Classification-based Content Sensitivity Analysis (DISCUSSION PAPER)

Elena Battaglia, Livio Bioglio, and Ruggero G. Pensa

University of Turin, Dept. of Computer Science, Turin, Italy
{elena.battaglia,livio.bioglio,ruggero.pensa}@unito.it

Abstract. With the availability of user-generated content in the Web, malicious users have access to huge repositories of private (and often sensitive) information regarding a large part of the world's population. In this paper, we propose a way to evaluate the harmfulness of text content by defining a new data mining task called *content sensitivity analysis*. According to our definition, a score can be assigned to any text sample according to its degree of sensitivity. Even though the task is similar to sentiment analysis, we show that it has its own peculiarities and may lead to a new branch of research. Thanks to some preliminary experiments, we show that content sensitivity analysis can not be addressed as a simple binary classification task.

CEUR-WS.org/Vol-2646/12-paper.pdf

1 Introduction

Internet privacy has gained much attention in the last decade due to the success of online social networks and other social media services that expose our lives to the wide public. Consequently, understanding and measuring the exposure of user privacy in the Web has become crucial [6] and many different metrics and methods have been proposed with the goal of assessing the risk of privacy leakage in posting activities [1]. Most research efforts, however, focus on measuring the overall exposure of users according to their privacy settings [12] or position within the network [11]. However, in addition to personal and behavioral data collected more or less legitimately by companies and organizations, many websites and mobile/web applications store and publish tons of user-generated content, which, very often, capture and represent private moments of our life. The availability of user-generated content is a huge source of relatively easy-to-access private (and often very sensitive) information concerning habits, preferences, families and friends, hobbies, health and philosophy of life, which expose the authors of such contents (or any other individual referenced by them) to many (cyber)criminal risks, including identity theft, stalking,

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). This volume is published and copyrighted by its editors. SEBD 2020, June 21-24, 2020, Villasimius, Italy.

burglary, frauds, cyberbullying or "simply" discrimination in workplace or in life in general. Sometimes users are not aware of the dangers due to the uncontrolled diffusion of their sensitive information and would probably avoid publishing it if only someone told them how harmful it could be.

In this discussion paper, we address this problem by proposing a way to assess the sensitivity of user-generated content. To this purpose, we define a new data mining task that we call *content sensitivity analysis* (CSA), inspired by sentiment analysis [7]. The goal of CSA is to assign a score to any text sample according to the amount of sensitive information it potentially discloses. The problem of private content analysis has already been investigated as a way to characterize anonymous vs. non anonymous content posting in specific social media [4, 9] or question-and-answer platforms [8]. However, the link between anonymity and sensitive contents is not that obvious: users may post anonymously because, for instance, they are referring to illegal matters (e.g., software/steaming piracy, black market and so on); conversely, fully identifiable persons may post very sensitive contents simply because they are underestimating the visibility of their action [12, 11]. Although CSA has some points in common with anonymous content analysis and the well-known sentiment analysis task, we show that it has its own peculiarities and may lead to a brand new branch of research, opening many intriguing challenges in several computer science and linguistics fields.

Through some preliminary but extensive experiments on a large annotated corpus of social media posts, we show that content sensitivity analysis can not be addressed straightforwardly. In particular, we design a simplified CSA task leveraging binary classification to distinguish between sensitive and non sensitive posts by testing several bag-of-words and word embedding models. According to our experiments, the classification performances achieved by the most accurate models are far from being satisfactory. This suggests that content sensitivity analysis should consider more complex linguistic and semantic aspects, as well as more sophisticated machine learning models. A more in-depth discussion on how to address these issues is reported in the full version of this paper [2].

2 Content Sensitivity Analysis

In this section, we introduce the new data mining task that we call *content sensitivity analysis* (CSA), aimed at determining the amount of privacy-sensitive content expressed in user-generated text content. We distinguish two cases, namely *binary CSA* and *continuous CSA*, according to the outcome of the analysis (binary or continuous). Before introducing the technical details of CSA, we briefly provide the intuition behind CSA by describing a motivating example.

2.1 Motivating example

To explain the main objectives of CSA and the scientific challenges associated to them, we consider the post given as an example in Figure 1. This particular post discloses information about the author and his friend Alice Green. Moreover, the

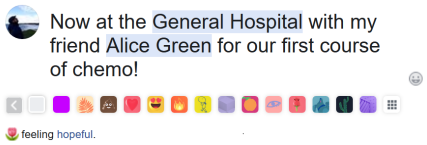

Fig. 1. An example of a potentially privacy-sensitive post.

post contains spatiotemporal references ("now" and "General Hospital"), which are generally considered intrinsically sensitive, and mentions "chemo", a potentially sensitive term. Finally, the sentence is related to "cancer", a potentially sensitive topic, and its structure suggests that the two subjects of disclosure have cancer and they are both about to start their first course of chemotherapy.

It is clear that, reducing sensitivity to anonymity, as done in previous research work [8, 4], is only one side of the coin. Instead, CSA has much more in common with the famous *sentiment analysis* (SA) task, where the objective is to measure the "polarity" or "sentiment" of a given text [7, 5]. However, while SA has already a well-established theory and may count on a set of easy-to-access and easy-to-use tools, CSA has never been defined before. Therefore, apart from the known open problems in SA (such as sarcasm detection), CSA involves three new scientific challenges.

1. **Definition of sensitivity**. A clear definition of sensitivity is required. Sensitivity is often defined in the legal systems, such as in the EU General Data Protection Regulation (GDPR), as a characteristic of some personal data (e.g., criminal or medical records), but a cognitive and perceptive explanation of what can be defined as "sensitive" is still missing [13].

2. **Sensitivity-annotated corpora**. Large text corpora need to be annotated according to sensitivity and at multiple levels: at the sentence level ("I got cancer" is more sensitive than "I got some nice volleyball shorts"), at the topic level ("health" is more sensitive than "sports") and at the term level ("cancer" is more sensitive than "shorts").

3. **Context-aware sensitivity**. Due to its subjectivity, a clear evaluation of the context is needed. The fact that a medical doctor talks about cancer is not sensitive per se, but if she talks about some of her patients having cancer, she could disclose very sensitive information.

2.2 Definitions

Here, we provide the details regarding the formal framework of *content sensitivity analysis*. We will propose a definition of "sensitivity" further in this section. The simplest way to define CSA is as follows:

Definition 1 (binary content sensitivity analysis). *Given a user-generated text object $o_i \in \mathcal{O}$, with $\mathcal{O}$ being the domain of all user-generated contents, the*

binary content sensitivity analysis *task consists in designing a function* $f_s : \mathcal{O} \rightarrow \{sens, ns\}$*, such that* $f_s(o_i) = sens$ *iff* o_i *is privacy-sensitive,* $f_s(o_i) = ns$ *iff* o_i *is not sensitive.*

In some cases, sensitivity is not the same for all sensitive objects: a post dealing with health is certainly more sensitive than a post dealing with vacations, although both can be considered as sensitive. This suggests that, instead of considering sensitivity as a binary feature of a text, a more appropriate definition of CSA should take into account different degrees of sensitivity, as follows:

Definition 2 (continuous content sensitivity analysis). *Let* $o_i \in \mathcal{O}$ *be a user-generated object, with* $\mathcal{O}$ *being the domain of all user-generated contents. The* continuous content sensitivity analysis *task consists in designing a function* $f_s : \mathcal{O} \rightarrow [-1, 1]$*, such that* $f_s(o_i) = 1$ *iff* o_i *is maximally privacy-sensitive,* $f_s(o_i) = -1$ *iff* o_i *is minimally privacy-sensitive,* $f_s(o_i) = 0$ *iff* o_i *has unknown sensitivity. The value* $\sigma_i = f_s(o_i)$ *is the* ***sensitivity score*** *of object* o_i*.*

According to this definition, sensitive objects have $0 < \sigma \leq 1$, while non sensitive posts have $-1 \leq \sigma < 0$. In general, when $\sigma \approx 0$ the sensitivity of an object cannot be assessed confidently. Of course, by setting appropriate thresholds, a continuous CSA can be easily turned into a binary CSA task.

At this point, a congruent definition of "sensitivity" is required to set up the task correctly. Although different characterizations of privacy-sensitivity exist, there is no consistent and uniform theory [13]; so, in this work, we consider a more generic, flexible and application-driven definition of privacy-sensitive content.

Definition 3 (privacy-sensitive content). *A generic user-generated content object is privacy-sensitive if it makes* ***the majority of users*** *feel uncomfortable in writing or reading it because it may reveal some aspects of their own or others' private life to unintended people.*

Notice that "uncomfortableness" should not be guided by some moral or ethical judgement about the disclosed fact, but uniquely by its harmfulness towards privacy. Such a definition allows the adoption of the "wisdom of the crowd" principle in contexts where providing an objective definition of what is sensitive (and what is not sensitive) is particularly hard. Moreover, it has also an intuitive justification. Different social media may have different meaning of sensitivity. For instance, in a professional social networking site, revealing details about one's own job is not only tolerated, but also encouraged, while one may want to hide detailed information about her professional life in a generic photo-video sharing platform. Similarly, in a closed message board (or group), one may decide to disclose more private information than in open ones. Sensitivity towards certain topics also varies from country to country. As a consequence, function f_s can be learnt according to an annotated corpus of content objects as follows.

Definition 4 (sensitivity function learning). *Let* $O = \{(o_i, \sigma_i)\}_{i=1}^{N}$ *be a set of* N *annotated objects* $o_i \in \mathcal{O}$ *with the related sensitivity score* $\sigma_i \in [-1, 1]$*. The goal of a sensitivity function learning algorithm is to search for a function* $f_s : \mathcal{O} \rightarrow [-1, 1]$*, such that* $\sum_{i=1}^{N} \left(f_s(o_i) - \sigma_i\right)^2$ *is minimum.*

The simplest way to address this problem is by setting a regression (or classification, in the case of binary CSA) task. However, we will show in Section 3 that such an approach is unable to capture the actual manifold of sensitivity accurately.

3 Preliminary experiments

In this section, we report the results of some preliminary experiments aimed at showing the feasibility of content sensitivity analysis together with its difficulties. The experiments are conducted under the binary CSA framework (see Definition 1 in Section 2). We set up a binary classification task to distinguish whether a given input text is privacy-sensitive or not. Before presenting the results, in the following, we first introduce the data, then we provide the details of our experimental protocol.

3.1 Annotated corpus

Since all previous attempts of identifying sensitive text have leveraged user anonymity as a discriminant for sensitive content [8, 4], there is no reliable annotated corpus that we can use as benchmark. Hence, we construct our own dataset by leveraging a crowdsourcing experiment. We use one of the datasets described in [3], consisting of 9917 anonymized social media posts, mostly written in English, with a minimum length of 2 characters and a maximum length of 435 (the average length is 80). Thus, they well represent typical social media short posts. On the other hand, they are not annotated for the specific purpose of our experiment and, because of their shortness, they are also very difficult to analyze. Consequently, after discarding all useless posts (mostly uncomprehensible ones) we have set up a crowdsourcing experiment by using a Telegram bot that, for each post, asks whether it is sensitive or not. As third option, it was also possible to select "unable to decide". We collected the annotations of 829 posts from 14 distinct annotators. For each annotated post, we retain the most frequently chosen annotation. Overall, 449 posts where tagged as non sensitive, 230 as sensitive, 150 as undecidable. Thus, the final dataset consists of 679 posts of the first two categories (we discarded all 150 undecidable posts).

3.2 Datasets

We consider two distinct document representations for the dataset, a bag-of-words and four word vector models. To obtain the bag-of-word representation we perform the following steps. First, we remove all punctuation characters of terms contained in the input posts as well as short terms (less than two characters) and terms containing digits. Then, we build the bag-of-words model with all remaining 2584 terms weighted by their *tfidf* score. Differently from classic text mining approaches, we deliberately exclude lemmatization, stemming and stop word removal from text preprocessing, since those common steps would affect content

sensitivity analysis negatively. Indeed, inflections (removed by lemmatization and stemming) and stop words (like "me", "myself") are important to decide whether a sentence reproduces some personal thoughts or private action/status. Hereinafter, the bag-of-words representation is referred to as *BW2584*.

The word vector representation, instead, is built using word vectors pre-trained with two billion tweets (corresponding to 42 billion tokens) using the *GloVe* (Global Vector) model [10]. In detail, we use three representation, here called *WV25*, *WV50* and *WV100* with, respectively, 25, 50 and 100 dimensions. Additionally, we build an ensemble by considering the concatenation of the three vector spaces. The latter representation is named *WVEns*. Finally, from all five datasets we removed all posts having an empty bag-of-words or word vector representation. Such preprocessing step further reduces the size of the dataset down to 611 posts (221 sensitive and 390 non sensitive), but allows for a fair performance comparison.

3.3 Experimental settings

Each dataset obtained as described beforehand is given in input to a set of six classifiers. In details, we use k-NN, decision tree (DT), Multi-layer Perceptron (MLP), SVM, Random Forest (RF), and Gradient Boosted trees (GBT). We do not execute any systematic parameter selection procedure since our main goal is not to compare the performances of classifiers, but, rather, to show the overall level of accuracy that can be achieved in a binary content sensitivity analysis task. Hence, we use the following default parameter for each classifier.

- k**NN**: we set $k = 3$ in all experiments;
- **DT**: for all datasets, we use C4.5 with Gini Index as split criterion, allowing a minimum of two records per node and minimum description length as pruning strategy;
- **MLP**: we train a shallow neural network with one hidden layer; the number of neurons of the hidden layer is 30 for the bag-of-words representation and 20 for all word vector representations;
- **SVM**: for all datasets, we use the polynomial kernel with default parameters;
- **RF**: we train 100 models with Gini index as splitting criterion in all experiments;
- **GBT**: for all datasets, we use 100 models with 0.1 as learning rate and 4 as maximum tree depth.

All experiments are conducted by performing ten-fold cross-validation, using, for each iteration, nine folds as training set and the remaining fold as test set.

3.4 Results and discussion

The summary of the results, in terms of average F1-score, are reported in Table 1. It is worth noting that the scores are, in general, very low (between 0.5826, obtained by the neural network on the bag-of-words model, and 0.6858, obtained

Table 1. Classification in terms of average F1-score for different post representations.

Dataset	Type	kNN	DT	MLP	SVM	RF	GBT
BW2584	bag-of-words	0.6579	0.6743	0.5826	0.6481	0.6776	0.6678
WV25	word vector	0.6203	0.6317	0.6497	0.6383	0.6628	0.6268
WV50	word vector	0.6121	0.6105	0.6530	0.6448	**0.6858**	0.6399
WV100	word vector	0.6367	0.6088	0.6497	0.6563	0.6694	0.6497
WVEns	word vector	0.6432	0.5859	0.6481	0.6547	0.6628	0.6416

by Random Forest on the word vector representation with 50 dimensions). Of course, these results are biased by the fact that data are moderately unbalanced (64% of posts fall in the non-sensible class). However they are not completely negative, meaning that there is space for improvement. We observe that the winning model-classifier pair (50-dimensional word vector processed with Random Forest) exhibits high recall on the non-sensitive class (0.928) and rather similar results in terms of precision for the two classes (0.671 and 0.688 for the sensitive and non-sensitive classes respectively). The real negative result is the low recall on the sensitive class (only 0.258), due to the high number of false negatives. We recall that the number of annotated sensitive posts is only 221, i.e., the number of examples is not sufficiently large for training a prediction model accurately.

These results highlight the following issues and perspectives. First, negative (or not-so-positive) results are certainly due to the lack of annotated data (especially for the sensitive class). Sparsity is certainly a problem in our settings. Hence, a larger annotated corpus is needed, although this objective is not trivial. In fact, private posts are often difficult to obtain, because social media platforms (luckily, somehow) do not allow users to get them using their API. As a consequence, all previous attempts to guess the sensitivity of text or construct privacy dictionaries strongly leverage user anonymity in public post sharing activities [8, 4], or rely on focus groups and surveys [13]. Moreover, without a sufficiently large corpus, not even the application of otherwise successful deep learning techniques would produce valid results. Second, simple classifiers, even when applied to rather complex and rich representations, can not capture the manifold of privacy sensitivity accurately. So, more complex and heterogenous models should be considered. An accurate sensitivity content analysis tool should consider lexical, semantic as well as grammatical features. Topics are certainly important, but sentence construction and lexical choices are also fundamental. Therefore, reliable solutions would consist of a combination of computational linguistic techniques, machine learning algorithms and semantic analysis. Third, the success of picture and video sharing platforms (such as Instagram and TikTok), implies that any successful sensitivity content analysis tool should be able to cope with audiovisual contents and, in general, with multimodal/multimedia objects. Finally, provided that a taxonomy of privacy categories in everyday life exists (e.g., health, location, politics, religious belief, family, relationships, and so on) a more complex CSA setting might consider, for a given content object, the privacy sensitivity degree in each category.

4 Conclusions

In this paper, we have addressed the problem of determining whether a given text object is privacy-sensitive or not by defining the generic task of content sensitivity analysis (CSA). Although the task promises to be challenging, we have shown that it is not unfeasible by presenting a simplified formulation of CSA based on binary text classification. With some preliminary but extensive experiments, we have showed that, no matter the data representation, the accuracy of such classifiers can not be considered satisfactory. Thus, it is worth investigating more complex techniques borrowed from machine learning, computational linguistics and semantic analysis. Moreover, without a strong effort in building massive and reliable annotated corpora, the performances of any CSA tool would be barely sufficient, no matter the complexity of the learning model.

Acknowledgments This work is supported by Fondazione CRT (grant number 2019-0450).

References

1. Alemany, J., del Val Noguera, E., Alberola, J.M., García-Fornes, A.: Metrics for Privacy Assessment When Sharing Information in Online Social Networks. IEEE Access **7**, 143631–143645 (2019)
2. Battaglia, E., Bioglio, L., Pensa, R.G.: Towards content sensitivity analysis. In: Proceedings of IDA 2020. pp. 67–79 (2020)
3. Celli, F., Pianesi, F., Stillwell, D., Kosinski, M.: Workshop on Computational Personality Recognition: Shared Task. In: Proceedings of ICWSM 2013 (2013)
4. Correa, D., Silva, L.A., Mondal, M., Benevenuto, F., Gummadi, K.P.: The Many Shades of Anonymity: Characterizing Anonymous Social Media Content. In: Proceedings of ICWSM 2015. pp. 71–80 (2015)
5. Liu, B., Zhang, L.: A Survey of Opinion Mining and Sentiment Analysis. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 415–463. Springer (2012)
6. Oukemeni, S., Rifà-Pous, H., i Puig, J.M.M.: Privacy Analysis on Microblogging Online Social Networks: A Survey. ACM Comput. Surv. **52**(3), 60:1–60:36 (2019)
7. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval **2**(1-2), 1–135 (2007)
8. Peddinti, S.T., Korolova, A., Bursztein, E., Sampemane, G.: Cloak and Swagger: Understanding Data Sensitivity through the Lens of User Anonymity. In: Proceedings of IEEE SP 2014. pp. 493–508 (2014)
9. Peddinti, S.T., Ross, K.W., Cappos, J.: User Anonymity on Twitter. IEEE Security & Privacy **15**(3), 84–87 (2017)
10. Pennington, J., Socher, R., Manning, C.D.: Glove: Global Vectors for Word Representation. In: Proceedings of EMNLP 2014. pp. 1532–1543 (2014)
11. Pensa, R.G., di Blasi, G., Bioglio, L.: Network-aware privacy risk estimation in online social networks. Social Netw. Analys. Mining **9**(1), 15:1–15:15 (2019)
12. Pensa, R.G., Blasi, G.D.: A privacy self-assessment framework for online social networks. Expert Syst. Appl. **86**, 18–31 (2017)
13. Vasalou, A., Gill, A.J., Mazanderani, F., Papoutsi, C., Joinson, A.N.: Privacy dictionary: A new resource for the automated content analysis of privacy. JASIST **62**(11), 2095–2105 (2011)