374, 2000.

[25] Mark Girolami. On an Equivalence Between PLSI and LDA. 2003.

[26] Thomas Griths and Mark Steyvers. A Probabilistic Approach to Semantic Representation. *In Proc. of the 24th conference of the Cognitive Science Society*, 2004.

[27] Trevor Hastie and R. Tibshirani. *Generalized Additive Models*. John Wiley & Sons, Ltd, 2014.

[28] Trevor J Hastie and Robert J Tibshirani. *Generalized Additive Models*. Routledge, 2017.

[29] Jianming He, Wesley Chu, and Zhenyu Liu. Inferring Privacy Information From Social Networks. volume 3975, 2006.

[30] C. Hutto and Eric Gilbert. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis Of Social Media Text. *Proc. of the International AAAI Conference on Web and Social Media*, 8(1):216–225, 2014.

[31] Maritza Johnson, Serge Egelman, and Steven M. Bellovin. Facebook and Privacy: It's Complicated.

[32] Yanfei Kang, Feng Li, Rob J Hyndman, Mitchell O'Hara-Wild, and Bocong Zhao. gratis: GeneRAting TIme Series With Diverse and Controllable Characteristics, 2020.

[33] Gillian Kant, Christoph Weisser, and Benjamin Säfken. TTLocVis: A Twitter Topic Location Visualization Package. *Journal of Open Source Software*, 5(54):2507, 2020.

[34] M Kawakita, Mihoko Minami, S Eguchi, and CE Lennert-Cody. An Introduction to the Predictive Technique AdaBoost With A Comparison to Generalized Additive Models. *Fisheries research*, 76(3):328–343, 2005.

[35] Patrick Kelley, Robin Brewer, Yael Mayer, Lorrie Cranor, and Norman Sadeh. An Investigation Into Facebook Friend Grouping. volume 6948, pages 216–233, 2011.

[36] L. Sweeney. K-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.

[37] Lillian Lee. Measures of Distributional Similarity. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[38] Lee Humphreys, Phillipa Gill, and Balachander Krishnamurthy. *How Much Is Too Much? Privacy Issues on Twitter*. 2012.

[39] Bo Luo and Dongwon Lee. On Protecting Private Information in Social Networks: A Proposal. In *Proc. - 25th IEEE International Conference on Data Engineering (ICDE), 2009*.

[40] Ereni Markos, George R. Milne, and James W. Peltier. Information Sensitivity and Willingness to Provide Continua: A Comparative Privacy Study of the United States and Brazil. *Journal of Public Policy & Marketing*, 36(1):79–96, 2017.

[41] Rishabh Mehrotra, Scott P Sanner, Wray Lindsay Buntine, and Lexing Xie. Improving LDA Topic Models for Microblogs Via Pooling And Automatic Labeling. In Diane Kelly, Maarten de Rijke, and Tetsuya Sakai, editors, *Proc. of the 36th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR)*, 2013.

[42] Sai Teja Peddinti, Keith W. Ross, and Justin Cappos. "On the Internet, Nobody Knows You're a Dog": A Twitter Case Study of Anonymity in Social Networks. In *Proc. Of the Second ACM Conference on Online Social Networks (COSN)*, 2004.

[43] Alan Mislove, Bimal Viswanath, Krishna P. Gummadi, and Peter Druschel. You Are Who You Know: Inferring User Profiles in Online Social Networks. In *Proc. of the Third ACM International Conference on Web Search and Data Mining (WSDM)*, 2010.

[44] H. D. PATTERSON and R. THOMPSON. Recovery of Inter-Block Information When Block Sizes Are Unequal. *Biometrika*, 58(3):545–554, 1971.

[45] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-Learn: Machine Learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[46] Radim Rehurek and Petr Sojka. Gensim - Python Framework for Vector Space Modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2), 2011.

[47] Delphine Reinhardt, Franziska Engelmann, and Matthias Hollick. Can I Help You Setting Your Privacy? A Survey-Based Exploration Of Users' Attitudes Towards Privacy Suggestions. *Proc. of the 13th International Conference on Advances in Mobile Computing and Multimedia*, 2015.

[48] Eva-Maria Schomakers, Chantal Lidynia, Dirk Müllmann, and Martina Ziefle. Internet Users' Perceptions of Information Sensitivity – Insights From Germany. *International Journal of Information Management*, 46:142–150, 2019.

[49] Carson Sievert and Kenneth Shirley. LDAvis: A Method for Visualizing and Interpreting Topics. In *Proc. of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014.

[50] Manya Sleeper, Justin Cranshaw, Patrick Gage Kelley, Blase Ur, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. "I Read My Twitter the Next Morning and Was Astonished". *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, 2013.

[51] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 2013.

[52] Statista. Daily Social Media Usage Worldwide — Statista, 11/15/2022. Link: https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

[53] Internet Live Stats. Twitter Usage Statistics, 11/15/2022. Link: https://www.internetlivestats.com/twitter-statistics/

[54] Qiaozhi Wang, Hao Xue, Fengjun Li, Dongwon Lee, and Bo Luo. #Donttweetthis: Scoring Private Information in Social Networks. *Proc. on Privacy Enhancing Technologies*, 2019(4):72–92, 2019.

[55] Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. " I Regretted the Minute I Pressed Share": A Qualitative Study Of Regrets on Facebook. In *Proc. of the Seventh Symposium on Usable Privacy and Security*, SOUPS, 2011.

[56] Jason Wiese, Patrick Gage Kelley, Lorrie Faith Cranor, Laura Dabbish, Jason I. Hong, and John Zimmerman. Are You Close With Me? Are You Nearby? Investigating Social Groups, Closeness, and Willingness to Share. In *Proc. of the 13th International Conference on Ubiquitous Computing (UbiComp)*, 2011.

[57] S N Wood. Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models, In 3rd edn. Philadelphia: Society for Industrial and Applied Mathematics, 2011.

[58] Simon N Wood. Stable and Efficient Multiple Smoothing Parameter Estimation For Generalized Additive Models. *Journal of the American Statistical Association*, 99(467):673–686, 2004.

[59] Yuhao Yang, Jonathan Lutes, Fengjun Li, Bo Luo, and Peng Liu. Stalking Online. In Elisa Bertino and Ravi Sandhu, editors, *Proc. of the second ACM conference on Data and Application Security and Privacy (CODASKY)*, 2012.

[60] Z. Cai, Zaobo He, Xin Guan, and Yingshu Li. Collective Data-Sanitization for Preventing Sensitive Information Inference Attacks in Social Networks. *in IEEE Transactions on Dependable and Secure Computing, vol. 15, no. 4*, 2018.

[61] José Alemany, Elena del Val, Ana García-Fornes Empowering Users Regarding the Sensitivity of their Data in Social Networks through Nudge Mechanisms. *53rd Hawaii International Conference on System Sciences (HICSS)*, 2020.

[62] Mohsen Minaei, Mainack Mondal, Patrick Loiseau, & Krishna P. Gummadi & Aniket Kate Lethe: Conceal Content Deletion from Persistent Observers. *Proc. on Privacy Enhancing Technologies*, 2019.

[63] Mohsen Minaei, Mouli S Chandra, Mondal Mainack, Ribeiro Bruno and Aniket Kate. Deceptive Deletions for Protecting Withdrawn Posts

on Social Media Platforms. *The Network and Distributed System Security Symposium (NDSS)*, 2021.

[64] Kqiku, Lindrit, Jakob Dieterle, and Delphine Reinhardt. Exploration of a Mobile Design for a Privacy Assistant to Help Users in Sharing Content in Online Social Networks. *Proc. of the Eighth Usable Security und Privacy Workshop (MuC workshop)*, 2022.

[65] Twitter. About Twitter Circle, 11/15/2022, Link: https://help.twitter.com/en/using-twitter/twitter-circle.