ReDoS-M: A Dataset of Multi-Label Regrettable Disclosures on Social Media

Warning: This paper contains social media content that may be offensive or upsetting.

Hervais Simo, Michael Kreutzer and Javor Nikolov
Fraunhofer SIT, Darmstadt, Germany

Keywords:
(Self-)disclosures, social media, corpus, privacy, transformer, multi-label classification

Abstract:
Research on automated detection of regrettable disclosures in online social networks is limited by the lack of large-scale, semantically rich, and fine-grained annotated resources. Existing datasets often provide narrow coverage and conflate regret with related phenomena such as toxicity or hate speech, hindering robust modeling of regret-specific cues. To address these limitations, we introduce ReDoS-M, a large-scale, multi-source corpus constructed via a hybrid annotation pipeline that combines crowd-sourced labeling, transformer-based self-training, and enrichment with Sentiment–Moral–Emotion (SME) features. Starting from a collection of more than 5.5M user-generated posts and comments gathered from platforms such as Reddit and X (formerly Twitter), we derive four complementary corpora ranging from 4.27M to 5.13M annotated instances, reflecting different annotation and label-fusion strategies. We evaluate ReDoS-M in terms of label coverage and downstream utility by training six transformer-based models (DeBERTa and XLM-RoBERTa variants, with and without SME and Large Language Model-generated features). Across corpora, models achieve strong performance, with micro-F1 scores exceeding 0.98 and AUC values above 0.99 in the best settings, demonstrating that ReDoS-M supports effective and generalizable regret detection. Overall, ReDoS-M constitutes a comprehensive and scalable foundation for advancing research on fine-grained modeling and classification of regrettable disclosures in social media environments.

1 INTRODUCTION

Regrettable disclosures represent a consequential form of self-disclosure on online social networks (OSNs) with implications for privacy, safety, and well-being [Hu, 2013, Snow, 2015, Kaur et al., 2016, Xu et al., 2013, Stern, 2015, Wang et al., 2011, Sleeper et al., 2013b, Calo, 2011, **?**]. Prior work, e.g., [Wang et al., 2011] and [Sleeper et al., 2013a] shows that regret-inducing content spans a broad spectrum of social, emotional, and moral contexts and often arises from impulsive expression, heightened affect, or misjudgment regarding audience expectations. Accurate identification of such regrettable disclosures is therefore critical both for understanding online behavioral risks and for developing assistive, user-centric privacy mechanisms on OSN. Yet despite increasing research attention, progress in automated regret detection is fundamentally constrained by the absence of large, high-quality, and semantically rich corpora that adequately capture the diversity and nuance of regret-inducing content. Existing corpora are typically small, thematically narrow, or confined to specific platforms, which limits the generalizability of regret detection models and hinders consistent insights into both regret-inducing and regret-preventing behaviors across diverse online environments.

To address these shortcomings, we introduce ReDoS-M, a comprehensive and substantially expanded version of the REGRET dataset originally proposed by Simo et al. [Simo and Kreutzer, 2022a]. ReDoS-M is built from a large-scale collection of user-generated social media messages, relying on a new annotation approach that combines manual annotation with an automated, multi-stage labeling approach driven by transformer-based self-training. ReDoS-M is enriched with both Sentiment Moral Emotion (SME)–derived features and Large Language Model (deepseekai/ DeepSeek-V3)–generated annotations. This enriched feature space aims at increasing the granularity and expressiveness of re-

gret annotations and enables a more precise characterization of regret-inducing OSN disclosures. Our contributions are the following:

1. We propose ReDoS-M, a feature-enriched extension of the REGRET dataset. ReDoS-M comprises four multi-label corpora of regrettable disclosures, each derived from over 5M OSN messages, and constructed through distinct annotation and label-fusion strategies. Through the integration of SME-derived features and LLM-generated descriptions, ReDoS-M offers significantly enhanced granularity for modeling the linguistic, affective, and moral dimensions of regret-inducing disclosures.
2. A comprehensive analysis of dataset coverage and distribution. We systematically characterize the linguistic, emotional, and moral properties of ReDoS-M, including message-length distributions, sentiment and emotion profiles, and moral foundation attributes. We further analyse coverage across regret categories and multi-label combinations, demonstrating the high diversity and representational depth of each of the four corpora.
3. A large-scale downstream evaluation of ReDoS-M. We train and evaluate six transformer-based models (text-only and feature-augmented variants) on each ReDoS-M corpus. Using standard multi-label metrics (precision, recall, micro-/macro-F1, AUC), we assess the suitability of each ReDoS-M corpus for building effective models for regrettable disclosure detection. Our results show that SME and LLM-derived features substantially enhance downstream performance and that ReDoS-M enables reliable modeling even for minority regret categories.

Overall, these contributions establish ReDoS-M as the most comprehensive and contextually enriched dataset currently available for studying regrettable disclosures on OSN. By combining large-scale data collection, expert-informed features, LLM-driven contextualization, and transformer-based automated labeling, ReDoS-M provides a robust foundation for advancing research on regrettable disclosures on OSN and computational tools for detecting and mitigating such harmful self-disclosures.

2 RELATED WORK

Research on regrettable disclosures in OSNs spans psychology and HCI studies of regretful experiences [Guo et al., 2025, Xie and Kang, 2015, Wang et al., 2011, Sleeper et al., 2013b], privacy harms stemming from online self-disclosure [Hu, 2013, Snow, 2015, Kaur et al., 2016, Xu et al., 2013, Stern, 2015, Calo, 2011, **?**], and computational approaches aiming to detect or predict regrettable content [Khan et al., 2025, Diaz Ferreyra et al., 2023, Balouchzahi et al., 2023, Acquisti et al., 2017, Zhou et al., 2015, Wang et al., 2013, Simo and Kreutzer, 2022b, Simo and Shulman, 2021, Wang et al., 2019]. A consistent finding across this literature is that regret often emerges from cognitive and contextual barriers (e.g., limited awareness of audience reach, misleading platform cues, and complex privacy settings) which shape disclosure decisions and post-hoc regret. Complementing this line of work, computational research has sought to operationalise regret as a detectable signal in user-generated text, developing models that identify regrettable disclosures either through direct manual annotation or by inferring regret signals from behavioral cues such as post deletion. Research on constructing corpora of "regrettable" OSN disclosures has developed along two main lines: (1) deletion-based corpora, treating removals as a proxy for regret, and (2) manually annotated corpora, enabling semantic regret labeling beyond what deletion signals can capture. Deletion-based studies provide scale but suffer from label noise because deletions include typos, spam, policy enforcement, or account changes unrelated to regret. While these efforts yield valuable insights into linguistic and contextual cues of regrettable disclosures, they also expose persistent gaps, particularly w.r.t. to the availability and quality of datasets suitable for training robust detection models.

Almuhimedi et al. [Almuhimedi et al., 2013] provided one of the earliest large-scale empirical characterizations of tweet deletion, motivated by the mismatch between Twitter's perceived ephemerality and the permanence of posted content. Using the Twitter Streaming API, they tracked a large cohort of English-speaking users for one week in March 2012 and collected 67.2M tweets, of which 1.6M (2.4%) were deleted. Deletions were common at the user level (roughly half of users deleted at least one tweet), and the authors distinguished superficial causes (e.g., typos,

spam removal) from potentially substantive ones. While foundational for quantifying deletion behavior as a privacy and self-presentation mechanism, the study did not explicitly identify which deletions were *regrettable*, and the dataset was not reported as publicly available. Petrović et al. [Petrovic et al., 2013] showed that deletions can be predicted to a limited extent, reinforcing deletion as a behavioral signal. From over 75M tweets collected in January 2012 (with deletion notices observed through February 2012), they labeled 2.4M tweets (3.2%) as deleted and framed deletion prediction as binary classification. Linguistic and behavioral correlates (e.g., profanity, first-person pronouns, celebrity/high-follower accounts) were associated with higher deletion likelihood. Classical classifiers (SVM, logistic regression, naïve Bayes) achieved modest performance (best F1 $\approx$ 0.27), highlighting both predictive signal and the heterogeneity of deletion causes; the corpus was likely not released due to restrictions on redistributing deleted content. Moving beyond treating all deletions as equally meaningful, Zhou et al. [Zhou et al., 2016] targeted *content-identifiable* regret signals. They they first collected 1.25 million tweets in May 2014, of which 440,000 were deleted, and then extended the collection window to two months, yielding 17.6 million tweets, including 3.2 million deletions. They filtered noisy accounts via user clustering to retain "normal" users. Manual inspection of deleted tweets produced a taxonomy of ten regret-related categories (e.g., negative sentiment, cursing, sex, alcohol/drugs, violence, health, race/religion, job, relationships). Only a minority of deletions (about 18%) fit these content-identifiable regret categories, and the distribution was highly skewed toward a few dominant types. Although the full tweet corpus was not released, the authors published lexicons enabling partial reproducibility. Subsequent work expanded the focus from *what* counts as regrettable to *when* and *under which conditions* deletions occur. Minaei et al. [Minaei et al., 2020] framed certain deletions as privacy-damaging and constructed a small annotated dataset (combined with re-labeled items from #DontTweetThis [Wang et al., 2019]) to study adversarial risks: damaging deletions were detectable with strong performance, but deceptive deletion strategies substantially degraded accuracy. Díaz Ferreyra et al. [Diaz Ferreyra et al., 2023] examined "self-cleaning" during COVID-19 by collecting health-related self-disclosures in early 2020 and later labeling tweets as deleted vs. not deleted as a regret proxy; benchmark results showed BERT outperforming classical models, and tweet IDs were released under ethical approval. Similarly, Xu et al. [Xu et al., 2013] tracked bullying-related tweets over time and used deletion as a regret proxy, finding only moderate predictive performance, underscoring that text alone captures limited regret-driven deletion signal. Wang et al. [Wang et al., 2019] addressed pre-publication risk via the #DontTweetThis dataset, created by lexicon-based filtering of a large crawl and crowd annotation of sensitiveness (later consolidated into fewer levels). The work produced keyword resources and enabled training of privacy-risk scoring systems (PrivScore), but full data availability remains constrained.

Recent corpora move beyond deletion proxies toward explicit regret annotation. Simo and Kreutzer [Simo and Kreutzer, 2022a] introduced REGRETS, a large-scale multi-label corpus constructed from a small expert-labeled seed set expanded via iterative self-training; benchmarking indicated that regret categories can be learned at scale, though full release is limited by privacy constraints. Complementarily, Balouchzahi et al. [Balouchzahi et al., 2023] presented ReDDIT, a smaller but carefully curated Reddit benchmark with explicit regret labels (regret by action/inaction/no regret) and domain annotations, achieving strong inter-annotator agreement. These datasets illustrate the spectrum of methodological trade-offs - deletion-based resources offer scale but confound regret with other deletion causes, whereas explicitly annotated corpora improve conceptual precision but are typically smaller and more costly to construct. ReDoS-M extends REGRETS [Simo and Kreutzer, 2022a] along three axes. First, ReDoS-M substantially broadens the scope and diversity of user-generated regrettable text content. In addition to REGRETS' multi-platform crawl, ReDoS-M integrates four public datasets - alcohol usage detection [Hossain et al., 2016] (6,513), profanities/personal attacks [Wulczyn et al., 2017] (115,864), German anti-refugee tweets [Ross et al., 2016] (469), and Kaggle insults [Kaggle.com, 2012] (8,829)—and adds 740,267 unannotated disclosures from Deleted Tweets Archive, GermEval2021 [Risch et al., 2021], and toxic spans resources [Pavlopoulos et al., 2021]. To balance coverage, we also add 261 neutral texts (167 EN, 94 DE) generated via GPT-4 (Mar 14, 2023). Second, we created ReDoS-M by refining the annotation pipeline

used for REGRETS. ReDoS-M replaces convenience sampling with experienced AMT crowdworkers for the seed ground truth and uses transformer-based self-training (DeBERTa Text-Only / Text+SME) instead of traditional ML, aligning with evidence that transformers outperform earlier neural and classical models across NLP tasks [Mienye et al., 2024, Rogers et al., 2020, Lin et al., 2022, Clark et al., 2020, Otter et al., 2020]. Third, ReDoS-M extends the annotation scheme beyond topical regret labels by introducing annotations for sentiment, emotion, and moral foundations. Beyond topical regret labels, ReDoS-M adds sentiment, emotion, and moral foundations, enabling multi-dimensional analysis of regret signals.

3 REDOS-M DATASET

Our ground-truth generation follows a three-stage process: (i) keyword-based raw data collection, (ii) extensive data cleansing, and (iii) a hybrid two-phase labeling strategy. Central to this process is a multi-label annotation framework that integrates human expertise with automated self-training. A representative subset of messages is first manually annotated by trained crowdworkers to form a high-quality seed corpus. This seed set is then iteratively expanded via semi-supervised self-training, enabling scalable label propagation to millions of unlabeled instances while maintaining annotation consistency. This hybrid approach combines the reliability of human judgment with the scalability required for larger datasets.

3.1 Raw Data Gathering

We collected publicly available text-only posts and comments from over 30 online platforms and more than 200 distinct data sources, including major social media platforms (e.g., Facebook, Twitter), discussion and self-help forums (e.g., Reddit, 4chan), YouTube channels, and news websites such as CNN, MSNBC, and Breitbart. These platforms contain dynamic, jargon-rich, and often short-form disclosures, which frequently include impulsive or potentially regrettable content [Hicks and Gasca, 2019, Breland, 2019, Arthur, 2019]. Such a diversity in sources and linguistic styles is of paramount importance to ensure a high-quality ReDoS-M.

Regret-related keyword lists. Data collection was guided by eight category-specific keyword lists corresponding to the primary regret categories T0–T7: Alcohol and Illegal Drug Use, Sexism and Misogyny, Personal and Family, Politics, Profanity and Obscenity, Religion, Sex-related Content, and Work and Company. These categories are derived from empirical studies of regret on OSNs [Wang et al., 2011, Sleeper et al., 2013b] and are summarized in Table 1. Additional labels capture residual content (T8: Others) and overall regret judgments (T9: Regrettable, T10: Non-Regrettable). Keywords were compiled in both English and German using Princeton WordNet [Miller, 1995], ODeNet [Siegel and Bond, 2021], and curated lexicons of insults, profanity, and hate speech [Ross et al., 2016], supplemented with hand-engineered domain knowledge. Our lists of keywords are publicly available at: `https://bit.ly/3InQn7W`.

Collection procedure. We used a combination of automated and manual crawling. Automated routines relied on extended versions of TweetScraper (`https://github.com/jonbakerfish/TweetScraper`) and PRAW (`https://praw.readthedocs.io/en/latest/`) to mitigate API rate limits, while manual scraping was applied where APIs were unavailable. For each identified source, we collected up to 100 posts and the top 50 comments per post. Data collection occurred in three waves. The first wave (Nov 2016–Feb 2017) yielded 1,122,728 posts; the second (May–Aug 2018) added over 4M posts. To improve coverage and mitigate class imbalance, we injected 131,675 messages from four widely used datasets: alcohol usage detection [Hossain et al., 2016], profanities and personal attacks [Wulczyn et al., 2017], German anti-refugee tweets [Ross et al., 2016], and the Kaggle insults dataset [Kaggle.com, 2012]. The third wave (May 2022) added 740,267 unannotated disclosures from the Deleted Tweets Archive [Salish-Coast and Karma, 2021], GermEval 2021 [Risch et al., 2021], and toxic spans datasets from ACL 2022 and SemEval 2021 [Pavlopoulos et al., 2021]. The collected texts, 6,160,367 in total, were subsequently cleaned and annotated, resulting in the basis of the ReDoS-M dataset.

Data Cleansing Given the heterogeneity and noise inherent in OSN data, extensive preprocessing was required. First, we applied a keyword-density heuristic: messages containing fewer than 20% regret category-specific keywords were removed. Second, we eliminated entries consisting only of names, URLs, images, or residual metadata, as well as retweets to avoid duplication bias.

Third, we removed non-English and non-German texts using a strict consensus on outputs from three distinct language identification tools - langdetect (`https://pypi.org/project/langdetect/`), langid (`https://pypi.org/project/langid/`) and polyglot (`https://pypi.org/project/polyglot/`). That is, we retained only messages classified identically by all three tools. Spam messages were filtered using techniques from [Dhingra and Mittal, 2015]. A final quality check helped to remove remaining duplicates and atypically long messages exceeding 450 tokens. We employed langid.py (`https://github.com/saffsd/langid.py`); Fasttext models for language identification (`https://fasttext.cc/docs/en/language-identification.html` [Joulin et al., 2016]), and langdetect (a port of Google's language-detection library to Python - `https://github.com/Mimino666/langdetect`) for additional language filtering. The resulting dataset comprises approximately 5.5M unique text-only messages, 5,177,133 in English and 323,683 in German. This cleaned corpus forms the basis for subsequent human and automated annotation.

Additional Dataset Statistics. Prior to annotation, we conducted three descriptive analyses to demonstrate the quality of the collected raw dataset. Message length analysis using SpaCy v3.2.3 shows an average of 34 tokens for English texts (median 24) and 26 tokens for German texts (median 21), with substantial variance across categories and sources. Sentiment analysis was performed using Python-SentiStrength [Thelwall et al., 2010] with language-specific lexicons. English texts exhibit predominantly negative sentiment (mean –0.168), whereas German messages are near-neutral (mean +0.012). Emotion analysis followed Ekman's taxonomy [Ekman,]. For English, we used FlairNLP models trained on GoEmotions [Demszky et al., 2020]. For German, we used Germotion [Klinger et al., 2016]. Across both languages, joy is the most frequent emotion, while negative emotions such as anger, sadness, fear, and disgust are also highly prevalent, reflecting affective patterns commonly associated with regrettable disclosures.

3.2 Data Annotation

The final cleaned raw dataset, comprising approximately 5.5M text-only posts and comments, was annotated using a two-stage hybrid labeling strategy that combines human judgment with semi-supervised self-training. This approach balances annotation reliability with scalability and enables the construction of multiple high-quality, multi-label corpora of regrettable disclosures. In the first stage, a compact but reliable human-annotated seed corpus is created via crowdsourcing (Section 3.2.1). In the second stage, this seed ground truth is iteratively expanded using a self-training paradigm [Yarowsky, 1995,Zhu, 2005, Chapelle et al., 2009], enabling automated label propagation to millions of previously unlabeled instances (Section 3.2.2). Table 1 presents the set of regret categories and labels used in ReDoS-M, derived from empirical findings from Wang et al. [Wang et al., 2011] and Sleeper et al. [Sleeper et al., 2013b].

3.2.1 Manual Annotation

Annotation sample. Manual labeling was conducted on *AnnotationSet*, a curated subset of the final raw dataset, with 3,618 messages (2,383 English, 1,235 German). The *"AnnotationSet"* was constructed by combining highly polarizing user-generated messages with intentionally neutral content, under the assumption that polarizing texts exhibit strong sentiment, emotional, or moral signals, whereas neutral texts lack such cues and are therefore less likely to trigger controversy or regret. To coverage of both controversial and benign content, we first constructed *PolarizationSet* a larger subset by filtering the raw dataset according to sentiment, emotion, and moral signals. Messages to be included in this set were selected if they met at least one of the following conditions: (i) strongly negative or positive sentiment ($[-4,-3]$ or $[+3,+4]$); (ii) non-zero emotion probability; or (iii) non-zero moral polarity (difference between positive and negative moral foundations). Messages were grouped into four clusters based on how many criteria they satisfied (Clusters #0–#3). The AnnotationSet is then created as subset of the *PolarizationSet*, relying on a stratified approach based on category, source, language and cluster severity, with oversampling of highly polarizing content (Cluster #3) and controlled inclusion of neutral messages. To further balance non-regrettable content, 400 neutral texts (200 EN, 200 DE) were added, including 261 OSN-like neutral messages generated via zero-shot prompting using GPT-4.

Annotators & Annotation Routine. We designed a manual annotation procedure in which annotators labeled messages from the *AnnotationSet*. Annotators were recruited via Amazon Mechanical Turk using purposive sam-

Table 1: ReDoS-M's regret categories and labels.

Category	Label	Explanation
`Alcohol and Illegal Drug Use`	T0	Posts referencing alcohol or illegal drug use.
`Sexism & Misogyny`	T1	Content containing derogatory, objectifying, or implicitly discriminatory language toward women.
`Personal & Family`	T2	Disclosures about personal or family matters (e.g., illness, loss, relationship difficulties) that may reveal sensitive information.
`Politics`	T3	Posts expressing political opinions, which may provoke conflict or negative reactions.
`Profanity and Obscenity`	T4	Content containing profanity or obscene expressions, regardless of target or context.
`Religion`	T5	Posts expressing religious beliefs or discussing religious issues, often eliciting strong reactions.
`Sex-related Content`	T6	Sexual content that is explicit or implicit but not categorized as profanity or obscenity.
`Work & Company`	T7	Posts discussing workplace issues, colleagues, or internal company matters.
`Others`	T8	Regret-relevant content not captured by the above categories.
`Regrettable`	T9	Overall judgment that the post is regrettable.
`Non-Regrettable`	T10	Overall judgment that the post is not regrettable.

pling [Campbell et al., 2020]. Participation was restricted to U.S.-based workers with a HIT approval rate of at least 95% and more than 100 approved HITs, without further demographic filtering to avoid introducing selection bias. Annotators received a short briefing on the study goals and were compensated at an effective rate of $13 per hour. Annotation was conducted using a custom web-based tool hosted on a European server (Fig. **??**). Upon accepting the task, annotators reviewed detailed labeling guidelines, provided informed consent, and completed a short questionnaire capturing demographics, OSN usage, and prior experience with online regret. Multiple attention-check questions were embedded in the questionnaire to ensure data quality. Over a 24-day period (Dec. 13, 2023–Jan. 5, 2024), annotators labeled the English portion of the AnnotationSet, comprising 2,383 messages. Each message was annotated by multiple workers and labeled either as non-regrettable (T10) or with one or more regret categories (T0–T9). The annotation task followed a three-step scheme: (i) identification of a negative undertone, (ii) binary judgment of potential regret, and (iii) assignment of topical regret labels. This design enabled reliability assessment both across the full decision pipeline and for topical labels alone. In total, 19 workers initiated the task, of whom 7 produced valid submissions after quality filtering, i.e., approximately 63% were excluded due to incomplete surveys or failed attention checks. All study materials, including consent forms, annotation guidelines, and questionnaires, are available to reviewers upon request.

Output. In total, 2,383 English-language messages were manually annotated, resulting in 16,674 individual judgments. Final labels were determined using majority voting, requiring agreement from at least four out of seven annotators per label. Messages that failed to meet this threshold were discarded. The resulting consensus annotations constitute the seed ground-truth corpus $\mathcal{D}_0$ used in subsequent processing. On average, each message received 1.86 labels (SD = 0.54; median = 2). The corpus is nearly perfectly balanced, with 49.85% of messages labeled as regrettable and 50.15% as non-regrettable. Items in $\mathcal{D}_0$ typically contained one or two topical regret labels, with profanity, politics, religion, and personal/family disclosures among the most frequent categories. The average undertone (negativity) score across the corpus is 0.66 (SD = 1.09), indicating a mild overall negative tone. Inter-annotator reliability was assessed using both percent agreement and Fleiss' Kappa. Across all labels, annotators achieved a pooled percent agreement of 70.17% and a Fleiss' Kappa of 0.501, corresponding to a moderate level of agreement according to Landis and Koch's interpretation scale [Landis and Koch, 1977]. Agreement varied across labels, with higher consistency for non-

regrettable decisions (e.g., w.r.t. negativity scoring) and lower, but still meaningful, agreement for regret categories. Given the inherent ambiguity and noisiness of social media text, these results indicate that the resulting ground truth is sufficiently reliable and of high quality for downstream modeling and large-scale self-training.

3.2.2 Automated Labeling

To overcome the scale and reliability limits of manual annotation for noisy social media text, we employ an automated, semi-supervised *self-training* pipeline to propagate regret labels from a small seed ground truth to millions of unlabeled posts. Manual annotation is costly and typically yields limited coverage; moreover, linguistic ambiguity often prevents strong annotator consensus, and models trained on small ground-truth sets may overfit. Self-training is a widely used semi-supervised paradigm [Amini et al., 2022]that has proven effective in a range of NLP applications (e.g., summarization and NER) [He et al., 2019, Amini and Gallinari, 2002, Liao and Veeramachaneni, 2009].

Overview. Starting with the seed corpus $\widetilde{\mathcal{D}}_0$ (2,383 human-labeled items; with half of the non-regrettable items not considered), we iteratively train and apply two transformer-based multi-label classifiers: DeBERTa-Text-Only and DeBERTa-Text+SME. In each iteration, the current best-performing model variants label previously unseen raw data; high-confidence predictions are added to the labeled pool, yielding an expanded ground truth that is then used for retraining and refinement. This iterative loop continues until the full raw dataset has been processed. To ensure stable execution under hardware constraints, the raw pool $\mathcal{D}$ is partitioned into two disjoint subsets, $\mathcal{D}_1$ and $\mathcal{D}_2$, which are labeled sequentially. The automated routine begins with an *initialization phase*, where both classifiers are trained and validated on $\widetilde{\mathcal{D}}_0$ using k-fold cross-validation. It then enters the *iteration phase*. For each subset $\mathcal{D}_i$ ($i \in \{1, 2\}$), optimized model variants generate label distributions for all messages. Items meeting confidence requirements form $\mathcal{D}_i^+$ and are incorporated into the labeled pool: $\widetilde{\mathcal{D}}_i = \widetilde{\mathcal{D}}_{i-1} \cup \mathcal{D}_i^+$. Items not meeting the confidence threshold, $\mathcal{D}_i^-$, are not discarded; instead, they are deferred and re-evaluated with the next subset. Unlike the human-annotation stage (which uses the overarching regret flag T9), the automated pipeline predicts only the fine-grained labels T0-T8 and T10 (non-regrettable), ensuring consistent focus on granular regret types. All experiments were conducted on a workstation with a 16-core CPU, 128 GB RAM, and an NVIDIA GeForce RTX 3090 (24 GB VRAM).

Underlying Transformer Architecture. Both self-training models share the same backbone architecture. The *input layer* receives preprocessed text. In Text+SME settings, each message is augmented with SME signals that are converted into textual SME labels (sentiment polarity, dominant emotion, and moral foundation dimensions) based on mapping rules proposed in [Simo et al., 2025]. A pretrained DeBERTa-v3-base encoder serves as the *feature extraction layer*, producing contextual token representations. The *pooling layer* (ContextPooler with GeLU) compresses the final hidden state (via the [CLS] token) into a sentence-level embedding, followed by a *dropout layer* (default $p = 0.1$). The *classifier head* projects the pooled embedding to 10 sigmoid outputs corresponding to T0-T8 and T10, enabling independent multi-label probabilities per class.

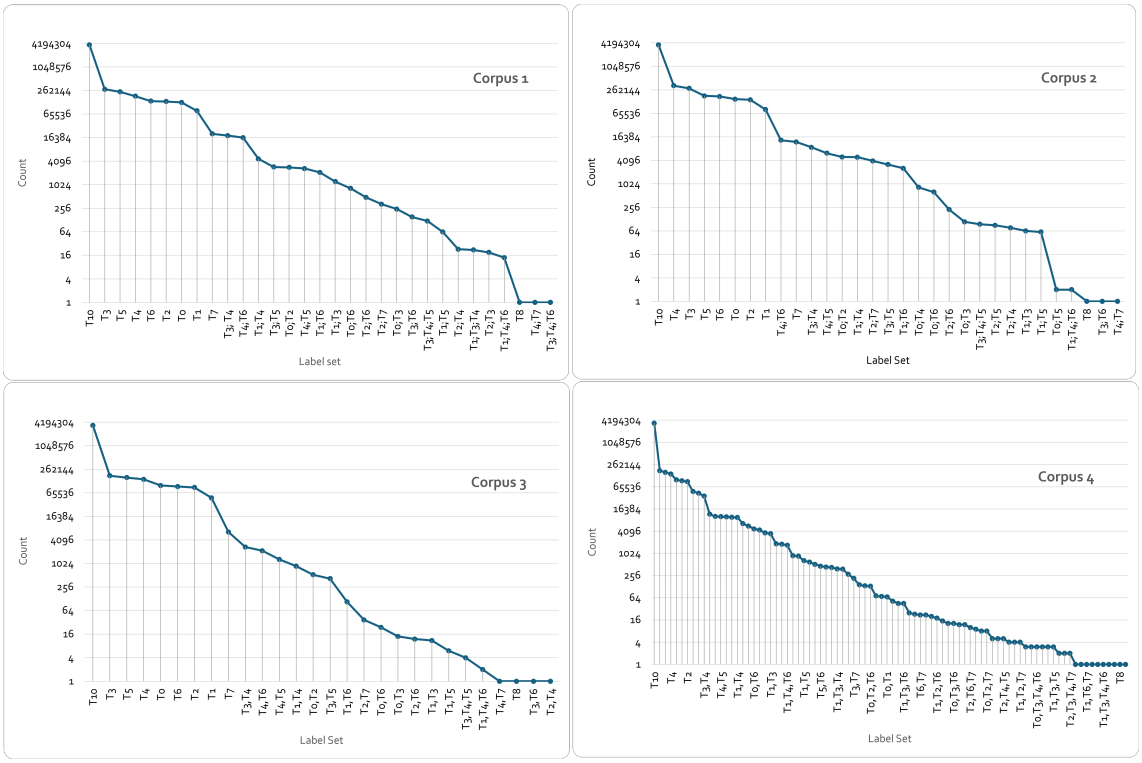
Self-Training Pipeline. Our automated data annotation pipeline consists of three stages.

Step 1: Data preprocessing. Because OSN texts are noisy and heterogeneous, each message is normalized prior to tokenization. Emojis and emoticons are converted to textual descriptors (e.g., CLDR short names), repeated-character sequences are reduced (more than two occurrences mapped to two), and stopwords, punctuation, special characters, user mentions, and HTML artifacts are removed. URL prefixes (e.g., *https://*) are stripped while retaining the domain string. The resulting text is lemmatized and tokenized using Hugging Face's DeBERTa tokenizer; special tokens ([CLS], [SEP]) and attention masks are created. Inputs are padded/truncated to a maximum length of 512 tokens. In the Text+SME settings, SME labels are concatenated to the text using [SEP] separators.

Step 2: Feature extraction. Tokenized sequences are encoded with DeBERTa-v3-base to obtain contextualized hidden states, which are pooled and passed to the classifier head.

Step 3: Training, evaluation, and iterative expansion. At $i = 0$, both models are trained on $\widetilde{\mathcal{D}}_0$ using k-fold cross-validation to maximize data usage under limited supervision. The best-performing variants are then used to label $\mathcal{D}_1$, producing expanded labeled sets. From round $i \geq 1$ onward, as the corpus becomes

Figure 1: Distribution of labels in ReDoS-M corpora.



sufficiently large, we use a standard 80/10/10 train-validation-test split (with label-distribution preservation). Training is performed epoch-wise with binary cross-entropy loss for multi-label prediction, Adam-based optimization [Kingma, 2014], and early stopping triggered by a plateau in validation loss. Hyperparameters are optimized via Population-Based Training (PBT) [Jaderberg et al., 2017]. The best model variants then label the next raw subset, and confidently labeled items are merged into $\widetilde{\mathcal{D}}_i$. Items that remain unlabeled after the final pass are excluded from the released corpora.

3.2.3 Final corpora

This process yields two primary corpora: **Corpus 1** (5,132,463 items) produced by the Text+SME pipeline, and **Corpus 2** (5,128,577 items) produced by the Text-Only pipeline. We further derive two additional corpora through label fusion of Corpus 1 and Corpus 2. **Corpus 3** (4,267,865 items) results from *strict fusion* strategy, i.e., an item is retained only if it appears in both corpora and both models assign an identical label set. On the other hand, **Corpus 4** (4,436,164 items) considers a *relaxed fusion*, i.e., items are retained if they appear in both corpora and there is no disagreement on T10 (i.e., cases where one model predicts T10 and the other predicts a regret label are discarded). This preserves consistent regret versus non-regret separation while allowing disagreements among fine-grained regret labels, resulting in a broader multi-label coverage than Corpus 3. Figure 1 summarizes the label distributions in each of the four final corpora. The self-training outputs (Corpus 1 and Corpus 2) and the two fusion-derived resources (Corpus 3 and Corpus 4) together form the ReDoS-M corpora. Table 2 provides some sample texts from the ReDoS-M dataset. We refer to the extended version of our paper [Simo et al., 2025] for supplementary artifacts documenting pipeline behavior, including hyperparameter logs, per-epoch losses, and evaluation summaries for all training, validation, and test phases. ReDoS-M and all supplementary resources are available online at (access for reviewers available upon request):[`https://shorturl.at/SmRcP`] .

Table 2: Sample texts from ReDoS-M dataset

Text	Sample
my younger brother is addicted to benzos and this comment just made me so fucking sad. it's real shit.	T0, T2
@user face_with_tears_of_joy you do not like the jewish people because you all claims they are holding all the money. sound familiar? your left wing libtard family is the family of the kk. you do not hate jews? that's funny. gun control; even funnier. good luck	T3, T4, T5
... also, you look like a 15. why would anyone fuck you? gosh. fucking stupid sluts.	T1, T4, T6
really i hope she get raped in the future. raping women should be legal i think	T1, T4, T6
this is so sinful. it says it in the bible you dumb fucktards. stop saying you will go to heaven if you are gay. because you will not go to heaven. trust me its a fucking choice my gf dated girls and i was her 2nd bf she dated like 12 girls and she is 100% straight. you can change to the right side	T4, T5

4 EVALUATION

We evaluate the quality and practical utility of the four ReDoS-M corpora along two complementary dimensions: i) *Dataset Coverage and Distribution*, which assesses representativeness, balance, and multi-label diversity of regret categories, and ii) *Downstream Utility*, which measures how effectively the corpora support training robust multi-label regret classifiers. Together, these analyses characterize both the intrinsic structure of the ReDoS-M corpora and their suitability for automated regret detection.

4.1 Dataset Coverage and Distribution

We analyze the distribution of regret labels T0–T8 and their co-occurrence patterns across the four corpora to assess coverage, diversity, and balance. Although all corpora are dominated by the non-regrettable class (T10), substantial differences emerge in minority-class representation and multi-label richness. Corpus 1 provides strong coverage of major regret categories, with T4 (profanity/obscenity) and T3 (politics) being most frequent, followed by T5 (religion), T6 (sex), T0 (alcohol/drugs), and T2 (personal/family). Minority categories such as T1 (misogyny) are underrepresented, while T7 (work/company) and T8 (other) are rare. Multi-label diversity is present but limited, with dominant overlaps such as T4;T6 and T3;T4. Corpus 2 exhibits broader and more balanced coverage. While T3 and T5 dominate, minority classes (notably T7) are better represented than in Corpus 1. Importantly, Corpus 2 contains substantially richer multi-label structures, including several frequent label pairs and a small number of triple-label combinations, reflecting more complex regret expressions. On the other hand, Corpus 3 prioritizes annotation consistency by retaining only instances with identical labels from both classifiers. This results in reduced coverage of minority classes and sharply diminished multi-label diversity. While major categories remain well represented, overlaps are an order of magnitude smaller than in Corpora 1 and 2. Corpus 4 combines the strengths of Corpora 2 and 3. It preserves the high label consistency of Corpus 3 while reintroducing extensive multi-label diversity, including numerous pairwise, triple, and quadruple overlaps. As a result, Corpus 4 offers the most balanced trade-off between annotation reliability and expressive richness, making it particularly well suited for modeling complex regret phenomena.

4.2 Downstream Utility

To assess downstream utility, we train and evaluate six transformer-based multi-label classifiers on each corpus. The models vary along three axes: encoder architecture (DeBERTa versus XLM-RoBERTa), inclusion of SME features, and integration of LLM-derived regret-relevant features.

4.2.1 Model Architectures

Models 1 and 2 integrate LLM-derived features extracted by prompting deepseek-ai/DeepSeek-V3 with the input text to identify potentially

sensitive topics, strong sentiment, lies, secrets and potentially for regret. These features are fused with the original text and SME labels before encoding. Across all models, tokenization is handled via Hugging Face AutoTokenizer with a maximum length of 512 tokens, truncation and padding enabled. Models 3 and 4 omit LLM features and rely solely on text and SME augmentation. Models 5 and 6 are text-only models without any auxiliary features. Encoders are based on `deberta-v3-base` or `xlm-roberta-base`, each followed by a classification head producing independent probabilities for all 10 regret labels via sigmoid activation.

4.2.2 Training and Evaluation Protocol

For each corpus, we use an 80/10/10 train–validation–test split. Models are trained with AdamW and BCEWithLogitsLoss, using a linear learning-rate schedule with warm-up. Validation is performed after each epoch, and early stopping is applied via a plateau-based criterion. Hyperparameters are optimized using PBT across parallel trials. The final evaluation is reported on the held-out test split using micro/macro/weighted F1, AUC, and related multi-label metrics.

4.2.3 Comparative Results

Across all corpora, all six models achieve strong performance, indicating that ReDoS-M supports reliable and learnable regret classification. Performance varies systematically with corpus characteristics, with Models 1, 3, and 5 typically performing best. Corpora 3 and 4 yield the highest scores, consistent with their higher label agreement and/or refined fusion strategy. Indeed, Corpus 3 yields the highest overall scores. Several models achieve micro-F1 values between 0.98 and 0.99 with AUCs approaching 1.0 and extremely low Hamming loss. This confirms that the strict-consensus annotations in Corpus 3 produce a particularly clean and internally consistent learning signals. Corpus 4 also demonstrates excellent performance, with the best models achieving micro-F1 scores around 0.97-0.98 and AUC values above 0.98. Compared to Corpus 3, macro-F1 scores here are slightly lower, reflecting increased linguistic and semantic diversity, yet overall generalization remains strong. In contrast, Corpora 1 and 2 show slightly lower, but still robust, performance. Best models reach micro-F1 scores in the mid-0.95 range with consistently high AUC values (>0.97). However, macro-F1 is noticeably lower ($\approx$ 0.80), indicating that minority labels such as T7 and T0 are slightly harder to predict. These effects are more pronounced in Corpus 1. The divergence between micro- and macro-averaged metrics across corpora underscores the potential impact of class imbalance. While dominant classes (especially T10) are classified with near-perfect accuracy, minority regret categories remain indeed more challenging. Nonetheless, even these challenging classes achieve reasonably strong F1 scores in the best-performing models (often between 0.76 and 0.90 for T7), indicating that they are not random or incoherent, but rather more difficult and thus more sensitive to corpus size and annotation consistency. Overall, the evaluation demonstrates that all four ReDoS-M corpora support high-quality automated regret detection. Corpus 3 emerges as the cleanest benchmark resource, while Corpus 4 offers the best balance between annotation reliability and expressive diversity. Corpora 1 and 2, although more heterogeneous, remain valuable for studying regret in realistic, large-scale social media settings. However, the lower macro-F1 in Corpora 1 and 2, especially for rare labels (e.g., T7), highlights concrete opportunities for future improvements in data collection and annotation strategies aimed at increasing minority-class coverage and consistency. Additional details regarding the quality and practical utility of the ReDoS-M corpora are available in [Simo et al., 2025], the extended version of this paper.

5 LIMITATIONS & FUTURE WORK

While our research demonstrate strong empirical results, our approach comes with several limitations that must be acknowledged.

Raw data collection. Our raw data were retrieved using regret-related keyword queries, which introduces two forms of bias. First, keyword popularity varies by topic, producing uneven coverage across regret categories and contributing to pronounced label imbalance. Training on such distributions risks reinforcing topic-linked majority/minority associations. Second, keyword retrieval misses regrettable disclosures that lack explicit lexical markers, excluding relevant content whose surface forms do not match our predefined terms. Hence, although the approach ensure a high concentration of regret-related content, the resulting raw data may not

reflect the true prevalence or distribution of regret on typical OSN platforms where regret is comparatively rare. Future work will explore diversified data sources and hybrid retrieval strategies that combine keyword search with semantic retrieval to reduce topical and lexical bias.

Annotation strategy. Our two-phase annotation strategy introduces additional limitations. The manual seed corpus, labeled via crowdsourcing, is inherently sensitive to variability in annotator expertise, cultural background, and attention, and thus is typically noisier than expert annotation despite quality controls. The subsequent semi-supervised self-training phase scales the corpus efficiently, but can propagate errors and biases from the seed set. Indeed, annotation errors or biases introduced early may persist across iterations, affecting the reliability and interpretability of the final corpora. Future work will investigate hybrid annotation pipelines combining crowd-sourcing with expert validation, active learning to focus human effort on ambiguous cases, and periodic re-annotation to detect label drift and bias.

Self-training models. Although both transformer-based self-training models perform well overall, they struggle with low-support classes, especially during early training stages. The small evaluation size per fold also inflates variance: single errors can substantially distort per-class scores. On $D^{+}_{1,\text{Only}}$ and $D^{+}_{1,\text{SME}}$, performance is dominated by well-represented classes, leading to high micro-F1 scores but substantially lower macro-F1 (gaps up to $\approx$0.18), indicating persistent weaknesses for rare and sensitive categories. This is critical because rare classes may be the most important to detect in practice; relying on micro-averaged metrics risks masking systematic blind spots. Future work will address these limitations with (i) class-balanced optimization (e.g., class-balanced focal loss [Cui et al., 2019]), (ii) targeted data augmentation for minority classes [Shamsolmoali et al., 2021], and (iii) sampling strategies that emphasize underrepresented cases [Cai et al., 2022], aiming to improve robustness and equity across regret categories.

ReDoS-M corpora. Each ReDoS-M corpus reflects trade-offs inherent to its construction. Corpus 1 and Corpus 2 reflect the model-dependent limitations discussed above, including uneven minority-class coverage. Corpus 3, derived via strict label fusion, improves reliability by retaining only exact cross-model agreement, but reduces diversity by discarding borderline or ambiguous cases that may be informative for studying nuanced regret. Corpus 4 relaxes this requirement to broaden coverage, but may reintroduce inconsistencies. Across all corpora, we rely exclusively on text. Multimodal content which can be central to regrettable disclosures [Guo et al., 2025], are not considered. Finally, because self-training is iterative, any early-stage bias or systematic error can propagate and become embedded in the final corpora.

Evaluation models. All six downstream models used to assess corpus utility also have intrinsic constraints. The text-only baselines (Models 5-6) cannot capture contextual factors beyond language, such as user behavior, temporal dynamics, or social interactions that may shape regret. Models incorporating auxiliary signals (Models 1-4) are more robust, but these features still provide only partial coverage of the multifaceted nature of regrettable disclosures. Future work should integrate multi-modal inputs, temporal behavior (e.g., posting frequency, deletion timing), and network context (e.g., interaction patterns) into more expressive architectures. Additional improvements may come from ensemble methods to balance complementary strengths across classifiers and continual learning to adapt to evolving linguistic norms on social media.

Ethical Considerations. All data used in this work were collected from publicly accessible online sources. Nevertheless, public availability does not eliminate ethical or privacy risks. To mitigate these concerns, we collected no user-level metadata beyond post text and removed or pseudonymized all personally identifiable information prior to processing. Datasets are stored on secured, access-controlled servers, and data sharing is strictly restricted. The data collection and annotation procedures were reviewed and approved by both the Data Protection Coordinator and Institutional Review Board at our Institute.

6 CONCLUSION

This paper introduced ReDoS-M, a large-scale, feature-enriched extension of the REGRET dataset [Simo and Kreutzer, 2022a] designed to advance automated detection of regrettable disclosures on online social networks. To construct ReDoS-M, we developed a new multi-stage annotation pipeline that combines crowd-sourced labeling, transformer-based self-training, and semantic enrichment through SME features. Ap-

plying this pipeline to a collection of more than five million user-generated posts and comments drawn from over three dozen online sources resulted in four structurally distinct multi-label corpora. These corpora, ranging from 4.27M to 5.13M annotated instances, substantially exceed previous resources in scale, topical breadth, and contextual richness, and collectively represent the most comprehensive dataset of regret-related OSN disclosures to date. We conducted an extensive empirical evaluation of ReDoS-M, assessing both dataset coverage and distribution and the downstream utility of each corpus for multi-label classification of regrettable disclosures. Six transformer-based models (DeBERTa and XLM-RoBERTa variants, with and without SME- and LLM-generated feature augmented variants) were trained and tested on each of the four ReDoS-M corpora. Across settings, the transformer-based models achieved consistently strong performance, with micro-F1 exceeding 0.98 and AUC values above 0.99 in the best configurations. These results confirm that the regret labels in ReDoS-M are highly learnable and that the corpora provide a reliable foundation for developing robust detection systems for regrettable OSN disclosures. At the same time, discrepancies between micro- and macro-F1 scores highlighted systematic challenges associated with minority regret categories in ReDoS-M, pointing to opportunities for improved data balancing and targeted corpus expansion. These insights highlight the need for further research to address bias in raw data collection and label assignment, and on developing architectures that extend beyond text-only content. Indeed, incorporating multi-modal user-generated content, temporal posting dynamics, and other behavioral signals into both future version of our dataset and detection models may enable more reliable generalization across platforms, communities, and regret types. Overall, ReDoS-M constitutes the most comprehensive and semantically rich resource to date for computational research on regrettable self-disclosures.

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.

Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N., and Acquisti, A. (2013). Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 897–908.

Amini, M.-R., Feofanov, V., Pauletto, L., Devijver, E., and Maximov, Y. (2022). Self-training: A survey. *arXiv preprint arXiv:2202.12040*.

Amini, M.-R. and Gallinari, P. (2002). The use of unlabeled data to improve supervised learning for text summarization. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 105–112.

Arthur, R. (2019). We analyzed more than 1 million comments on 4chan. hate speech there has spiked by 40% since 2015 - violent threats against minorities have also proliferated on the anonymous message board. *vice.com*.

Balouchzahi, F., Butt, S., Sidorov, G., and Gelbukh, A. (2023). Reddit: Regret detection and domain identification from text. *Expert Systems with Applications*, 225:120099.

Breland, A. (2019). Why reddit is losing its battle with online hate - new research shows how the message board keeps giving bigotry a home. *MotherJones*.

Cai, W., Encarnacion, R., Chern, B., Corbett-Davies, S., Bogen, M., Bergman, S., and Goel, S. (2022). Adaptive sampling strategies to construct equitable training datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1467–1478.

Calo, R. (2011). The boundaries of privacy harm. *Ind. LJ*, 86:1131.

Campbell, S., Greenwood, M., Prior, S., Shearer, T., Walkem, K., Young, S., Bywaters, D., and Walker, K. (2020). Purposive sampling: complex or simple? research case examples. *Journal of research in Nursing*, 25(8):652–661.

Chapelle, O., Scholkopf, B., and Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Dhingra, A. and Mittal, S. (2015). Content based spam classification in twitter using multi-layer perceptron learning. *International Journal of*

Latest Trends in Engineering and Technology, 5(4).

Diaz Ferreyra, N. E., Shahi, G. K., Tony, C., Stieglitz, S., and Scandariato, R. (2023). Regret, delete,(do not) repeat: an analysis of self-cleaning practices on twitter after the outbreak of the covid-19 pandemic. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Ekman, P. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Guo, L., Fu, Y., Lin, X., Xu, X., Chang, Y.-J., and Hiniker, A. (2025). What social media use do people regret? an analysis of 34k smartphone screenshots with multimodal llm. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

He, J., Gu, J., Shen, J., and Ranzato, M. (2019). Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788*.

Hicks, D. and Gasca, D. (2019). A healthier twitter: Progress and more todo. *Blog.twitter.com*.

Hossain, N., Hu, T., Feizi, R., White, A. M., Luo, J., and Kautz, H. (2016). Inferring fine-grained details on user activities and home location from social media: Detecting drinking-while-tweeting patterns in communities. *arXiv preprint arXiv:1603.03181*.

Hu, E. (2013). When social sharing goes wrong: Regretting the facebook post.

Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., et al. (2017). Population based training of neural networks. *arXiv preprint arXiv:1711.09846*.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kaggle.com (2012). Detecting insults in social commentary.

Kaur, P., Dhir, A., Chen, S., and Rajala, R. (2016). Understanding online regret experience using the theoretical lens of flow experience. *Comput. Hum. Behav.*, 57(C):230–239.

Khan, M. T., Dimitrov, D., and Dietze, S. (2025). Characterization of tweet deletion patterns in the context of covid-19 discourse and polarization. In *Proceedings of the 36th ACM Conference on Hypertext and Social Media*, HT '25, page 43–47, New York, NY, USA. Association for Computing Machinery.

Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klinger, R., Suliya, S. S., and Reiter, N. (2016). Automatic emotion detection for antitative literary studies.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Liao, W. and Veeramachaneni, S. (2009). A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 58–65.

Lin, J., Nogueira, R., and Yates, A. (2022). *Pretrained transformers for text ranking: Bert and beyond*. Springer Nature.

Mienye, I. D., Swart, T. G., and Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, 15(9):517.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Minaei, M., Mouli, S. C., Mondal, M., Ribeiro, B., and Kate, A. (2020). Deceptive deletions for protecting withdrawn posts on social platforms. *arXiv preprint arXiv:2005.14113*.

Otter, D. W., Medina, J. R., and Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624.

Pavlopoulos, J., Sorensen, J., Laugier, L., and Androutsopoulos, I. (2021). Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69.

Petrovic, S., Osborne, M., and Lavrenko, V. (2013). I wish i didn't say that! analyzing and predicting deleted messages in twitter. *arXiv preprint arXiv:1305.3107*.

Risch, J., Stoll, A., Wilms, L., and Wiegand, M. (2021). Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12. Association for Computational Linguistics.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In Beißwenger, M., Wojatzki, M., and Zesch, T., editors, *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum.

SalishCoast and Karma (2021). deleted-tweets-archive. `https://github.com/salcoast/deleted-tweets-archive`.

Shamsolmoali, P., Zareapoor, M., Shen, L., Sadka, A. H., and Yang, J. (2021). Imbalanced data

learning by minority class augmentation using capsule adversarial networks. *Neurocomputing*, 459:481–493.

Siegel, M. and Bond, F. (2021). Odenet: Compiling a germanwordnet from other resources. In *Proceedings of the 11th Global Wordnet Conference*, pages 192–198.

Simo, H. and Kreutzer, M. (2022a). Regrets: A new corpus of regrettable (self-) disclosures on social media. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–2. IEEE.

Simo, H. and Kreutzer, M. (2022b). Towards automated detection and prevention of regrettable (self-) disclosures on social media. In *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 638–645. IEEE.

Simo, H., Kreutzer, M., and Nikolov, J. (2025). Redos-m: A dataset of multi-label regrettable disclosures on social media (extended version). `https://github.com/AvatarPriv/ReDoS-M/blob/main/ReDoS_M_ExtendedVersion.pdf`.

Simo, H. and Shulman, H. (2021). Poster: Wallguard-a deep learning approach for avoiding regrettable posts in social media. In *2021 IEEE 41st international conference on distributed computing systems (ICDCS)*, pages 1142–1143. IEEE.

Sleeper, M., Balebako, R., Das, S., McConahy, A. L., Wiese, J., and Cranor, L. F. (2013a). The post that wasn't: exploring self-censorship on facebook. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 793–802. ACM.

Sleeper, M., Cranshaw, J., Kelley, P. G., Ur, B., Acquisti, A., Cranor, L. F., and Sadeh, N. (2013b). "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, pages 3277–3286, New York, NY, USA. ACM.

Snow, S. (2015). Don't post that! why half of americans regret their social media posts.

Stern, S. (2015). Regretted online self-presentations: U.s. college students' recollections and reflections. *Journal of Children and Media*, 9:248–265.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558.

Wang, Q., Xue, H., Li, F., Lee, D., and Luo, B. (2019). # donttweetthis: Scoring private information in social networks. *Proceedings on Privacy Enhancing Technologies*, 2019(4):72–92.

Wang, Y., Leon, P. G., Chen, X., and Komanduri, S. (2013). From facebook regrets to facebook privacy nudges. *Ohio St. LJ*, 74:1307.

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, pages 10:1–10:16, New York, NY, USA. ACM.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. International World Wide Web Conferences Steering Committee.

Xie, W. and Kang, C. (2015). See you, see me: Teenagers' self-disclosure and regret of posting on social network site. *Computers in Human Behavior*, 52.

Xu, J.-M., Burchfiel, B., Zhu, X., and Bellmore, A. (2013). An examination of regret in bullying tweets. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 697–702.

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

Zhou, L., Wang, W., and Chen, K. (2015). Identifying regrettable messages from tweets. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 145–146, New York, NY, USA. ACM.

Zhou, L., Wang, W., and Chen, K. (2016). Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones. In *Proceedings of the 25th International Conference on World Wide Web*, pages 603–612.

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