IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS) | 978-1-6654-0926-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/INFOCOMWKSHPS54753.2022.9798377

REGRETS: A New Corpus of Regrettable (Self-)Disclosures on Social Media

Hervais Simo and Michael Kreutzer
Fraunhofer SIT, Darmstadt, Germany.

***Abstract*—In the past few years, researchers have shown a growing interest in techniques for automated detection of regrettable disclosures (things people wish they had not shared) on social media. Most of these proposals formulate the task of automatically detecting potentially regrettable disclosures as a supervised classification problem. In such a setting, the underlying classification model is trained and validate on a dataset labeled accordingly. However, despite growing efforts, existing approaches remain limited, partly due to the lack of high-quality corpus of regrettable messages and comments shared on social media. Previous work tend to confuse regrettable disclosure with related concepts such as hate speech, profanity and offensive language, ignoring empirical findings on the reasons, the types of contents, and disclosure contexts that often lead to regrets. Moreover, corpora used in prior work are typically limited in size and w.r.t. their source domains (i.e., social media platforms) and scope (i.e., range of regret-related topical content used as labels). The goal of this paper is to contribute towards lowering the barrier for developing effective systems for automated detection of regret-related posts. We propose a novel methodology for large-scale data collection and semi-automated annotation. We introduce *REGRETS*, a new large-scale corpus of 4,7 million regrettable text-only posts and comments with high-quality annotations. Further, we propose regret-specific embeddings models pre-trained on our corpus of user-generated social media texts which were extracted from various popular social media ecosystems. Lastly, we report on analyses that demonstrate the feasibility of partly automating the annotation of social media texts, and the richness of the resulting corpus. We release our findings as resources to facilitate further interdisciplinary research: https://bit.ly/3fO36Ex.**

***Index Terms*—(Self-)disclosures, social media, corpus, privacy**

I. Introduction

When engaging in self-presentation on social media, users often make disclosures that they subsequently regret. Such regrets have been shown to typically resolve around disclosures on sensitive topics and content with strong sentiment, lies, and secrets. As such, regrettable self-disclosures do not only jeopardize peoples' privacy but are also damaging to their reputation and relationships. Proposals to address this issue typically rely on automated detection models trained and evaluated on a large number of regret-related posts with high quality annotations. However, there has been a relative lack of high-quality datasets of regrettable disclosures on social media, i.e., datasets that cover various platforms and regret-related topics uniformly. To remedy this problem, we introduce a new corpus of 4,700,284 regrettable user-generated social media texts annotated with detailed information about different types of topical content that people regret posting. Inspired by Wang et al.'s [3] empirical study of regrets on Facebook, we propose eight regret-related labels for data annotation: Alcohol & drug use (T1), Misogyny (T2), Personal & family (T3), Politics (T4), Profanity & obscenity (T5), Religion (T6), Sex (T7), and Work & company (T8). In order to create such a much-needed high quality corpus, we start by collecting a large amount of publicly available user-generated texts from 30 platforms and 206 sources. We subsequently performed extensive cleaning, and tagged each disclosure with one or several of the eight regret-related labels. This is, to the best of our knowledge, the first large-scale effort to annotate social media texts with regret labels at such a fine level of detail.

This work has been co-funded by: the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

	First round		Second round	
Topical categories	Raw Set	Clean Set	Raw Set	Clean Set
Alcohol & Drugs	145,810	112,040	541,184	512,507
Misogyny	115,320	83,338	456,564	443,671
Personal & Family	128,190	107,989	491,794	450,674
Politics	179,022	157,246	565,419	550,840
Profanity & Obscenity	132,439	103,770	667,830	594,238
Religion	142,494	114,088	571,049	552,155
Sex	159,262	140,685	455,014	412,520
Work & Company	120,191	105,384	542,647	519,389
Total	**1,122,728**	**924,540**	**4,297,372**	**4,077,114**

Table I: Our final raw dataset

II. Methodology

Our methodology consists of the following key steps: gathering raw data, data cleaning and a 2-phase labeling strategy.

Gathering raw data: We collected data from various popular social media platforms including Facebook, Twitter, emerging discussion and self-help boards such as 4chan (4chan.org), Reddit (reddit.com) and Youtube channels, as well as from popular news websites - e.g., Msnbc.com, CNN.com and Breitbart. For data collection, we rely on a combination of automated and non-automated approaches, using topic-related keywords (see https://bit.ly/3InQn7W) as query strings. Non-automated data collection mostly relied on data scraping by hand. For automated data collection approaches, we relied on official API calls, extending tools such as TweetScraper and PRAW. This allows to overcome rate limits and other APIs restrictions. From each identified source, we extract available

text-only posts and comments, and sent them to our backend server. Per source, we only crawled the first 100 posts and the top-50 comments. We collected data twice in a row, first from November 2016 to February 2017, capturing a total of 1,122,728 posts. The second wave took place from May to August 2018 resulting in a total of 4,165,2697 additional posts. The output, 5,420,100 text-only posts and comments, was subsequently cleaned and annotated.

		T1	T2	T3	T4	T5	T6	T7	T8	Avg.
CM-AB	Acc.	0.93	0.94	0.91	0.93	0.92	0.91	0.87	0.86	0.91
	Prec.	0.78	0.81	0.64	0.7	0.72	0.67	0.74	0.74	0.72
	Rec.	0.47	0.71	0.75	0.78	0.54	0.61	0.66	0.65	0.65
	F1	0.58	0.76	0.69	0.73	0.62	0.64	0.70	0.69	0.68
CM-DT	Acc.	0.96	0.96	0.95	0.96	0.95	0.95	0.89	0.88	0.94
	Prec.	0.85	0.88	0.79	0.85	0.78	0.79	0.57	0.58	0.76
	Rec.	0.74	0.82	0.81	0.85	0.80	0.8	0.60	0.60	0.75
	F1	0.79	0.85	0.80	0.85	0.79	0.79	0.58	0.59	0.76
CM-RF	Acc.	0.96	0.97	0.96	0.97	0.97	0.96	0.92	0.91	0.95
	Prec.	0.90	0.91	0.86	0.91	0.86	0.86	0.65	0.63	0.82
	Rec.	0.76	0.83	0.84	0.87	0.84	0.83	0.73	0.75	0.81
	F1	0.82	0.87	0.85	0.89	0.85	0.84	0.69	0.68	0.81
CM-RT	Acc.	0.93	0.95	0.93	0.95	0.94	0.93	0.86	0.85	0.92
	Prec.	0.68	0.77	0.74	0.79	0.73	0.73	0.48	0.47	0.67
	Rec.	0.69	0.78	0.74	0.79	0.71	0.71	0.50	0.52	0.68
	F1	0.68	0.78	0.74	0.79	0.72	0.72	0.50	0.50	0.68
CM-SVM	Acc.	0.96	0.95	0.91	0.96	0.96	0.94	0.89	0.86	0.93
	Prec.	0.82	0.81	0.68	0.89	0.81	0.82	0.58	0.50	0.74
	Rec.	0.76	0.74	0.55	0.79	0.82	0.71	0.64	0.72	0.72
	F1	0.79	0.77	0.61	0.84	0.81	0.76	0.61	0.59	0.72

Table II: Multi-label classifiers' performances on $\widetilde{\mathcal{D}}_{11}$. On a Ubuntu (v. 18.04) machine with 252 GB Memory and Intel(R) Xeon(R) Gold 5118 CPU with 48 cores.

Data cleansing: Duplicates, non-English and non-German entries and entries consisting solely of names, or of a single string (e.g., Amen), or solely of URLs or images were removed. The resulting cleaned dataset consists of 5,001,654 text-only message samples, see Table I.

Our 2-phase labeling strategy: Our post annotation strategy consists of two phases: i) *Labeling by human experts*, i.e., creating an initial small-sized ground truth as result of a manual data labeling process with the help of a group of "experts" humans; and ii) *Labeling through self-training*, i.e., extending our initial ground truth employing self-training [4] in a semi-supervised learning [5] fashion. Specifically, the first annotation phase entails two rounds of experiments in which a total of 16 human experts were asked to assign regrets labels to messages sampled from our raw dataset. To assist our experts and reduced the chance for human errors, we designed and implemented a web tool - see Figure 1. The resulting ground truth consists of 1,915 messages from the first experiment and 1,180 messages from the second. This output was achieved with high inter-annotator agreement, with Fleiss' kappa coefficients of 0,744 and 0,745, respectively. The second phase of our labeling strategy aims to overcome the limitation of labeling by human experts - the lack of human resources to manually

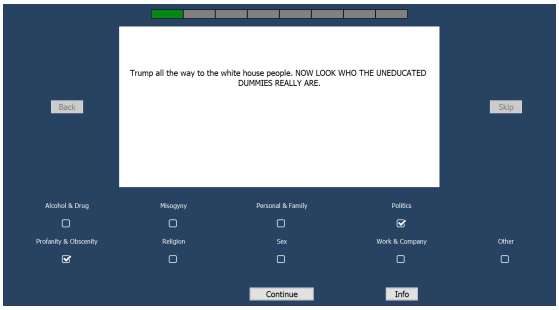

Figure 1: Standard user interface of our labeling tool

label a large raw dataset. Here, our initial ground truth (3095 expert-labeled items) is used to iteratively train and evaluate several multi-label classifiers in a semi-supervised manner on 11 subsets of our raw data. Classification maximation (CM) by Read et al. [2] in combination with domain-specific word embeddings as features and five off-the-shelf classifiers constitutes the general framework for our self-training strategy. At each iteration, the best classifier in combination with domain-specific word embeddings obtained by incrementally training Word2vec [1] on our growing corpus of regrettable messages, are used to predict labels of posts from the remaining raw dataset. The successfully labeled messages are then added to the previously labeled items, hence creating the next set of training data. This step is repeated until every message m_i in our raw dataset $\mathcal{D}$ is assigned a distribution of labels $T_{m_i} \in \{0,1\}^{|\mathbb{T}|}$ with $\mathbb{T} = \{T_1, T_2, \cdots, T_8\}$. For this study, we rely on five well established algorithms, namely AdaBoost (AB), Decision Tree (DT), Random Tree (RT), Random Forest (RF) and SVM.

III. Outcomes

Based on the aforementioned approach, we created a corpus of 4,700,284 messages with rich annotations based on a set of empirically validated regret-labels. During our self-training driven corpus generation experiments, we capture various performance metrics, including *accuracy*, *precision*, *recall*, and *weighted F1-Score*. The performances of all five classifiers on the last subset of raw data $\widetilde{\mathcal{D}}_{11}$ are summarized in Table II.

References

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[2] J. Read, P. Reutemann, B. Pfahringer, and G. Holmes. Meka: A multi-label/multi-target extension to weka. *J. Mach. Learn. Res.*, 17(1):667–671, Jan. 2016.

[3] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. 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*, page 10. ACM, 2011.

[4] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, 1995.

[5] X. J. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005.