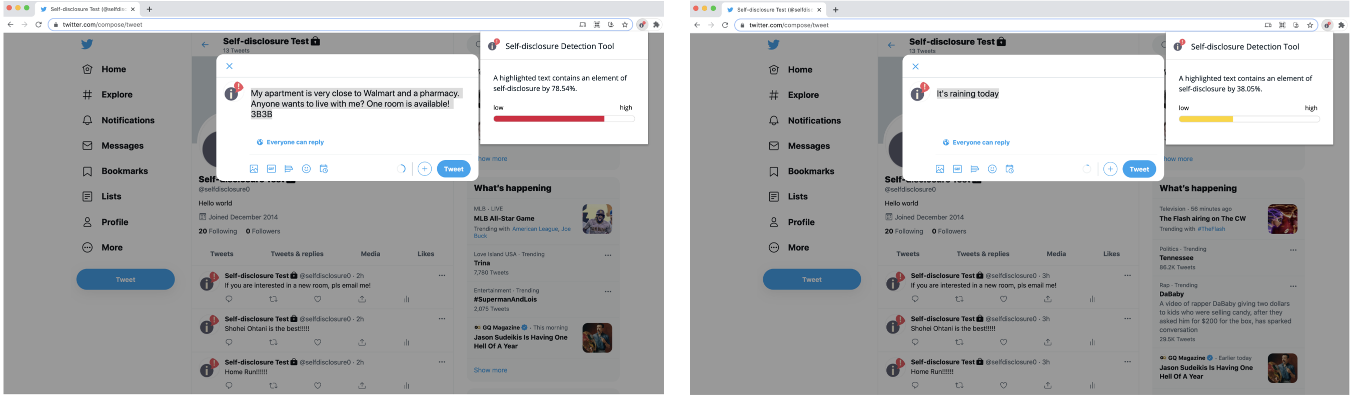
WiSDoM: Guiding Internet Users Toward Safer Privacy Decisions

Jooyoung Lee*
jfl5838@psu.edu
Pennsylvania State University
University Park, PA, USA

Hee Jeong Han*
heejeonghan@psu.edu
Pennsylvania State University
University Park, PA, USA

Michiharu Yamashita*
michiharu@psu.edu
Pennsylvania State University
University Park, PA, USA



twitter.com/compose/tweet

Self-disclosure Test
13 Tweets

My apartment is very close to Walmart and a pharmacy. Anyone wants to live with me? One room is available! 3B3B

Self-disclosure Detection Tool

A highlighted text contains an element of self-disclosure by 78.54%.

@selfdisclosure0
Hello world
Joined December 2014
20 Following 0 Followers

Self-disclosure Test @selfdisclosure0 · 2h
If you are interested in a new room, pls email me!

Self-disclosure Test @selfdisclosure0 · 2h
Shohei Ohtani is the best!!!!!

Self-disclosure Test @selfdisclosure0 · 2h
Home Run!!!!!!

It's raining today

Self-disclosure Detection Tool

A highlighted text contains an element of self-disclosure by 38.05%.

Self-disclosure Test (@selfdisc
Home
Explore
Notifications
Messages
Bookmarks
Lists
Profile
More
Tweet
Everyone can reply
Tweet
low
high
Tweets
Tweets & replies
Media
Likes
What's happening
MLB · LIVE
MLB All-Star Game
Trending with American League, Joe Buck
Love Island USA · Trending
Trina
7,780 Tweets
Entertainment · Trending
#SupermanAndLois
2,075 Tweets
GQ Magazine · This morning
Jason Sudeikis Is Having One Hell Of A Year
Show more
Television · 56 minutes ago
The Flash airing on The CW
Trending with #TheFlash
Politics · Trending
Tennessee
86.2K Tweets
Rap · Trending
DaBaby
A video of rapper DaBaby giving two dollars to kids who were selling candy, after they asked him for $200 for the box, has sparked conversation
29.5K Tweets
GQ Magazine · Earlier today
Jason Sudeikis Is Having One Hell Of A Year
3h

Figure 1: Examples using WiSDoM on Twitter

ABSTRACT

Alongside the surge in individuals' online activities involving Social Network Sites (SNSs), users reveal various types of information related to themselves on such platforms. While self-disclosure can assist users in forming deeper connections with others, oversharing may harm their privacy and security protection. Recent studies have emphasized a need for a self-disclosure mitigation tool that is publicly available and is not domain-specific. Therefore, we present a prototype of WiSDoM (a **W**arn**i**ng tool of **S**elf-**D**isclo**s**ure to **M**itigate privacy harms), which is a Chrome extension tool that informs users about their disclosing behaviors. The proposed tool is powered by a BERT-based objective disclosure detection model. WiSDoM responds fast to any highlighted texts and works on multiple domains without the help of external servers or APIs. Results of a user experiment on Amazon Mechanical Turk indicate the benefits of our tool in assisting users to distinguish self-disclosing content.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing**; *Collaborative and social computing systems and tools.*

KEYWORDS

Self-Disclosure Detection; Privacy; Chrome Extension

ACM Reference Format:
Jooyoung Lee, Hee Jeong Han, and Michiharu Yamashita. 2021. WiSDoM: Guiding Internet Users Toward Safer Privacy Decisions. In *Proceedings of UIST '12: ACM Symposium on User Interface Software and Technology (UIST '21).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Living in a data-driven society, 3.96 billion people are using social media worldwide in 2021, and 58.11% of the world's population engages in social media to stay connected with others [6]. Given the nature of those platforms, self-disclosure — a communication process in which individuals expose information about themselves to another [5] — has appeared to be prevalent in cyberspace. While a body of literature has focused on understanding the positive influence of the voluntary public discourse in relational development, social connectedness, and identity clarification [1, 2, 8, 14, 16, 20], oversharing may also harm users by causing privacy or security vulnerability [3, 17], damaged self-image [7, 19], or emotional stress [23]. This is explainable by humans' limited cognitive ability, which blocks them from considering a wide range of risk variables associated with their disclosures [11]. Therefore, increasing individuals' risk awareness resulting from their sharing behaviors is critical for assisting them in avoiding mistakes, unethical conduct, and, ultimately, protecting themselves from undesired occurrences [9].

*Each author contributed equally to this research.

UIST '21, October 10–13, 2021, Virtual
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

In this paper, we propose a prototype of WiSDoM, a Chrome extension tool that alerts the dangers of self-disclosure to users across online platforms. Our tool aims to guide online users whether their texts contain personal sensitive information about themselves in real-time. WiSDoM can detect objective information such as age, sexual orientation, and employment status.

2 SYSTEM OVERVIEW

2.1 Automated Self-Disclosure Detection Model

2.1.1 Dataset. We use a publicly available dataset [15][1] collected from Reddit as gold-standard labels. The Reddit dataset is released as part of the AAAI 2019 workshop and contains 12,860 labeled comments, and 5,000 unlabeled comments crawled from two support-based subreddits, 'r/CasualConversation' and 'r/OffMyChest'. For the purpose of this study, we only utilize the 'information disclosure' label which is equivalent to objective disclosure. The dataset is split into a ratio of 8:2 for training and testing.

Table 1: A comparison of our models' performance

	LR	RF	KNN	AdaBoost	LightGBM	RNN	CNN	BERT
Accuracy	0.68	0.69	0.58	0.66	0.68	0.62	0.62	**0.71**
Precision	0.65	0.65	0.43	0.54	0.62	0.49	0.50	**0.62**
Recall	0.37	0.37	0.32	0.63	0.44	0.45	0.42	**0.57**
F1	0.62	0.62	0.53	0.58	0.64	0.54	0.46	**0.68**

2.1.2 Model Selection. Our machine learning models are as follows: logistic regression, random forest, K-nearest Neighbors, AdaBoost, and LightGBM using the scikit-learn library.[2] Inspired by Wang et al. [21]'s findings, features of our ML models include TF-IDF, word count, character count, readability score, sentiment scores (e.g., negative, neutral, and positive), the inclusion of first-person pronouns, and topic distributions. Readability of posts is measured by the Flesch Reading Ease (FRE) metric [12], which is calculated based on the number of words per sentence and the number of syllables per word. For sentiment analysis, we use the VADER algorithm [13] to capture positive, negative, neutral, and compound sentiment scores from a text. Lastly, we employ the Latent Dirichlet Allocation algorithm to extract latent topics when a set of documents is given in an unsupervised way [4]. To improve the robustness of ML models, we additionally generate three advanced deep learning models, convolutional neural network (CNN), recurrent neural network (RNN), and BERT. For CNN and RNN models, we use each text as a sequence and directly predict the class based on a text. Among different BERT model, we choose a BERT base (uncased) model from Hugging Face [22].

Due to the uneven distribution of labels in our dataset, we implement cost-sensitive learning [10] in our BERT model. It assigns a higher cost for misclassification of the minority class (self-disclosure texts), instead of decreasing or increasing the total sample size. All models are evaluated based on accuracy, recall of the minority class, precision of the minority class, and F1 score. Models' performance on test set is displayed in Table 1. Given the fact that our BERT model outperforms all of the baseline models, results confirm the efficacy of a fine-tuned language model and the effectiveness of cost-sensitive in mitigating Precision-Recall tradeoffs. Hyperparameters of our best performing model are as follows: lr = 3e-5; max_grad_norm = 1.0; num_total_steps = 300; num_warmup_steps = 30; warmup_proportion = 0.1; weight_decay = 1e-5.

2.2 Chrome Extension Implementation

For easy installation, WiSDoM is deployed as a Google Chrome extension. To incorporate our best performing model (BERT) into the Flask application, we convert our model to an ONNX[3] model and upload it on the back-end. Hence, WiSDoM can visually represent the likelihood of given texts containing objective information about a user, delivering a stronger message than plain texts. In addition, a response time of WiSDoM to display warning messages takes less than one second. Furthermore, WiSDoM do not need any APIs and external servers. Once users download WiSDoM, they can easily use WiSDoM without connecting with external servers so that their text draft before posting is not shared. Figure 1 illustrates a design of the prototype working on Twitter.

3 USABILITY EXPERIMENT

To verify the WiSDoM's effectiveness, we conducted a participant experiment on Amazon Mechanical Turk. We first randomly selected 50 posts in our Reddit dataset, in which half of the comments are self-disclosing, and the other half are not self-disclosing. One hundred workers participated in our experiment wherein 50 workers are in a control group, and the other 50 participants are in a treatment group. We asked them to classify whether a provided text contains an element of objective disclosure. The control group classified sentences themselves, whereas the treatment group was provided to use WiSDoM.

Our experiment results show that the control and treatment groups successfully identified self-disclosing texts by 78.24% and 81.12%, respectively. Therefore, we conclude that WiSDoM plays a positive role in relatively increasing awareness of self-disclosure by 3.7% (p = 0.07). Yet, annotators' success rate in distinguishing non-self-disclosure texts did not differ statistically significantly (42.08% vs. 43.92%, p = 0.35). We speculate that this result is attributable to a relatively weak warning message. Describing potential privacy harms that may occur when particular pieces of personal data are revealed has been reported to have a positive impact on self-disclosure regulation [9, 18].

4 CONCLUSION AND FUTURE WORK

In this paper, we present the prototype of WiSDoM deployed as a Google Chrome extension for objective disclosure detection powered by a BERT-based prediction model. The prototype can give an alarm to users based on their text across multiple online platforms in real-time. Results of the usability experiment highlight the possible implication of WiSDoM in enhancing users' privacy awareness. Finally, we plan to conduct future studies to explore the impact of design factors (e.g., warning messages, notifications) on their actual sharing behaviors and publish our tool in the future.

[1] Their labeling instructions and annotated dataset can be found in this Github repository: https://github.com/kj2013/claff-offmychest.

[2] https://scikit-learn.org/stable/

[3] https://github.com/onnx/onnx

REFERENCES

[1] JinYeong Bak, Suin Kim, and Alice Oh. 2012. Self-disclosure and relationship strength in twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 60–64.

[2] Azy Barak and Orit Gluck-Ofri. 2007. Degree and reciprocity of self-disclosure in online forums. *CyberPsychology & Behavior* 10, 3 (2007), 407–417.

[3] Natalya N Bazarova and Yoon Hyung Choi. 2014. Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites. *Journal of Communication* 64, 4 (2014), 635–657.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[5] Paul C Cozby. 1973. Self-disclosure: a literature review. *Psychological bulletin* 79, 2 (1973), 73.

[6] Brian Dean. 2021. Social Network Usage &amp; Growth Statistics: How Many People Use Social Media in 2021? https://backlinko.com/social-media-users#how-many-people-use-social-media

[7] Bernhard Debatin, Jennette P Lovejoy, Ann-Kathrin Horn, and Brittany N Hughes. 2009. Facebook and online privacy: Attitudes, behaviors, and unintended consequences. *Journal of computer-mediated communication* 15, 1 (2009), 83–108.

[8] Valerian J Derlaga and John H Berg. 1987. *Self-disclosure: Theory, research, and therapy*. Springer Science & Business Media.

[9] Nicolás E Díaz Ferreyra, Tobias Kroll, Esma Aïmeur, Stefan Stieglitz, and Maritta Heisel. 2020. Preventative Nudges: Introducing Risk Cues for Supporting Online Self-Disclosure Decisions. *Information* 11, 8 (2020), 399.

[10] Charles Elkan. 2001. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, Vol. 17. Lawrence Erlbaum Associates Ltd, 973–978.

[11] Gerard Emilien, Rolf Weitkunat, and Frank Lüdicke. 2017. *Consumer perception of product risks and benefits*. Springer.

[12] Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221.

[13] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

[14] Emmi Ignatius and Marja Kokkonen. 2007. Factors contributing to verbal self-disclosure. *Nordic Psychology* 59, 4 (2007), 362–391.

[15] Kokil Jaidka, Iknoor Singh, Jiahui Lu, Niyati Chhaya, and Lyle Ungar. 2020. A report of the CL-Aff OffMyChest Shared Task: Modeling Supportiveness and Disclosure. In *Proceedings of the AAAI-20 Workshop on Affective Content Analysis*. AAAI, New York, USA.

[16] Adam N Joinson, Ulf-Dietrich Reips, Tom Buchanan, and Carina B Paine Schofield. 2010. Privacy, trust, and self-disclosure online. *Human–Computer Interaction* 25, 1 (2010), 1–24.

[17] Jooyoung Lee, Sarah Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. 2021. Digital Inequality Through the Lens of Self-Disclosure. *Proceedings on Privacy Enhancing Technologies* 3 (2021), 372–392.

[18] Vincent Marmion, Felicity Bishop, David E Millard, and Sarah V Stevenage. 2017. The cognitive heuristics behind disclosure decisions. In *International Conference on Social Informatics*. Springer, 591–607.

[19] Ann E Schlosser. 2020. Self-disclosure versus self-presentation on social media. *Current opinion in psychology* 31 (2020), 1–6.

[20] Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2019. Detection and analysis of self-disclosure in online news commentaries. In *The World Wide Web Conference*. 3272–3278.

[21] Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling self-disclosure in social networking sites. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 74–85.

[22] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45.

[23] Wenjing Xie and Cheeyoun Kang. 2015. See you, see me: Teenagers' self-disclosure and regret of posting on social network site. *Computers in Human Behavior* 52 (2015), 398–407.