Towards a Multi-Stage Approach to Detect Privacy Breaches in Physician Reviews

Frederik S. Bäumer, Joschka Kersting,
Matthias Orlikowski and Michaela Geierhos

Semantic Information Processing Group, Paderborn University, Germany
{fbaeumer,jkers,morlikow,geierhos}@mail.upb.de
https://go.upb.de/seminfo

Abstract. Physician Review Websites allow users to evaluate their experiences with health services. As these evaluations are regularly contextualized with facts from users' private lives, they often accidentally disclose personal information on the Web. This poses a serious threat to users' privacy. In this paper, we report on early work in progress on "Text Broom", a tool to detect privacy breaches in user-generated texts. For this purpose, we conceptualize a pipeline which combines methods of Natural Language Processing such as Named Entity Recognition, linguistic patterns and domain-specific Machine Learning approaches which have the potential to recognize privacy violations with wide coverage. A prototypical web application is openly accesible.

Keywords: Detection of Privacy Violations, Physician Reviews

1 Motivation

Due to progressive semantic enrichment, the Web is becoming a vast resource for exciting data-driven applications. However, this also poses a threat to individual users. As newly exposed data can be linked with existing resources more and more effectively, even implicitly disclosed individual pieces of personal information may have harmful consequences for users. An important example in this context are Physician Review Websites (PRWs), which enable users to rate medical services on the Web. To provide an authentic rating, patients often add private information, e.g. about locations, diseases or medication. This makes them potentially identifiable by third parties [3]. In this paper, we present work in progress on Text Broom[1], a tool which detects and highlights privacy breaches in user-generated texts to prevent accidental disclosure of information. While possible privacy threats are obvious with respect to specific entities (e.g. locations), they are much less obvious in full texts [1, 2, 5]. Natural language (NL) allows us to share information in numerous subtle ways, so that information is more than a sum of words. Therefore, in order to prevent violations of privacy and personality rights [4, 7, 6], a lot of open challenges need to be

[1] A prototype of TextBroom is available under `https://bit.ly/2vrDd5Q`

adressed [3]. Related work [8] uses existing Named Entity Recognition (NER) methods to detect explicitly mentioned entities in texts and remove or randomly replace them. We substantially extend this approach by also considering *inherent* private information. For example, the sentence“*As mother of three girls*” discloses information about family relations and gender without stating them explicitly. Similar sentences are very frequent in physicians’ reviews [3]. Therefore, we combine Natural Language Processing (NLP) methods (NER, linguistic rules and patterns), knowledge resources (linked data, gazetteers) and domain-specific machine learning models to detect potential privacy violations. Note, that we report on early work in progress and do not provide quantitative evaluations.

2 Text Broom: A Multi-Stage Approach

Text Broom basically adapts the idea of Bäumer et al. [3], who recognize privacy breaches in physician reviews using a pipeline of NLP approaches (multi-stage approach) which provide different perspectives and levels of granularity. Bäumer et al. [3] use a series of patterns which reach high precision, but low recall. They also do not take into account the problem of ambiguities and the complexity of grammatical constructions. For this reason, the Text Broom pipeline processes a much wider range of linguistic information. We also follow Kleinberg and Mozes [8], who focus on the visualization of potential violations, and extend this approach with ideas from explainability research.

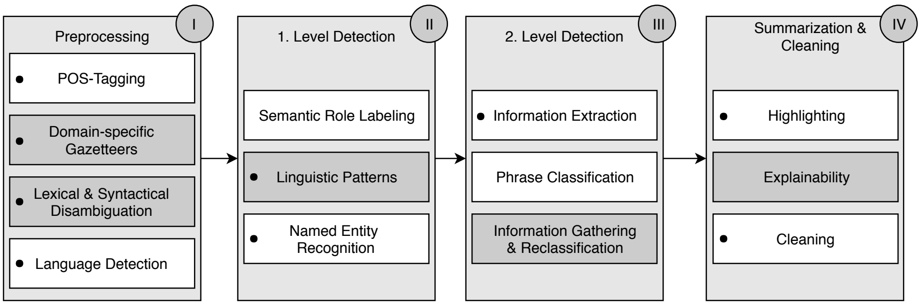
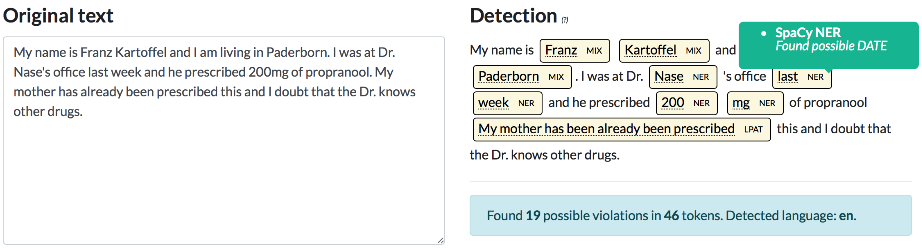


Fig. 1. Processing pipeline for detection and highlighting of privacy breaches

Fig. 1 shows the current processing pipeline of Text Broom, which is divided into four stages. Already implemented components are marked with a dot. The componentes are organized into separate stages which allow each component to use all the information generated in the previous stage. For example, the linguistic patterns component uses the results of POS tagging, the gazetteers and syntactic information to detect possible phrases containing private information. Because of limited space, we will only give outlines of the central components (see grey boxes in Fig. 1).

Gazetters are used in the preprocessing stage and contain extensive lists of terms for drugs and diseases. These are technical terms which are not covered by domain-unspecific methods (e.g. general NER). In addition to doctor portals and pharmacy websites, we also use Wikidata for the maintenance and extension of the drug gazetteers. Since the tokens defined in the gazetteers are applied without taking into account further context, the found matches are exclusively used as input for other components (e.g. linguistic patterns). **Lexical and syntactic disambiguation** allows us to improve detection quality. Lexical disambiguation provides information about which reading of a word is meant in the given context and can minimize recognition errors. An example is "*The doctor also treats my* ***mom***", where "mom" stands for mother (breach of privacy towards a third party) and not e.g. Mars Orbiter Mission. More importantly, disambiguation allows us to consider additional surface forms which way refer to the mentioned entity (e.g. mom, mother, female parent). These may be more likely to occur in gazetteers and thus potentially increase recall. **Linguistic patterns** such as "`(My)?+GAZETTEER+(ADV)?+VERB`" use gazetteers in a predefined context [3]. In this example, the context is defined by the optional adjective "My" and a following verb. Theoretically, suitable verbs could also be maintained in a gazetteer, but we aim to detect a broader range of candidate privacy breaches. Consequently, this type of pattern tends to detect a lot of false positives. Therefore, we combine several different detection methods. For example, in the case of "*My mother also goes to this doctor*", the word "mother" would also be recognized by the NER component as a person, so that there is additional evidence of a potential breach. **Information gathering and reclassification** is an important processing step, since, as mentioned, different components discover many and sometimes contradictory evidence for privacy breaches. This component can combine and re-evaluate indications of potential privacy violations that are related, merge and re-evaluate the different types of evidence and, in principle, classify the identified entities. For example, disclosing a real name is clearly more serious than merely mentioning the name of a drug in isolation and should be marked accordingly. **Explanation generation** or explainability is located in the fourth stages. This component generates explanations for why certain words, phrases or sentences are potentially harmful from a privacy perspective. It is motivated by research in the field of fair, accountable and transparent (FAT) machine learning and explainable artificial intelligence. Since enabling users to understand the privacy implications of indvidual statements is an explicit design goal of Text Broom, we regard this component to be very important. It is still in very early stages. Currently, we highlight segments of text which have been detected as potential privacy breaches directly based on the output of the other components.

Fig. 2 shows Text Broom's web interface with exemplary input and output. In the given example, our system detects a notable amount of evidence for potential breaches, but there are also obvious problems. For example, the system has not detected the drug "propranool" (Propranolol) due to a spelling mistake and ignored the co-reference between "Dr. Nase" and "he". To tackle these and similar problems in a rigorous fashion, we are currently working on annotating a comprehensive evaluation dataset. But even at this early stage, Text Broom showcases a number of promising approaches to improve online privacy. As soon as the tool has matured, we will provide a server-sided interface (API). Professional providers can then integrate Text Broom into their services to help users to protect their privacy and also prevent their business from encountering legal issues due to data protections laws, especially within the European Union.



Original text

My name is Franz Kartoffel and I am living in Paderborn. I was at Dr. Nase's office last week and he prescribed 200mg of propranool. My mother has already been prescribed this and I doubt that the Dr. knows other drugs.

Detection (?)

My name is Franz MIX Kartoffel MIX and Paderborn MIX . I was at Dr. Nase NER 's office last NER week NER and he prescribed 200 NER mg NER of propranool My mother has been already been prescribed LPAT this and I doubt that the Dr. knows other drugs.

- SpaCy NER *Found possible DATE*

Found **19** possible violations in **46** tokens. Detected language: **en.**

Fig. 2. Current version of Text Broom

References

1. S. Afroz, A. C. Islam, A. Stolerman, R. Greenstadt, and D. McCoy. Doppelgänger finder: Taking stylometry to the underground. In *Proceedings of the 2014 IEEE Symposium on Security and Privacy*. IEEE, 2014.
2. M. Balduzzi, C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel. Abusing Social Networks for Automated User Profiling. In *Proceedings of the 13th International Conference on RAID*, pages 422–441, Berlin / Heidelberg, 2010. Springer.
3. F. S. Bäumer, N. Grote, J. Kersting, and M. Geierhos. Privacy matters: Detecting nocuous patient data exposure in online physician reviews. In *Proceedings of the 23rd ICIST 2017, Communications in Computer and Information Science*, volume 756, pages 77–89, Druskininkai, Lithuania, 2017. Springer.
4. R. Bild, K. A. Kuhn, and F. Prasser. SafePub: A truthful data anonymization algorithm with strong privacy guarantees. *Proceedings on PET*, 2018(1), 2018.
5. G. Danezis and C. Troncoso. You cannot hide for long. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*. ACM, 2013.
6. O. Ferrández, B. R. South, S. Shen, F. J. Friedlin, M. H. Samore, and S. M. Meystre. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. *JAMIA*, 20(1):77–83, 2012.
7. M. Geierhos and F. S. Bäumer. Erfahrungsberichte aus zweiter Hand: Erkenntnisse über die Autorschaft von Arztbewertungen in Online-Portalen. In *Book of Abstracts der DHd-Tagung 2015*, pages 69–72, Graz, 2015. DHd.
8. B. Kleinberg and M. Mozes. Web-based text anonymization with node.js: Introducing NETANOS (named entity-based text anonymization for open science). *The Journal of Open Source Software*, 2(14):293, 2017.