Towards Exploring User Perception of a Privacy Sensitive Information Detection Tool

Vanessa Bracamonte[1], Welderufael B. Tesfay[2] and Shinsaku Kiyomoto[1]
[1]*KDDI Research, Inc., Saitama, Japan*
[2]*Goethe University Frankfurt, Frankfurt am Main, Germany*

Keywords: Privacy Tools, User Perception, Privacy Sensitive Information.

Abstract: Users reveal privacy sensitive information when they post on social media, which can have negative consequences. To help these users make informed decisions, different tools have been developed that detect privacy sensitive information (PSI) and provide alerts. However, how users would perceive this type of tool has not yet been evaluated. In this position paper, we take the first steps to address this gap, by exploring user intention, perceived usefulness and attitude towards a PSI detection tool. We designed an experiment and showed participants examples of the PSI detection tool alerts, and quantitatively and qualitatively evaluated their response. The results showed that participants perceived the PSI detection tool as useful, had positive interest, and a low level of concern about it, although they had a neutral level of intention of using the tool. The participants' open-ended responses revealed that they considered the PSI detection tool useful, but mostly for other people and not for themselves. In addition, they were concerned about the privacy risks of using the tool and about its effectiveness. The findings reveal the challenges that PSI detection tools have to overcome to gain acceptance among users that would benefit from this type of privacy protection.

1 INTRODUCTION

Over the last couple of decades, information technology and digital services have become an integral part of the online society. Always connected devices such as the smartphone have fostered the transfer of many offline services and activities to online. As a result, internet users often release a huge amount of data while using these services. As such, the user concerns over the consequences of privacy and liberty have also grown. To address these concerns, research efforts have focused on devising privacy enhancing technologies, enacting data protection laws and studying user information disclosure behaviours.

Regulations such as the EU General Data Protection Regulation (GDPR) aim to protect user's privacy by enforcing different requirements. Privacy-by-design and privacy default are among the key principles enshrined in the GDPR. In particular, Article 9 of the GDPR highlights that certain information types need special care. Previous studies have demonstrated that when users (un)intentionally divulge privacy sensitive information (PSI), they often regret having done so (Sleeper et al., 2013; Wang et al., 2011). Furthermore, Acquisti and Fong (2020) have shown that the disclosure PSI such as information related to religious affiliations and sexual orientation, can be utilized by different parties to discriminate users, e.g., in job applications screening processes.

To help users make informed decisions with respect to their PSI disclosures, Privacy Enhancing Tools (PETs) such as Privacy Detective (Caliskan Islam et al., 2014a) and PrivacyBot (Tesfay et al., 2019) have been developed, which detect PSI in user-generated unstructured texts. These tools have a promising potential for empowering users; however, users' perception of these tools has not yet been evaluated. In this paper, our main objective is to explore how users perceive a PSI detection tool, whether PSI type and explanation information influences their perception, and what are their most frequent comments and concerns, in order to improve the design of these tools.

2 RELATED WORK

This section gives an overview of the existing work both in PSI detection in social networking sites, and

Bracamonte, V., Tesfay, W. and Kiyomoto, S.
Towards Exploring User Perception of a Privacy Sensitive Information Detection Tool.
DOI: 10.5220/0010319706280634
In *Proceedings of the 7th International Conference on Information Systems Security and Privacy (ICISSP 2021)*, pages 628-634
ISBN: 978-989-758-491-6

user evaluations of Privacy Enhancing Technologies (PETs) based on Machine Learning (ML) techniques.

2.1 PSI Detection

Detecting PSI, especially in social networking sites, is a challenging yet growing area of research (Tesfay et al., 2016). Wang et al. (2011) presented two models for the prediction of personally identifiable information (PII) in emails. Similarly, Bier and Prior (2014) developed a process that focuses on the automated recognition of PII. The aim of their study was to enable companies to uncover PII in incoming email communication. Sokolova et al. (2009) focused on health information. Their work presented a mechanism to detect privacy sensitive health information.

Castillo and Chen (2016) proposed a transfer learning approach to detect PSI in tweets. Mao et al. (2011) demonstrated the application of classification techniques to identify three types of privacy leaks in tweets; namely, revealing dates of vacation plans, tweeting under the influence of alcohol, and revealing medical conditions. Tesfay et al. (2019) presented the PrivacyBot PSI detection tool that identifies 14 different information types that are defined in Art. 9 of the EU GDPR and Caliskan Islam et al. (2014b)'s work.

2.2 User Evaluation of ML-based Nudging PETs

When users navigate through different configurations in online services, they encounter a large number of decision points. These decisions often have a big impact on their privacy. Nudging tools based on machine learning algorithms for PSI detection in unstructured texts could help users with these decisions, but user evaluation remains unattended research arena, and therefore we do not yet understand how these tools may be perceived.

In general, however, nudging research has shown promising results in influencing users towards more informed privacy-related decisions (Acquisti et al., 2017). Vishwamitra et al. (2017) studied the effect of privacy-enhancing obfuscation such as "blurring" and "blocking" on user perceptions measured by image satisfaction, information sufficiency, enjoyment, and social presence. Bracamonte et al. (2019) found that users have a positive interest in PETs such as privacy policy summarization tools (Tesfay et al., 2018), and that adding explanation information to the results of these tools, such as highlighting privacy policy segments, can increase perceptions of trustworthiness and usefulness, as well as intention of use (Bracamonte et al., 2020).

These studies demonstrate that users, in general, have a positive perception of PETs. However, to the best of our knowledge, there is a limited research effort in applying similar user evaluation studies for PSI detection tools.

3 METHOD

3.1 Experiment Design

We designed the experiment to explore user perception of the PSI detection tool, as follows. We defined two factors, *Explanation* and *PSI type*, with two levels each. For *Explanation*, the levels were No Explanation (Control) and Explanation. For PSI type, we chose two types from the list defined by Tesfay et al. (2019): Health and Family.

The combination of factors resulted in four experimental conditions (between-subjects design), and we prepared two examples of the PSI detection tool alert for each of the conditions. The content of the social media posts for the examples was related to Health or Family, according to the definition by Tesfay et al. (2019), and were taken from Twitter.

In the alert examples, we used highlighting of important words to simulate the explanation of the result of the PSI detection tool. Although existing PSI detection tools do not currently provide explanations for their results, the technique of keyword highlighting for explanation visualization is commonly used for text data. For the conditions with explanation, we highlighted words in the text which were important to its classification as privacy sensitive within its type (Health or Family). The conditions without explanation (Control) used the same texts, but without highlights. Figure 1 shows how the alert examples were presented to the participants in the survey.

3.2 Questionnaire

After viewing the privacy alert examples, we asked the participants questions regarding *intention of use*, *perceived usefulness* and attitude (*interest* and *concern about the tool*) regarding the PSI detection tool. Table 1 shows the detail of the questions. We included an open-ended response question for the participants to indicate their opinion about the privacy alert tool (*"Please write your opinions, comments or concerns about the privacy alert tool described."*).

The questionnaire also included questions about the participants' social media use and social media privacy concerns, to help characterize the sample. We asked questions on frequency of social media

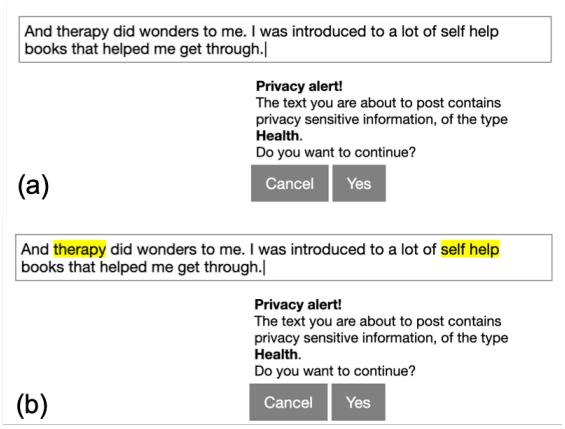

Figure 1: Example of the privacy sensitive information detection tool alert included in the questionnaire, corresponding to the Health PSI type. (a) Base design (Control). (b) Alert with highlighted words (Explanation).

Table 1: Questionnaire items for the main variables of interest in this study. The responses are on a 7-point scale from Strongly disagree to Strongly agree.

Variable	Question
Intention	I would use this tool to receive this type of privacy alert.
Usefulness	This tool would be useful to decide whether to share privacy sensitive information.
Interest	I am interested in trying out this tool.
Concern	I have concerns about trying out this tool.

use (*"I use social media (Twitter, Facebook, etc.)"*) and personal information posting on social media (*"I post personal information on social media (Twitter, Facebook, etc.)"*) with a 7-point response scale ranging from *Never* to *Very frequently*. We included a question on privacy concerns when using social media (*"I worry about the consequences of posting personal information on social media"*) with a 7-point response scale ranging from *Strongly disagree* to *Strongly agree*, and a question on whether the participants used any privacy tools (*"Do you currently use any tools for protecting your privacy on social media?" (Yes/No)*). For this last question, we asked participants who had responded *Yes* to indicate which privacy tools they used.

Finally, the questionnaire included an attention check question, and age and gender questions.

3.3 Participants

We conducted the study on the Amazon Mechanical Turk (AMT) platform, on October 14-15, 2020. AMT workers with the following characteristics were recruited: had worked a minimum of 1000 tasks in the platform (HITs), had an 99% task approval rate, and were from the USA, Canada, Australia or the UK. The participants were compensated with $1.25 for completing the task. While running the study, we identified responses with answers which were completely unrelated to the questions and which failed the attention check question. These responses were not approved and were not included in the data.

4 RESULTS

4.1 Sample Characteristics

The survey initially obtained 160 responses. We analyzed the response to the attention check question and identified 4 cases with incorrect answers. These cases were eliminated from the analysis, which resulted in a valid sample of 156 participants.

The gender distribution of participants was as follows (number of participants in parentheses): female 38.7% (60), male 60.6% (94), 1 other and 1 NA. The age distribution was: 20-29 y/o 21% (33), 30-39 y/o 52% (81), 40-49 y/o 17% (27), 50-59 y/o 7% (11), 60+ y/o 2% (3), with 1 NA.

4.2 Social Media Use and Social Media Privacy Concerns

As Figure 2 shows, the majority of participants frequently used social media, with the highest proportion corresponding to a *Very frequently* response.

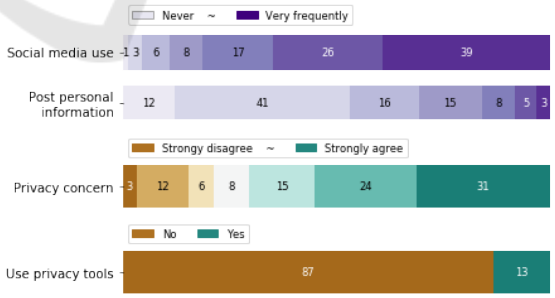

Figure 2: Distribution of responses to the questions about the participants' frequency of social media use and personal information posting, social media privacy concerns and use of privacy tools. Numbers indicate percentage.

On the other hand, the majority of participants indicated that they seldom posted personal information on social media. Participants indicated concern about the consequences of posting personal information on social media: the median of responses was 6,

significantly greater than the neutral point (Wilcoxon signed-rank test, $p<0.001$). The vast majority of participants did not report the use of any privacy protection tools. Of the participants that indicated the use of a privacy protection tool, 7 mentioned the use of privacy settings in the social media platform itself (e.g. *"only the privacy settings in the actual apps. like friends can see posts only.."*); 4 mentioned the use of security settings (e.g *"password"*); and 5 mentioned the use of security/privacy tools (e.g *"VPN", "Norton 360"*).

4.3 Perception of the Privacy Alert Tool

4.3.1 Quantitative Analysis

We used the non-parametric method Aligned Ranks Transform ANOVA (Wobbrock et al., 2011) to test whether the factors of *Explanation* and *PSI type* had an effect on the variables of interest: *intention of use*, *perceived usefulness*, *interest* and *concern* regarding the PSI detection tool alert. The results of the analysis indicated that there were no significant effects of either of the factors for any of the variable.

We proceeded to analyze the whole sample. Figure 3 shows the distribution of responses.

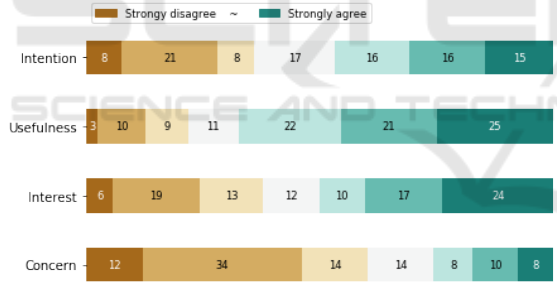

Figure 3: Distribution of responses to the questions about the participants' intention of use, perceived usefulness, interest and concern regarding the PSI detection tool. Numbers indicate percentage.

Figure 4 shows the box plots with the median for each of the variables of interest; the dashed line indicates the neutral point.

We conducted the non-parametric one-sample Wilcoxon signed-rank test, to evaluate whether the scores for *intention of use*, *perceived usefulness* and *interest* were significantly greater than the neutral score of 4. In the case of *concern about the tool*, we tested whether the scores were significantly less than the neutral score.

The results indicate that the participants had a positive perception and attitude towards the tool: usefulness ($p<0.001$) and interest in the tool ($p=0.001$)

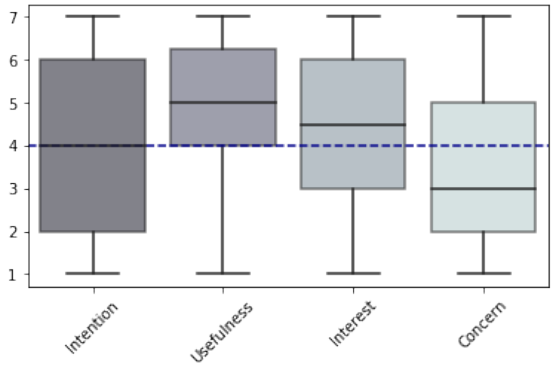

Figure 4: Box plots for intention of use, perceived usefulness, interest and concern regarding the tool. The dashed line indicates the middle point of the response scale (neutral).

were significantly positive. In addition, participants had a low level of concern about trying out the tool ($p<0.001$). On the other hand, participants intention of use of the PSI detection tool was not positive ($p=0.11$), with a median of 4 (neutral middle of the scale) as can be observed in Figure 4.

4.3.2 Open-ended Comments

We qualitatively analyzed the content of the responses to the open-ended question on the participants' opinions and concerns regarding the PSI detection tool. We first reviewed each response and classified them by whether they contained all positive, all negative or both positive and negative comments. During the review we also identified responses that did not correspond to the question (which were classified as Other) and blank responses. Table 2 shows the results by experiment condition.

The results show that in all conditions, the proportion of participants that gave an all positive comment was similar than those who gave an all negative comment. The exception is the case of the Control-Family condition, where the number of all positive comments was more than twice the number of all negative comments. The number of participants that mentioned both positive and negative aspects was similar for all conditions.

On the positive side, participants mentioned usefulness most frequently, followed by a general perception of the tool being good or a good idea and interesting (Table 3). With regards to usefulness, it is interesting to note that participants frequently mentioned it as useful for others, not themselves.

On the negative side (Table 4), the aspects mentioned were more varied. Concern about the privacy risk of the PSI detection tool itself was most frequently mentioned: participants were concerned

Table 2: Classification of open-ended comments from participants, by condition. The table includes the category Other (unrelated comments) and Blank (no answer).

Condition		n	%
Control-Health	Positive	15	38%
	Both	8	20%
	Negative	13	33%
	Other	3	8%
	Blank	1	3%
Explanation-Health	Positive	15	38%
	Both	6	15%
	Negative	16	41%
	Other	1	3%
	Blank	1	3%
Control-Family	Positive	20	53%
	Both	8	21%
	Negative	8	21%
	Other	2	5%
Explanation-Family	Positive	12	31%
	Both	10	26%
	Negative	13	33%
	Other	3	8%
	Blank	1	3%

Table 3: Positive aspects of the PSI detection tool mentioned by participants in the open-ended comments.

Type	n	Example
useful	56	”I think that this tool would be useful to get people thinking about the type of things they post online.”
good	22	”I believe this is a good tool to help those who are not very aware of what sensitive information they may be posting online ”
interesting	11	”I didn’t know it was a thing so it’s interesting”
performance	2	”Based on the examples, I’d say this privacy alert tool is doing its job.”

about how the PSI detection tool would handle their private data and what it might do with it.

Participants mentioned the PSI detection tool not being useful as the second most frequent negative aspect; specifically, participants considered that the PSI detection tool was not useful for themselves, because they already took the necessary precautions to avoid posting personal information on social media. This type of comment was frequently accompanied by positive comments regarding the potential usefulness of the PSI detection tool for other people. The categories of others that participants consider might benefit from

Table 4: Negative aspects of the PSI detection tool mentioned by participants in the open-ended comments.

Type	n	Example
privacy	31	” There is still worries about that, if it analyses my data, does it store it? If so the tool itself is a privacy risk.”
not useful	21	”However I myself would likely have no use for it as I don’t post everything about my life like some others. ”
too sensitive	14	”If the tool were configurable to be less sensitive when it makes sense, it could be handy without having too many warnings”
performance	12	”Concerns that I have on the tool include how effective the tool is in detecting privacy information and how and why the tool detects certain words as sensitive.”
inconvenient	9	”Also, I would find it annoying after awhile I think.”
censorship	3	”At first glance, it just seems like another form of censorship. A tool attempting to get us to censor ourselves.”
over-reliance	2	”In fact, I’m kind of afraid that having a reminder like this might make me less careful.”
monitoring	2	” I feel like it’s just monitoring me, which I don’t like.”
cost	1	”(. . .) but I would want to know what I would have to provide/download/etc to use it.”

the PSI detection tool includes: children, young people, older people, and in general people who are ”less aware” of privacy risks on social media.

Next, participants were concerned about the PSI detection tool being ”too sensitive”, and about the general performance of effectiveness of the PSI detection tool. In particular, the information related to family and health was considered not sensitive by some participants, and some of them made suggestions about how the PSI detection tool could deal with this issue. For example, they suggested having sensitivity settings, whitelists or considering the context of the post.

5 DISCUSSION

The quantitative results shows a positive perception of the usefulness of the PSI detection tool, and the qualitative analysis, where usefulness was the most frequent positive aspect mentioned, supports this result. Comments about usefulness were often accompanied by phrases that provide evidence of the nudging effect of the PSI detection tool: "useful to get people thinking", "It would give them a second to pause and think about if it is a good thing to post or not.". On the other hand, usefulness was also mentioned in negative terms and the examination of the content of the responses reveal a dual nature in the participants' perception of usefulness of the PSI detection tool. In summary, participants considered the tool useful, in particular for others but not so much for themselves. This contradiction can also explain the finding that the level of intention of use was neutral and not significantly positive.

We cannot judge the accuracy of the participant's self-perceived privacy risk awareness, but this type of response suggests that it would be beneficial for PSI detection tools to provide information on historical data. This could be done through calculating privacy risk scores (Aghasian et al., 2017), or by detecting examples from the user's past social media posts where (or if) PSI had been disclosed.

Participant's comments show that the foremost concern is the privacy risk of the PSI detection tool itself. This type of privacy tool relies on an important assumption that the users' will grant access to their private information, and users rightly wish to have assurances that their data will not be misused. The trustworthiness of the PSI detection tool in terms of security and privacy protection should be established, through technical means or reputable providers for example, in order to allow users to safely grant it access to current or historical data.

Participants' concern regarding the tool being "too sensitive" also poses an interesting challenge for the design of PSI detection. Current tools have a predetermined classification of what exactly constitutes "sensitive information", but this classification may not be compatible with the users' own concept of PSI. Regardless of whether the user is correct, this incompatibility could result in false alarms (as perceived by the user) and annoyance. PSI detection tools should consider how to offer some flexibility without compromising their goal.

Finally, we did not find a significant effect (positive or negative) of explanations or PSI type in the perception of the tool. In the case of explanation, one possibility may be that the explanations are superfluous with such short texts. Future research will include validation of these results.

5.1 Limitations

This study has the following limitations. First, the design of the study is not comprehensive, in the sense that we did not include all possible PSI types as conditions in the experiment. The main objective of the study was to explore user perception and to identify avenues of future research. Next, the data used to construct the example alerts that we showed to participants were taken from Twitter, a social media platform that is characterized by short text length per post. The responses from participants may be different for alerts corresponding to longer or more complex social media posts, where the privacy sensitive information contained in the text may be harder to judge at a glance. Finally, we asked questions regarding opinions and experience related to privacy sensitive and personal information, but their exact definition may be different for each participant. Future work will aim to clarify and explore users' understanding of these concepts.

6 CONCLUSIONS

In this paper, we conducted a study to explore user perception about a privacy sensitive information (PSI) detection tool for social media. We showed participants examples of the alert provided by the PSI detection tool, with and without explanation of the result, which warned them that they might be posting PSI about health or family. We quantitatively and qualitatively evaluated the participants' response.

The results showed that participants perceived the PSI detection tool as useful, had positive interest, and a low level of concern about it. On the other hand, participants were neutral regarding whether they would use the PSI detection tool, and we did not find significant effects of explanations or between health and family type of PSI for any of the variables. The analysis of the open-ended comments indicated that participants thought that the PSI detection tool was useful (although mostly for others and not themselves), but were concerned about the privacy risks posed by the use of the tool itself. The participants' comments also included suggestions on possible features of the PSI detection tool, which suggests that there was a level of engagement with the idea. Future work is planned to quantitatively validate the influence of the factors identified in this study, and to apply the findings to the design of a PSI detection tool.

REFERENCES

Acquisti, A., Adjerid, I., Balebako, R., Brandimarte, L., Cranor, L. F., Komanduri, S., Leon, P. G., Sadeh, N., Schaub, F., Sleeper, M., et al. (2017). Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)*, 50(3):1–41.

Acquisti, A. and Fong, C. (2020). An experiment in hiring discrimination via online social networks. *Management Science*, 66(3):1005–1024.

Aghasian, E., Garg, S., Gao, L., Yu, S., and Montgomery, J. (2017). Scoring users' privacy disclosure across multiple online social networks. *IEEE Access*, 5:13118–13130.

Bier, C. and Prior, J. (2014). Detection and labeling of personal identifiable information in e-mails. In *IFIP International Information Security Conference*, pages 351–358. Springer.

Bracamonte, V., Hidano, S., Tesfay, W. B., and Kiyomoto, S. (2019). User study of the effectiveness of a privacy policy summarization tool. In *International Conference on Information Systems Security and Privacy*, pages 186–206. Springer.

Bracamonte, V., Hidano, S., Tesfay, W. B., and Kiyomoto, S. (2020). Evaluating the effect of justification and confidence information on user perception of a privacy policy summarization tool. In *ICISSP*, pages 142–151.

Caliskan Islam, A., Walsh, J., and Greenstadt, R. (2014a). Privacy detective: Detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, WPES '14, page 35–46, New York, NY, USA. Association for Computing Machinery.

Caliskan Islam, A., Walsh, J., and Greenstadt, R. (2014b). Privacy Detective: Detecting Private Information and Collective Privacy Behavior in a Large Social Network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, WPES '14, pages 35–46. Association for Computing Machinery.

Castillo, S. R. M. and Chen, Z. (2016). Using Transfer Learning to Identify Privacy Leaks in Tweets. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, pages 506–513.

Mao, H., Shuai, X., and Kapadia, A. (2011). Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, pages 1–12.

Sleeper, M., Cranshaw, J., Kelley, P. G., Ur, B., Acquisti, A., Cranor, L. F., and Sadeh, N. (2013). "i read my twitter the next morning and was astonished": A conversational perspective on twitter regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 3277–3286, New York, NY, USA. Association for Computing Machinery.

Sokolova, M., El Emam, K., Rose, S., Chowdhury, S., Neri, E., Jonker, E., and Peyton, L. (2009). Personal health information leak prevention in heterogeneous texts. AdaptLRTtoND '09, page 58–69, USA. Association for Computational Linguistics.

Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., and Serna, J. (2018). I read but don't agree: Privacy policy benchmarking using machine learning and the eu gdpr. In *Companion Proceedings of the The Web Conference 2018*, pages 163–166.

Tesfay, W. B., Serna, J., and Pape, S. (2016). Challenges in detecting privacy revealing information in unstructured text. In *PrivOn@ ISWC*.

Tesfay, W. B., Serna, J., and Rannenberg, K. (2019). Privacybot: Detecting privacy sensitive information in unstructured texts. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 53–60.

Vishwamitra, N., Knijnenburg, B., Hu, H., Kelly Caine, Y. P., et al. (2017). Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 39–47.

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., and Cranor, L. F. (2011). "i regretted the minute i pressed share": A qualitative study of regrets on facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11, New York, NY, USA. Association for Computing Machinery.

Wobbrock, J. O., Findlater, L., Gergle, D., and Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 143–146. Association for Computing Machinery.