




Towards Content Sensitivity Analysis

Elena Battaglia, Livio Bioglio, and Ruggero G. Pensa^(✉) 

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

Abstract. With the availability of user-generated content in the Web, malicious users dispose of huge repositories of private (and often sensitive) information regarding a large part of the world’s population. The self-disclosure of personal information, in the form of text, pictures and videos, exposes the authors of such contents (and not only them) to many criminal acts such as identity thefts, stalking, burglary, frauds, and so on. In this paper, we propose a way to evaluate the harmfulness of any form of content by defining a new data mining task called *content sensitivity analysis*. According to our definition, a score can be assigned to any object (text, picture, video...) 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.

Keywords: Privacy · Text mining · Text categorization

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. 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 in the form of text posts and comments, pictures and videos 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, 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 paper, we address exactly this problem by proposing a way to measure the degree of sensitivity of any type 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 [13]. The goal of CSA is to assign a score to any object (text, picture, video...) 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 [5, 15, 16] or question-and-answer platforms [14]. 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 [18, 19]. 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.

The remainder of the paper is organized as follows: we report a short analysis of the related scientific literature in Sect. 2 and Sect. 3 provides the definition of content sensitivity analysis and presents some challenging aspects of this new task together with some hints for possible solutions; the preliminary experiments are reported and discussed in Sect. 4; finally, Sect. 5 concludes by also presenting some open problems and suggestions for future research.

2 Related Work

With the success of online social networks and content sharing platforms, understanding and measuring the exposure of user privacy in the Web has become crucial [11, 12]. Thus, many different metrics and methods have been proposed with the goal of assessing the risk of privacy leakage in posting activities [1, 23]. Most research efforts, however, focus on measuring the overall exposure of users according to their privacy settings [8, 19] or position within the network [18].

Very few research works address the problem of measuring the amount of sensitivity of user-generated content, and yet different definitions of sensitivity are adopted. In [5], for instance, the authors define sensitivity of a social media post as the extent to which users think the post should be anonymous. Then, they try to understand the nature of content posted anonymously and analyze the differences between content posted on anonymous (e.g., Whisper) and non-anonymous (e.g., Twitter) social media sites. They also find significant linguistic

differences between anonymous and non-anonymous content. A similar approach has been applied on posts collected from a famous question-and-answer website [14]. The authors of this work identify categories of questions for which users are more likely to exercise anonymity and analyze different machine learning model to predict whether a particular answer will be written anonymously. They also show that post sensitivity should be viewed as a nuanced measure rather than as a binary concept. In [2], the authors propose a ranking-based method for assessing the privacy risk emerging from textual contents related to sensitive topics, such as depression. They use latent topic models to capture the background knowledge of an hypothetical rational adversary who aims at targeting the most exposed users. Additionally, the results are exploited to inform and alert users about their risk of being targeting.

Similarly to sentiment analysis [13], valuable linguistic resources are needed to identify sensitive content in texts. To the best of our knowledge, the only works addressing this issue are [6, 22], where the authors leverage prototype theory and traditional theoretical approaches to describe and evaluate a dictionary of privacy designed for content analysis. Dictionary categories are evaluated according to privacy-related categories from an existing content analysis tool, using a variety of text corpora.

The problem of sensitive content detection has been investigated as a pattern recognition problem in images as well. In [25], the authors leverage massive social images and their privacy settings to learn the object-privacy correlation and identify categories of privacy-sensitive object automatically. To increase the accuracy and speed of the classifier, they propose a deep multi-task learning architecture that learn more representative deep convolutional neural networks and more discriminative tree classifier. Additionally, they use the outcomes of such model to identify the most suitable privacy settings and/or blur sensitive objects automatically. This framework is further improved in [24], where the authors add a clustering-based approach to also incorporate trustworthiness of users being granted to see the images in the prediction model.

Contrary to the above-mentioned works, in this paper we formally define the general task of *content sensitivity analysis* independently from the type of data to be analyzed. Additionally, we provide some suggestions for improving the accuracy of the results and show experimentally that the task is challenging, and deserves further investigation and greater research efforts.

3 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 any user-generated content. We first distinguish two cases, namely *basic CSA* and *continuous CSA*, according to the outcome of the analysis (binary or continuous). Then, we identify a set of subtasks and discuss their theoretical and technical details. Before introducing the technical details of CSA, we briefly provide the intuition behind CSA by describing a motivating example.

3.1 Motivating Example

To explain the main objectives of CSA and the scientific challenges associated to them, we consider the example in Fig. 1. To decide whether (and to what extent) the sentence is sensitive, an inference algorithm should be able to answer the following questions:

1. **Subjects:** whose information is going to be disclosed?
2. **Information types:** does the post refer to any potentially sensitive information type?
3. **Terms:** does the post mention any sensitive term?
4. **Topics:** does the post mention any sensitive topic?
5. **Relations:** is sensitive information referred to any of the subjects?

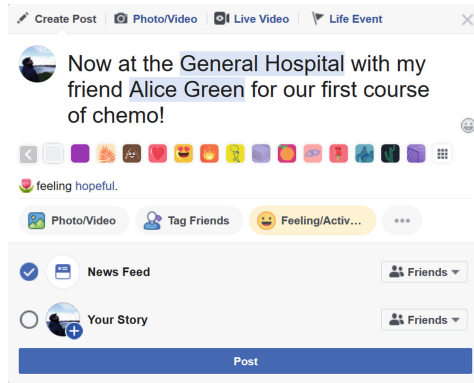


Fig. 1. An example of a potentially privacy-sensitive post.

By observing the post in Fig. 1, it is clear that: the post discloses information about the author and his friend Alice Green (1); the post contains spatiotemporal references (“now” and “General Hospital”), which are generally considered intrinsically sensitive; the post mentions “chemo”, a potentially sensitive term (3); the sentence is related to “cancer”, a potentially sensitive topic (4); the sentence structure suggests that the two subjects of disclosure have cancer and they are both about to start their first course of chemotherapy (5).

It is clear that, reducing sensitivity to anonymity, as done in previous research work [5, 14], 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, 13]. 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 [22].
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.

In the following, we will provide the formal definitions concerning CSA and provide some preliminary ideas on how to address the problem.

3.2 Definitions

Here, we provide the details regarding the formal framework of *content sensitivity analysis*. To this purpose, we consider generic user-generated contents, without specifying their nature (whether textual, visual or audiovisual). We will propose a definition of “sensitivity” further in this section. The simplest way to define CSA is as follows:

Definition 1 (basic content sensitivity analysis). *Given a user-generated object $o_i \in \mathcal{O}$, with \mathcal{O} being the domain of all user-generated contents, the basic content sensitivity analysis task consists in designing a function $f_s : \mathcal{O} \rightarrow \{\text{sens}, na, ns\}$, such that $f_s(o_i) = \text{sens}$ iff o_i is privacy-sensitive, $f_s(o_i) = ns$ iff o_i is not sensitive, otherwise $f_s(o_i) = na$.*

The *na* value is required since the assignment of a correct sensitivity value could be problematic when dealing with controversial contents or borderline topics. In some cases, assessing the sensitivity of a content object is simply impossible without some additional knowledge, i.e., the conversation a post is part of, the identity of the author of a post, and so on. In addition, 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 basic 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 [22]; 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 $\sum_{i=1}^N (f_s(o_i) - \sigma_i)^2$ is minimum.*

The simplest way to address this problem is by setting a regression (or classification, in the case of basic CSA) task. However, we will show in Sect. 4 that such an approach is unable to capture the actual manifold of sensitivity accurately. Hence, in the following sections, we present a fine-grained definition of CSA together with a list of open subproblems related to CSA and provide some hints on how to address them.

3.3 Fine-Grained Content Sensitivity Analysis

In the previous section, we have considered contents as monolithic objects with a sensitivity score associated to them. However, in general, any user-generated

content object (text, video, picture) may contain both privacy-sensitive and privacy-unsensitive elements. For instance, a long text post (or video) may deal with some unsensitive topic but the author may insert some references to her or his private life. Similarly, a user may post a picture of her own desk deemed to be anonymous but some elements may disclose very private information (e.g., the presence of train tickets, drug paraphernalia, someone else’s photo and so on). Moreover, the same object (or some of its elements) may violate the privacy of multiple subjects, including the author and other people mentioned in the corpus, in a different way. For all these reasons, here we propose a fine-grained definition of content sensitivity analysis. The definition is as follows:

Definition 5 (fine-grained content sensitivity analysis). *Let $o_i \in \mathcal{O}$ be a user-generated content object. Let $E_i = \{e_j^i\}_{j=1}^{m_i} \subset \mathcal{E}$ be a set of $m_i \geq 1$ elements (or components) that constitutes the object o_i , with \mathcal{E} being the domain of all possible elements. Let $P_i = \{p_k^i\}_{k=1}^{n_i} \subset \mathcal{P}$ be the set of $n_i \geq 1$ persons (or subjects) mentioned in o_i , with \mathcal{P} being the domain of all subjects. The fine-grained content sensitivity analysis task consists in designing a function $f_s : \mathcal{E} \times \mathcal{P} \rightarrow [-1, 1]$, such that $f_s(e_j^i, p_k^i) = 1$ iff e_j^i is maximally privacy-sensitive for subject p_k^i , $f_s(e_j^i, p_k^i) = -1$ iff e_j^i is minimally privacy-sensitive for subject p_k^i , $f_s(e_j^i, p_k^i) = 0$ iff e_j^i has unknown sensitivity for subject p_k^i . The value $\sigma_{jk}^i = f_s(e_j^i, p_k^i)$ is the **sensitivity score** of element e_j^i towards subject p_k^i .*

Notice that $|E_i| \geq 1$ since each object contains at least one element (when $|E_i| = 1$, the only element e_1^i corresponds the object o_i itself). Similarly $|P_i| \geq 1$ because each object has at least the author as subject. In the example reported in Fig. 1, the post contains only one element (there is only one sentence) and concerns two subjects (the author and Alice Green). According to Definition 5 (and to what we said in Sect. 3.1), the sensitivity score of the post towards both the author and Alice Green will be high.

3.4 Challenges and Possible Solutions

Fine-grained content sensitivity analysis presents many scientific and technical challenges, and may benefit of the cross-fertilization of computational linguistics, machine learning and semantic analysis. Addressing the problem of connecting sensitivity to specific subjects in texts requires the solution of many NLP tasks such as named entity recognition, relation extraction [21], and coreference resolution [4]. Additionally, concept extraction and topic modeling are important to understand whether a given text deals with sensitive content. To this purpose, privacy dictionaries [22] could provide a valid support for tagging certain topics/terms as sensitive or non-sensitive. Sentiment analysis and emotion detection could also reveal private personality traits if related to contents associated to certain topics, persons or categories of persons. Furthermore, elements in a sentence cannot be simply considered as separated entities, but the connection between different parts of a text play an important role in determining the correct fine-grained sensitivity. It is clear that such a complex problem requires the

availability of massive annotated text corpora and the design of robust machine learning algorithms to cope with the sparsity of the feature space. All these considerations apply to the case of visual and audiovisual content as well, but, in addition, the intrinsic difficulty of handling multimedia data makes the above mentioned challenge even harder and more computationally expensive.

In the next section, we will show how the basic content sensitivity analysis settings can be modeled as a binary classification problem on text data using different approaches with scarce or moderate success, thus showing the necessity of a more systematic and in-depth investigation of the problem.

4 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 basic CSA framework (see Definition 1 in Sect. 3) with the only difference that we do not consider the “na” class. 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.

4.1 Annotated Corpus

Since all previous attempts of identifying sensitive text have leveraged user anonymity as a discriminant for sensitive content [5, 14], 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 were 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).

4.2 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 [17]. We use this word embedding method as it consistently outperforms both continuous bag-of-words and skip-gram model architectures of *word2vec* [10]. In detail, we use three representation, here called *WV25*, *WV50* and *WV100* with, respectively, 25, 50 and 100 dimensions¹. 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.

4.3 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 basic content sensitivity analysis task. Hence, we use the following default parameter for each classifier:

- **kNN**: 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.

¹ Pre-trained vectors are available at <https://nlp.stanford.edu/projects/glove/>.

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.

4.4 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 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 negatives². 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.

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

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 [5, 14], or rely on focus groups and surveys [22]. Moreover, without a sufficiently large corpus, not even the application of otherwise successful deep learning techniques (e.g., RNNs for sentiment analysis [9]) 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.

² Due to space limitations, we do not report detailed precision/recall results.

So, more complex and heterogenous models should be considered. Probably, 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 (an open problem in sentiment analysis as well [20]). 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.

5 Conclusions

In this paper, we have addressed the problem of determining whether a given content object is privacy-sensitive or not by defining the generic task of content sensitivity analysis (CSA). Then, we have declined it according to increasing complexity of the problem settings. Although the task promises to be challenging, we have shown that it is not unfeasible by presenting a simplified formulation of CSA based on text categorization. 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. The authors would like to thank Daniele Scanu for implementing the Telegram bot used by the annotators. 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. Biega, J.A., Gummadi, K.P., Mele, I., Milchevski, D., Tryfonopoulos, C., Weikum, G.: R-Susceptibility: an IR-centric approach to assessing privacy risks for users in online communities. In: *Proceedings of ACM SIGIR 2016*, pp. 365–374 (2016)
3. Celli, F., Pianesi, F., Stillwell, D., Kosinski, M.: Workshop on computational personality recognition: shared task. In: *Proceedings of ICWSM 2013* (2013)
4. Clark, K., Manning, C.D.: Improving coreference resolution by learning entity-level distributed representations. In: *Proceedings of ACL 2016* (2016)

5. 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)
6. Gill, A.J., Vasalou, A., Papoutsis, C., Joinson, A.N.: Privacy dictionary: a linguistic taxonomy of privacy for content analysis. In: Proceedings of ACM CHI 2011, pp. 3227–3236 (2011)
7. 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, Heidelberg (2012). https://doi.org/10.1007/978-1-4614-3223-4_13
8. Liu, K., Terzi, E.: A framework for computing the privacy scores of users in online social networks. *TKDD* **5**(1), 6:1–6:30 (2010)
9. Ma, Y., Peng, H., Khan, T., Cambria, E., Hussain, A.: Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. *Cogn. Comput.* **10**(4), 639–650 (2018). <https://doi.org/10.1007/s12559-018-9549-x>
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of NIPS 2013, pp. 3111–3119 (2013)
11. Oukemeni, S., Rifà-Pous, H., i Puig, J.M.M.: IPAM: information privacy assessment metric in microblogging online social networks. *IEEE Access* **7**, 114817–114836 (2019)
12. 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)
13. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Found. Trends Inf. Retrieval* **2**(1–2), 1–135 (2007)
14. 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)
15. Peddinti, S.T., Ross, K.W., Cappos, J.: Finding sensitive accounts on Twitter: an automated approach based on follower anonymity. In: Proceedings of ICWSM 2016, pp. 655–658 (2016)
16. Peddinti, S.T., Ross, K.W., Cappos, J.: User anonymity on Twitter. *IEEE Secur. Privacy* **15**(3), 84–87 (2017)
17. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of EMNLP 2014, pp. 1532–1543 (2014)
18. Pensa, R.G., di Blasi, G., Bioglio, L.: Network-aware privacy risk estimation in online social networks. *Soc. Netw. Anal. Mining* **9**(1), 15:1–15:15 (2019)
19. Pensa, R.G., Blasi, G.D.: A privacy self-assessment framework for online social networks. *Expert Syst. Appl.* **86**, 18–31 (2017)
20. Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A.F., Hussain, A.: Multimodal sentiment analysis: addressing key issues and setting up the baselines. *IEEE Intell. Syst.* **33**(6), 17–25 (2018)
21. Surdeanu, M., McClosky, D., Smith, M., Gusev, A., Manning, C.D.: Customizing an information extraction system to a new domain. In: Proceedings of RELMS@ACL 2011, pp. 2–10 (2011)
22. Vasalou, A., Gill, A.J., Mazanderani, F., Papoutsis, C., Joinson, A.N.: Privacy dictionary: a new resource for the automated content analysis of privacy. *JASIST* **62**(11), 2095–2105 (2011)
23. Wagner, I., Eckhoff, D.: Technical privacy metrics: a systematic survey. *ACM Comput. Surv.* **51**(3), 57:1–57:38 (2018)

24. Yu, J., Kuang, Z., Zhang, B., Zhang, W., Lin, D., Fan, J.: Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing. *IEEE Trans. Inf. Forensics Secur.* **13**(5), 1317–1332 (2018)
25. Yu, J., Zhang, B., Kuang, Z., Lin, D., Fan, J.: iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Trans. Inf. Forensics Secur.* **12**(5), 1005–1016 (2017)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

