What Clued the AI Doctor In? On the Influence of Data Source and Quality for Transformer-Based Medical Self-Disclosure Detection

Mina Valizadeh, Xing Qian, Pardis Ranjbar-Noiey, Cornelia Caragea, and **Natalie Parde**
Department of Computer Science
University of Illinois Chicago
{mvaliz2, xqian20, pranjb3, cornelia, parde}@uic.edu

Abstract

Recognizing medical self-disclosure is important in many healthcare contexts, but it has been under-explored by the NLP community. We conduct a three-pronged investigation of this task. We (1) manually expand and refine the only existing medical self-disclosure corpus, resulting in a new, publicly available dataset of 3,919 social media posts with clinically validated labels and high compatibility with the existing task-specific protocol. We also (2) study the merits of pretraining task domain and text style by comparing Transformer-based models for this task, pretrained from general, medical, and social media sources. Our *BERTweet* condition outperforms the existing state of the art for this task by a relative F_1 score increase of 16.73%. Finally, we (3) compare data augmentation techniques for this task, to assess the extent to which medical self-disclosure data may be further synthetically expanded. We discover that this task poses many challenges for data augmentation techniques, and we provide an in-depth analysis of identified trends.

1 Introduction

Self-disclosure is a complex communicative process (Kreiner and Levi-Belz, 2019) that involves sharing one's personal thoughts, feelings, or memories with another individual (Jourard and Friedman, 1970). Reciprocal self-disclosure between conversation partners may strengthen relationships (Altman and Taylor, 1973) and improve the communicative experience (Wang et al., 2016). Self-disclosure may take nuanced forms, such as *medical self-disclosure* (see Figure 1), or the act of disclosing symptoms, diagnoses, or other information related to mental or physical health problems (Valizadeh et al., 2021; Joinson, 2001). Some medical self-disclosures may be explicit (A), whereas others may be less direct (B and C).

Medical self-disclosure is helpful from a clinical perspective and reinforces therapeutic relationships during medical interactions (Jannat, 2018; Kadji and Mast, 2021). It also may enable earlier detection and treatment of medical issues (Joinson, 2001; Tidwell and Walther, 2002; Valizadeh et al., 2021). Detecting medical self-disclosure automatically could support clinicians and other medical practitioners in productively identifying helpful patient information from untapped sources or during review of medical conversations (Farber, 2003; Stricker, 2003; Valizadeh and Parde, 2022; Kaelin et al., 2021). However, research towards automated medical self-disclosure detection has been limited and may require different techniques from those used to detect self-disclosure in the general domain (Valizadeh et al., 2021; Reuel et al., 2022).

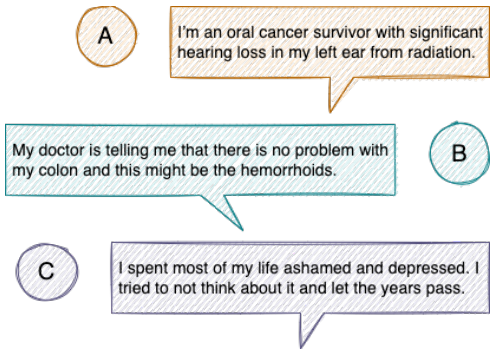

Figure 1: Examples of medical self-disclosure. Note that medical self-disclosure may be explicit (A) or less direct (B and C).

In this paper, we comprehensively investigate automated medical self-disclosure detection. We manually expand the only existing dataset (MEDSD) in this domain to include nearly 4,000 additional instances collected from publicly available forums, strengthening our understanding of this task and its relationship between performance and dataset size. In parallel, we experiment with data augmentation to study the feasibility of automated dataset expansion for this task. Finally, we perform a comparative analysis of transfer learning models, to disentangle the subtle distinctions and shared charac-

teristics between medical self-disclosure detection and other language tasks. Our primary contributions are as follows.

First (1), we create a publicly-available 3,919-instance expansion to MEDSD, sourced from health-related social media posts. The dataset is triple-annotated with continuous (0-5) and graded (NO SELF-DISCLOSURE, POSSIBLE SELF-DISCLOSURE, and CLEAR SELF-DISCLOSURE) labels. Next (2), we conduct transfer learning experiments using Transformer-based models to determine the utility of different source datasets in the context of detecting medical self-disclosure. We find that our highest-performing model outperforms the state of the art (Valizadeh et al., 2021) by relative percentage increases of 14.19% and 16.73% for accuracy and F_1, respectively. Finally (3), we conduct the first study of data augmentation in the context of medical self-disclosure. We find that this task poses unique challenges for data augmentation, and we explore these challenges in a detailed analysis. It is our hope that our findings add to the burgeoning knowledge base surrounding the detection and processing of medical self-disclosure, and can be used as a guide for others working within this domain. We make our data and source code publicly available to facilitate replication and rapid follow-up.

2 Related Work

Clinical literature has extensively examined the effects of patient self-disclosure with medical professionals, finding both positive benefits to the patient's health and overall wellbeing (Arroll and Allen, 2015), as well as negative experiences if the patient feels that professional or personal boundaries have been crossed (Lussier and Richard, 2007). Often self-disclosure in these studies is broadly construed, with less frequent exploration of medical self-disclosure specifically (Wagner, 1982). For instance, Oprescu et al. (2013) examine the correlation between information-seeking behaviors and self-disclosure, and Weisband and Kiesler (1996) investigate disclosure in online and in-person settings through statistical meta-analyses of standardized interviews, questionnaires, tests, and scales reported in existing studies. They conclude that disclosure of sensitive and private information occurs more frequently in online settings, highlighting the urgency of analyzing these settings more closely.

Bak et al. (2014) developed a semi-supervised algorithm to classify self-disclosure levels from Twitter conversations, and Blose et al. (2020) studied self-disclosure in a large dataset of Twitter conversations about the Coronavirus pandemic, proposing an unsupervised approach to detect voluntary disclosure of personal information. Yang et al. (2017) detected positive and negative self-disclosures based on linguistic features, such as LIWC and word embeddings, in a supervised model. Recently, Reuel et al. (2022) created a multi-task RoBERTa model to measure self-disclosure across varying domains, including both general and medical self-disclosure. They found poor generalization across datasets, with in-domain self-disclosure detection models performing much better than across-domain models due to widespread differences in the topics and targets of self-disclosure.

Only two existing works have focused on detecting medical self-disclosure specifically. Balani and De Choudhury (2015) predicted levels of self-disclosure of mental wellness in Reddit posts, and in our prior work we introduced the MEDSD dataset and established a Transformer-based performance benchmark for the task of medical self-disclosure detection (Valizadeh et al., 2021). Since the latter is the current state of the art and offers the only publicly available dataset and corresponding benchmark in this domain, we primarily build upon that work. We investigate medical self-disclosure specifically, and also study the manual and automated expansion of the MEDSD dataset.

3 Data

3.1 Data Collection

MEDSD, our source dataset, is a 6,639-instance dataset annotated with continuous and graded (NO SELF-DISCLOSURE, POSSIBLE SELF-DISCLOSURE, and CLEAR SELF-DISCLOSURE) medical self-disclosure labels. Since our work seeks to empirically compare the MEDSD benchmark (a fine-tuned DistilBERT model) and other Transformer-based models, we sought to expand it to a size more consistent with that used to fine-tune other Transformer-based models (typically 10k or more instances (Devlin et al., 2019)) while recognizing the time-consuming and challenging nature of the annotation task (Farzana et al., 2020). This also presented the opportunity to study the relative merits of manual and automated (via augmentation) dataset expansion for the task of medical

self-disclosure detection.

To ensure consistency with the original MEDSD, we followed the same annotation procedures outlined in our previous work (Valizadeh et al., 2021). We selected *patient.info*[1] as our sole data source, and downloaded publicly-available English-language posts from randomly-selected subforums. Our rationale for selecting *patient.info* as our sole data source (rather than our primary data source as in the original data collection process) was largely due to task-based motivations—we wanted to maximize our capacity to collect data with medical information, and we observed adequate diversity in terms of style and content when performing a preliminary review of *patient.info* posts. To a lesser extent, our decision was also motivated by site-specific data privacy and sharing policies. Since posts collected from other websites in our prior work already comprised a small minority of the dataset (less than 12%), we did not anticipate that this difference would result in meaningful distributional shifts between the original and expanded datasets in terms of data content or style.

We provide additional details and examples regarding our download process in the appendix (Tables 7 and 8). Posts were complete written utterances submitted by users, and we automatically segmented posts at the paragraph level following previously established guidelines. Collected data instances had an average length of 41 tokens. Overall, we added 3,919 instances to MEDSD to reach a raw, merged dataset size of 10,558 instances. To preserve user privacy, we did not download usernames or other identifying metadata and replaced any names written directly within the post text with generic name tokens. The terms and conditions of *patient.info* maintain public access to forum posts, and this research was approved by the Institutional Review Board (IRB) at the University of Illinois Chicago. Similarly to our prior work, we make our dataset expansion available upon request.

3.2 Data Annotation

We observed very high inter-annotator agreement ($\kappa = 0.88$), measured using averaged pairwise per-class Cohen's kappa scores similarly to our prior work to facilitate comparison with annotation quality in the original MEDSD. This suggests that the annotation guidelines were sufficient for consistent replication (Landis and Koch, 1977). We did not make any changes to the guidelines to maximize labeling compatibility between original and new samples. We trained three annotators (computer science graduate and undergraduate students; a mixture of fluent L2 and native English speakers) to follow these guidelines, collecting triple annotations for each post. Labels were first assigned using a graded scheme ranging from 0-5, with "0" indicating no self-disclosure and "5" indicating high self-disclosure. More specifically, we stipulated that numeric labels should be assigned as follows:

- **5:** The post discloses a specific illness, medication, surgery, or other specific medical variable(s) or event(s).
- **4:** The post discloses specific symptoms but does not further specify an illness, medication, or other diagnosis.
- **1-3:** The post ranges from very low (ambiguous hinting of possible, non-specific medical concerns) to moderate (clear reference to non-specific medical concerns) self-disclosure.
- **0:** The post does not disclose any health-related information.

Annotations were averaged across all three labels for each instance. Cases for which the distance between one or more individual annotators and the average was greater than 1.0 were forwarded to a third-party adjudicator. The adjudicator (a study lead with high task familiarity) was authorized to decide the gold standard value based on the provided annotations and their judgment of the instance itself. The continuous-valued (averaged or adjudicated) labels were binned into three discrete classes: [0-1] NO SD, (1-4) POSSIBLE SD, and [4-5] CLEAR SD. This resulted in 520 *No SD* instances (13.26% of the expanded dataset), 1,545 *Possible SD* instances (39.42%), and 1,854 *Clear SD* instances (47.30%).

When collecting our data, we ensured that none of our newly annotated posts were duplicates of those already present in the existing MedSD, to maximize the contribution of new information and content in this dataset expansion. We also spot-checked agreement early in the annotation process to identify and resolve systemic disagreements before they could take root. We worked with in-person (rather than crowdsourced) annotators to allow further oversight for compliance with annotation guidelines as needed. We observed very high

[1] https://patient.info, a popular online forum that offers publicly available information and posts on health, disease, and other medical topics.

Dataset	No SD	Pos. SD	Clear SD	Total
MEDSD	2651	1019	2969	6639
Expansion	520	1545	1854	3919
Merged	3171	2645	4823	10558
Final	**2945**	**2172**	**4650**	**9767**

Table 1: Medical self-disclosure datasets statistics. The final dataset refers to the merged dataset comprising both the original MEDSD and our new dataset expansion after refinement.

final agreement: 93.59% of records did not require adjudication, and the averaged pairwise Cohen's kappa (Landis and Koch, 1977) was 0.81 across the entire dataset, indicating high agreement in line with that originally reported for MEDSD. Per-class agreement was calculated at $\kappa = 0.85$, $\kappa = 0.74$, and $\kappa = 0.85$ for the *No SD*, *Possible SD*, and *Clear SD* classes, respectively. We release both the averaged scores and discretized class labels with our dataset.

3.3 Dataset Refinement

We observed that a sizable portion of mispredictions by the MEDSD baseline on our new expansion were adjudicated instances. After further manual review of the full MEDSD including our expansion, we discarded instances that required adjudication (791 total) from further inclusion in our experiments. Our rationale for this refinement step is that these cases may have arisen from ambiguous context, systemic and conflicted understanding of certain nuances in the annotation guidelines, or occasional annotation mistakes that produced labels that the adjudicator considered to have reasonable justification (Kilgarriff, 1998).

After refinement, the final, merged dataset comprising the pruned MEDSD and our expansion without adjudicated records includes 9,767 instances (4,650 *Clear SD*, 2,172 *Possible SD*, and 2,945 *No SD* instances). Table 1 provides more details regarding the dataset composition and class distribution. Table 9 in the appendix also illustrates the performance change observed when running the benchmark model on the final, refined dataset versus the raw merged version. We release our refined MEDSD alongside our other data as an additional resource for the community.

3.4 Clinical Validation

Since our annotators did not have external clinical expertise, we also recruited a clinical expert (a frequent collaborator who holds graduate degrees in the healthcare domain and has worked in clinical settings) to manually label a subset of our data. We randomly selected a sample of 1,465 data instances (15% of the final dataset size, evenly distributed across classes), and asked the expert to assign labels based on our annotation guidelines and drawing upon their own expertise.

We compared the expert labels with our gold standard, and observed high agreement. In total, 92.37% of the expert's labels (1353/1465 instances) matched those in the gold standard, confirming our annotation reliability from a clinical perspective. At a per-class level, we observed 95.97% label compatibility with *No SD* instances, 83.84% compatibility with *Possible SD* instances, and 97.31% compatibility with *Clear SD* instances. In the appendix (Table 15), we briefly discuss some examples of disagreement between the gold standard and the clinical expert for further analysis.

4 Methods

We compared fine-tuned Transformer models from varying source domains on the task of medical self-disclosure detection (§4.2), and also studied the utility of data augmentation for automated dataset expansion within this domain (§4.3). We describe these studies in the following subsections.

4.1 Data Preprocessing

We passed each instance in our final dataset through a two-step preprocessing pipeline. First, we converted all emojis to their American English CDLR short names (e.g., 🙂 → *:smiling face with smiling eyes:*), since insight into an author's emotional status may provide valuable clues to the presence of self-disclosure (Eisner et al., 2016; Felbo et al., 2017). Next, we replaced all numeric values with generic NUMBER_TOKENS. This prevented our models from drawing spurious conclusions regarding specific values, allowing them to recenter their focus on the presence of numeric content. To validate the utility of these preprocessing steps, we ran the MEDSD baseline (Valizadeh et al., 2021) on our final dataset without any preprocessing steps and with each step individually. We report our findings in Table 11 in the appendix, demonstrating that each step results in small but measurable performance improvements.

4.2 Models

One arm of our study focuses on the influence of pretraining style and task domain on task performance. This is naturally supported through the use of existing, externally-validated Transformer models that were originally leveraged for other research problems. We emphasize therefore that the novelty of our work is not in the model implementation itself, but in the study of its application to this new domain and the extent to which these models could sufficiently leverage and generalize from out-of-domain or near-domain data. We focused our comparison of Transformer-based models on those that are pretrained on health or social media data. In total, we consider seven models: *DistilBERT*, *RoBERTa*, *BioBERT*, *Bio-ClinicalBERT*, *Bio-RedditBERT*, *MentalBERT*, and *BERTweet*.

DistilBERT was the highest-performing model in our initial experiments on MEDSD, and is the current state of the art for this task. The model achieves comparable performance to larger Transformer-based models while requiring much less time and space through the use of knowledge distillation (Sanh et al., 2019), and is often used for lower-resource tasks. In contrast, *RoBERTa* replicates BERT but is trained on larger batches, a higher number of epochs, and more training data, often resulting in higher performance than the original BERT model (Liu et al., 2019). Both *DistilBERT* and *RoBERTa* are trained on general domain data (BookCorpus and Wikipedia for both, and additional news and web data for RoBERTa).

BioBERT leverages nearly the same architecture as BERT, but it is pretrained on large biomedical corpora (Lee et al., 2020). Prior work has shown that it outperforms other BERT-based models at biomedical text classification tasks (Mitra et al., 2021; Zhu et al., 2020). Although *BioBERT* is designed to perform well on biomedical tasks specifically, these may differ from tasks using clinical notes or more casual health discourse. *Bio-ClinicalBERT* is pretrained on two million clinical notes from the MIMIC-III v1.4 database (Johnson et al., 2016; Alsentzer et al., 2019), and *Bio-RedditBERT* is initialized from *BioBERT* (Lee et al., 2020) and then further pretrained on health-related Reddit posts (Basaldella et al., 2020). *MentalBERT* is pretrained on mental health posts collected from Reddit (Ji et al., 2022). Finally, *BERTweet* was pretrained using *RoBERTa*'s training procedures on a massive amount (80 GB) of uncompressed English tweet text, including 845,000,000 Tweets streamed from January 2012 to August 2019 and 5,000,000 Tweets related to the Covid-19 pandemic. It outperformed previous models on a wide variety of social media tasks (Nguyen et al., 2020).

We fine-tuned each included model separately for our task. We applied model gradual layer freezing, and optimized model hyperparameters using grid search. Table 13 provides additional details regarding the fine-tuned hyperparameters.

4.3 Data Augmentation

Manual dataset expansion, as described in Section 3, can be expensive and time-consuming. Data augmentation (DA) strategies can be used to automatically increase training set size, offering an attractive way to reduce overfitting or other issues causing poor predictive performance (Bayer et al., 2021; Feng et al., 2021). They also offer an opportunity to shift the class distribution of unbalanced datasets through generation of additional samples for specific classes, and they have grown more common in NLP recently (Feng et al., 2021).

Although DA techniques have not been explored in the context of self-disclosure detection, this task offers an intriguing testbed for these experiments since it relies on nuanced language with uneven class distribution (see Table 1). We experimented with the following DA techniques to synthetically expand our dataset and balance its distribution:

- **Backtranslation Augmentation (BT):** Data is translated to a different target language (in this case, German) and then back to the source language (English). This often paraphrases the original text (Beddiar et al., 2021).
- **WordNet Synonym Augmentation (WS):** One or more words, depending on a fine-tuned hyperparameter *aug_p* that controls the percentage of words to be augmented, are replaced with their synonyms from WordNet, a large English lexical database (Ramachandran and Parvathi, 2021).[2] We set *aug_p*=0.3 in our experiments.
- **Masked Language Model Augmentation (MLM):** A random sample of words is "masked out" and a pretrained Transformer model is used to restore the text to its original version, typically resulting in a paraphrase

[2] https://wordnet.princeton.edu/

(Wu et al., 2019; Kumar et al., 2020). We experiment with DistilBERT (Sanh et al., 2019) and RoBERTa (Liu et al., 2019) backbones for this technique.

- **Target-Aware Data Augmentation (TA):** The masked word is conditioned on both its context and the prepended auxiliary sentence that contains target and label information, producing alternative versions of the original text that are geared toward a specific target (in our case, the specific target is medical self-disclosure) (Li and Caragea, 2021).

We generated augmented data for the *Possible SD* and *No SD* classes using each of these techniques as a separate condition. We fine-tune the augmentation ratio, a hyperparameter indicating the multiplicative factor by which the dataset size is increased, setting it to 2 for *Possible SD* and 1.5 for *No SD*. Figure 5 in the appendix demonstrates that the DA process resulted in a relatively balanced training set (3,488 *Clear SD*, 3,258 *Possible SD*, and 3,488 *No SD* instances). Following standard protocol we did not augment the validation or test sets, to avoid introducing potential biases and allow for direct comparison with other models.

To compare DA approaches with one another, we train our best-performing model from our modeling experiments (§5.2) on separate combinations of the final, manual dataset and each of the synthetic dataset expansions (*BT*, *WS*, *DistilBERT MLM*, *RoBERTa MLM*, and *TA*). We additionally compare to a baseline *no augmentation* condition trained only on the final, manual dataset.

5 Evaluation

We compare model performance using accuracy, precision, recall, and macro-averaged F_1, following prior work on self-disclosure detection (Valizadeh et al., 2021; Balani and De Choudhury, 2015). For each experiment, we randomly split the specified data into training (80%), validation (10%), and test (10%) subsets. We measure the efficacy of our manual dataset expansion (§5.1), empirically compare the performance of the proposed Transformer models (§5.2), and evaluate the performance of DA techniques for medical self-disclosure (§5.3).

5.1 Does more (manual) data lead to better performance?

To investigate whether the manual expansion of MEDSD directly results in increased model performance, we train *DistilBERT*, the MEDSD benchmark, separately on the original MEDSD and the final dataset (Table 2).[3] The model trained on the final dataset exhibits relative percentage increases of 5.39% and 6.87% for accuracy and F_1, respectively, compared to the model trained on MEDSD. This suggests that manual data expansion boosts performance for this task, irregardless of model architecture or pretraining settings.

Model	Acc.	Precision	Recall	F_1
MEDSD	76.77	0.7497	0.7241	0.7313
Final	**80.91**	**0.7832**	**0.7810**	**0.7816**

Table 2: Comparison between DistilBERT models trained on the original MEDSD and final datasets, separately. Accuracy shown as a percentage (%).

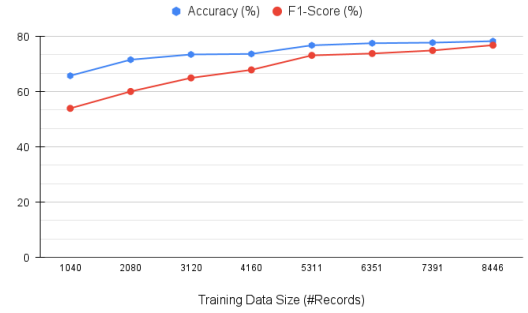

Figure 2: Training dataset size versus *DistilBERT* performance scores. Accuracy and F_1 both shown as a percentage (%) to facilitate presentation.

However, there also appears to be a limit to which more manual data results in meaningful improvements. Figure 2 shows *DistilBERT* performance at gradually increasing training dataset sizes, exhibiting a flattening performance curve approaching the final dataset size. Although this suggests a performance plateau approaching the full size of the manually-annotated dataset, further experiments are needed to justify a broader claim that training set size is detached from performance beyond a certain size threshold. We move closer towards validating this claim with the findings from our data augmentation experiments (§5.3).

[3]The MEDSD results reported in Table 2 (top row) are lower than we observed previously (Valizadeh et al., 2021). After replicating those experiments for this study with different random seeds to assess statistical significance, we found that when averaged across numerous runs, multinomial *DistilBERT* achieves the performance values reported here. These scores remain higher than the other baselines studied by Valizadeh et al. (2021), and the conclusions resulting from those earlier experiments still hold (and are statistically significant).

Model	Acc.	Precision	Recall	F_1
DistilBERT	80.91	0.7832	0.7810	0.7816
RoBERTa	84.40	0.8172	0.7927	0.7968
BioBERT	84.35	0.8117	0.8219	0.8173
Bio-ClinicalBERT	83.10	0.8042	0.7889	0.7954
Bio-RedditBERT	85.15	0.8241	0.8250	0.8245
MentalBERT	85.42	0.8357	0.8282	0.8208
BERTweet	**87.67**	**0.8528**	**0.8551**	**0.8537**

Table 3: Model performance comparison. Accuracy shown as a percentage (%).

Technique	Acc.	Precision	Recall	F_1
MLM-R	86.28	0.8388	0.8271	0.8325
MLM-D	85.88	0.8355	0.8228	0.8286
TA	85.48	0.8298	0.8362	0.8321
WS	86.01	0.8363	0.8196	0.8263
BT	**86.42**	**0.8389**	**0.8470**	**0.8420**
NA	**87.67**	**0.8528**	**0.8551**	**0.8537**

Table 4: Performance comparison for DA techniques. *MLM-R* and *MLM-D* use RoBERTa and DistilBERT backbones, respectively. *NA* refers to the *no augmentation* condition. Accuracy shown as a percentage (%).

5.2 How do pretraining style and task domain impact model performance?

We present the results of our model comparison in Table 3. *BERTweet* achieved the highest performance, with accuracy=87.67% and F_1=0.8537. The model outperformed the existing *DistilBERT* baseline by relative percent increases in accuracy and F_1 of 8.35% and 9.22%, respectively. We applied McNemar-Bowker's test ($\alpha = 0.05$), an extension of McNemar's test designed to accommodate more than two classes (McNemar, 1947), to assess the statistical significance of our results. We observe that all differences are statistically significant ($p <= 0.05$).

5.3 Does more (augmented) data lead to better performance?

To investigate whether the synthetic expansion of the dataset leads to further performance boosts beyond those observed from the manual dataset expansion, we retrained *BERTweet* on each augmented version of the dataset described in §4.3. We report our results in Table 4. Backtranslation resulted in the best model performance among all DA approaches with accuracy=86.42% and F_1=0.8537. However, *BERTweet* with *no augmentation* outperformed all other conditions. The outcomes from this experiment suggest that synthetic expansion beyond the current dataset size is unlikely to lead to substantial model improvement. This also offers supporting evidence for the insights from Figure 2, discussed in §5.1.

6 Discussion

6.1 Lessons Learned from Model Comparison

To further disentangle the differences in observed model performances, we computed per-class accuracy for each model (Table 5). The *DistilBERT* baseline had acceptable performance when predicting *No SD* and *Clear SD* instances with 88.70% and 87.60% accuracy, respectively, but poor performance in detecting *Possible SD* instances, dropping to 56.25% accuracy. *RoBERTa* resulted in higher accuracy for the *No SD* and *Clear SD* classes (relative percent increases of 7.04% and 2.79%, separately) while also offering a slight improvement in detecting *Possible SD* instances.

Our task's reliance on social media data in the medical domain drove our selection of models pretrained primarily on social media data (*BERTweet*), medical data (*BioBERT* and *Bio-ClinicalBERT*), and data at the intersection of both domains (*Bio-RedditBERT* and *MentalBERT*) for our experiments. We anticipated that *BioBERT* and *Bio-ClinicalBERT* would result in improved recognition of *Clear SD* instances, and these expectations were confirmed with per-class performance increased by 3.29% and 5.38%, respectively, relative to *DistilBERT*. *Bio-RedditBERT* and *MentalBERT* achieved higher performance still, with relative performance increases in the *Possible SD* class in particular of 20.01% and 21.15% over the baseline, emphasizing the importance of text style in addition to domain specificity.

Interestingly, our strongest model overall was *BERTweet*, which is pretrained primarily on social media data although a small subset of the data did have a specific health focus. This suggests that ultimately, stylistic patterns may be more important than task-based knowledge when recognizing specific forms of self-disclosure. At the class level, models trained entirely on healthcare datasets (e.g., *Bio-ClinicalBERT*) experience performance boosts when detecting *Clear SD* instances, and models trained on social media datasets (e.g., *BERTweet*) achieve substantial increases in their detection of *Possible SD* instances, which was the most challenging class for human annotators (§3.2).

Model	No	Possible	Clear
DistilBERT	88.70	56.25	87.60
RoBERTa	94.95	57.48	90.05
BioBERT	89.79	63.69	90.49
Bio-ClinicalBERT	88.86	58.27	92.32
Bio-RedditBERT	90.71	67.51	91.27
MentalBERT	91.64	68.15	90.83
BERTweet	89.38	78.66	91.45

Table 5: Accuracy per class for all the models implemented for detecting medical self-disclosure. Accuracy is shown as a percentage (%).

Dataset	No	Possible	Clear
BT	93.72	70.94	92.27
NA	89.38	78.66	91.45

Table 6: Comparison of per-class accuracy for the *BERTweet* model trained on the backtranslation-augmented (BT) and non-augmented (NA) datasets. Accuracy is shown as a percentage (%).

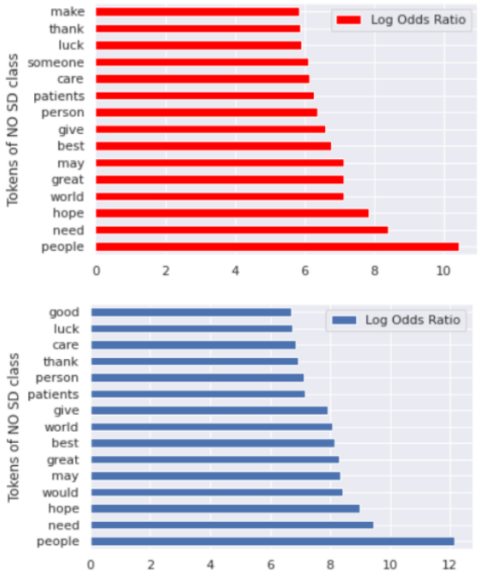

Figure 3: Words most closely associated with No SD class. Red and blue plots correspond to original and augmented data, respectively.

6.2 Factors Influencing Data Augmentation

Our experiments did not provide conclusive evidence that data augmentation can be effectively leveraged to detect medical self-disclosure. This is understandable since the task is known to rely on often-subtle language patterns (Valizadeh et al., 2021; Reuel et al., 2022). We examined the per-class performance of our *BERTweet* model fine-tuned on our backtranslation-augmented dataset compared to the *no augmentation* condition (Table 6) to develop a deeper understanding of opportunities for future improvement.[4]

We find that data augmentation resulted in small performance increases when predicting *No SD* and *Clear SD*, but larger performance reductions when predicting *Possible SD* instances. Although initially counterintuitive since our primary goal in augmenting data was to balance the class distribution (for which frequency was lowest in the *Possible SD* class), we conducted further analyses to pinpoint underlying factors. To analyze the linguistic patterns associated with augmented and non-augmented instances, we computed the log odds ratio with an informative Dirichlet prior (Monroe et al., 2008; Hessel, 2016) for both versions of *No SD* and *Possible SD* (Figures 3 and 4).

We found that after augmenting the dataset, the ratio of third-person nouns (e.g., "people" or "person") increased in the *No SD* class (Figure 3). This may have strengthened the model's inverse association between the use of external language and the confirmed disclosure of medical information, resulting in the observed 4.85% increase in accuracy for *No SD*. The ratio of "feeling" keywords in the *Possible SD* class also increased. Prior to DA, these words were more prominent in the *Clear SD* class than the *Possible SD* class (612 and 331 tokens prior to DA, respectively, and 621 and 622 tokens after), due to their use when expressing physical or mental symptoms. We found that 18.1% of mispredicted *Possible SD* instances contained "feeling" keywords prior to DA, and 29.54% of mispredicted *Possible SD* instances contained these keywords after DA. Thus, it appears that backtranslation created a harmful distributional shift in the expression of feeling language across classes. Finally, we suspect that despite its outperformance over other techniques, backtranslation is still limited in its ability to create convincing synthetic data for this task and may not have captured subtleties in writing style that (as observed in our model comparison) are important (Longpre et al., 2020; Beddiar et al., 2021). As a result, the introduction of synthetically augmented data to the learning process may have merely added noise.

Although our data augmentation experiments did not produce positive results, we note that this is the first exploration of augmentation in the context of

[4]Per-class performance of *BERTweet* fine-tuned on other augmented datasets is provided in Table 14 in the appendix.

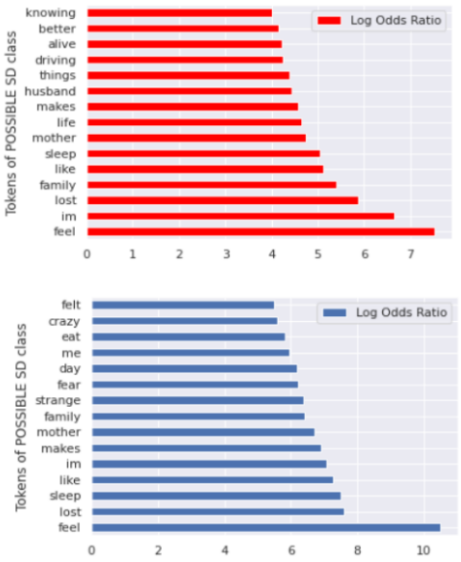

Figure 4: Words most closely associated with POSSIBLE SD class. Red and blue plots correspond to original and augmented data, respectively.

medical self-disclosure. Negative findings move the dial toward more fully understanding the performance boundaries of this and related tasks, and the recent movement towards leaderboards rather than hypothesis-driven investigations in NLP has contributed to the under-reporting of negative results (Tafreshi et al., 2022). This can slow progress as scientists repeat experiments that they are unaware have already been attempted, often with substantial effort and carbon footprint. It is our hope that through reporting the negative outcomes from our DA experiments, we add to the growing knowledge base surrounding data augmentation for NLP and also lay the groundwork for future experiments in self-disclosure detection.

7 Conclusion

In this work, we study data source, quantity, and quality as it pertains to detecting medical self-disclosure. First, we manually expanded MEDSD with 3,919 additional instances and clinically validated its labels in collaboration with a healthcare expert. Next, we compared Transformer models pretrained on varying source datasets for predicting medical self-disclosure, finding that our best-performing model outperforms the state of the art by relative percentage increases of 14.19% and 16.73% for accuracy and F_1, respectively. Our findings also suggest that stylistic patterns prove more revealing than task-specific trends. Finally, we study data augmentation in the context of this task, finding that it poses many DA challenges. We document these in our analysis, opening the door to intriguing follow-up studies. We make our dataset and models available to the research community upon request, and we hope that our work can be used as a roadmap for future experiments in self-disclosure detection.

Acknowledgements

We thank the anonymous reviewers for their helpful suggestions. We thank Dr. Vera Kaelin for providing clinical expertise for validating our labels, and we also thank Shayan Rasheed and Nidhi Bhupalam for their roles in creating the dataset. The first and last authors of this paper were fully or partially supported throughout the duration of this work by the National Science Foundation under Grant No. 2125411. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Limitations

Our work is limited by several factors. Both the original MEDSD dataset and our own expanded version are imbalanced, resulting in some known performance weaknesses. We attempted to correct for this using data augmentation techniques, but across the wide range of techniques tested, none were able to improve performance beyond that of the *no-augmentation* baseline. Thus, we cannot provide conclusive evidence that synthetic dataset expansion is a valuable pursuit for this task (although we note that this negative finding in itself provides worthwhile direction for other researchers). The manual expansion of the medical self-disclosure dataset clearly improved performance, highlighting the effectiveness of human gold standard labels for this task; therefore, manual or semi-supervised dataset expansion may be a promising avenue for future model improvements.

Our highest performing model was *BERTweet*, which is pretrained on a massive amount of primarily general-domain social media data, although a small subset of it was focused on the Covid-19 pandemic. This suggests that stylistic patterns may be more important than domain knowledge for the recognition of specific forms of self-disclosure. Therefore, instead of limiting the methodology to transfer learning models in clinical and medical domains, future work should extend the range of

domains and text styles that are studied. It may be the case that the most worthwhile domains or stylistic cues for supporting this task are further from the target task, and as yet undiscovered.

All data and experiments reported in this paper were conducted on English data, which constrains the extent to which our claims can be generalized. Future work examining medical self-disclosure in languages that are less-resourced than English or that differ greatly in their morphological typology may provide crucial insight into the generalizability of our findings. Finally, training and fine-tuning large Transformer-based models often requires costly GPU resources. This limits the accessibility of running these experiments at scale.

Ethics Statement

This research was approved by the Institutional Review Board (IRB) at the University of Illinois Chicago. Our primary data source was *patient.info*, for which the terms and conditions allow public access to forum posts.[5] As outlined in §3.1 and to protect privacy, we manually anonymized our data instances by removing any usernames or other identifying metadata, and replaced any names written directly within the post text with generic name tokens. Annotators were compensated for their work through paid internships and assistantship positions at a competitive rate for the cost of living in our area. We make our expanded dataset available upon request via email, following IRB protocol.

We intend for our dataset expansion and the proposed methods to be used as a tool to analyze the linguistic trends and other language behaviors associated with medical self-disclosure in online settings. Our experiments closely follow this intent, providing novel insight into the influence of different pre-training tasks and stylistic domains on the ability of our models to recognize possible and clear cases of medical self-disclosure. When the technology is being used as intended and functioning correctly, we anticipate that it may be of value to numerous downstream applications, primarily as a data analysis tool or as an avenue for providing information. When the technology is being used as intended but giving incorrect results, its value may decrease (since researchers or clinicians using the tool may need to discard their insights or may fail to replicate findings in subsequent experiments).

[5] https://patient.info/terms-and-conditions

A potential misuse of the technology would be to use the trained models to identify social media users disclosing medical information and apply targeted advertising or messaging to those users. Since neither our dataset nor our models are designed to recognize or predict specific medical conditions, we note that the extent to which they could be used for these purposes is limited. We do not condone this use of the technology, and we will monitor citations and uses of our dataset to ensure that others are using our data and models for their intended purpose.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. *CoRR*, abs/1904.03323.

Irwin Altman and Dalmas A Taylor. 1973. *Social penetration: The development of interpersonal relationships.* Holt, Rinehart & Winston.

Bruce Arroll and Emily-Charlotte Frances Allen. 2015. To self-disclose or not self-disclose? a systematic review of clinical self-disclosure in primary care. *British Journal of General Practice*, 65(638):e609–e616.

JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014. Self-disclosure topic model for classifying and analyzing twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1986–1996.

Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1373–1378.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. Cometa: A corpus for medical entity linking in the social media. *arXiv preprint arXiv:2010.03295*.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2021. A survey on data augmentation for text classification. *arXiv preprint arXiv:2107.03158*.

Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. 2021. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24:100153.

Taylor Blose, Prasanna Umar, Anna Cinzia Squicciarini, and Sarah Michele Rajtmajer. 2020. Privacy in crisis: A study of self-disclosure during the coronavirus pandemic. *CoRR*, abs/2004.09717.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.

Barry A Farber. 2003. Patient self-disclosure: A review of the research. *Journal of clinical psychology*, 59(5):589–600.

Shahla Farzana, Mina Valizadeh, and Natalie Parde. 2020. Modeling dialogue in conversational cognitive health screening interviews. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1167–1177, Marseille, France. European Language Resources Association.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Jake Hessel. 2016. Implementation: Fightin' words. `https://github.com/jmhessel/FightingWords`.

Khadiza Tul Jannat. 2018. *Self-Disclosure, Gender, and Patient Satisfaction in the Doctor-Patient Relationship*. Minnesota State University, Mankato.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Adam N Joinson. 2001. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European journal of social psychology*, 31(2):177–192.

Sidney M Jourard and Robert Friedman. 1970. Experimenter-subject" distance" and self-disclosure. *Journal of Personality and Social Psychology*, 15(3):278.

Keou Kadji and Marianne Schmid Mast. 2021. The effect of physician self-disclosure on patient self-disclosure and patient perceptions of the physician. *Patient Education and Counseling*, 104(9):2224–2231.

Vera C Kaelin, Mina Valizadeh, Zurisadai Salgado, Natalie Parde, and Mary A Khetani. 2021. Artificial intelligence in rehabilitation targeting the participation of children and youth with disabilities: Scoping review. *J Med Internet Res*, 23(11):e25745.

Adam Kilgarriff. 1998. Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech & Language*, 12(4):453–472.

Hamutal Kreiner and Yossi Levi-Belz. 2019. Self-disclosure here and now: combining retrospective perceived assessment with dynamic behavioral measures. *Frontiers in psychology*, 10:558.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yingjie Li and Cornelia Caragea. 2021. Target-aware data augmentation for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1850–1860.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shayne Longpre, Yu Wang, and Christopher DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? *arXiv preprint arXiv:2010.01764*.

Marie-Thérèse Lussier and Claude Richard. 2007. Self-disclosure during medical encounters. *Canadian family physician*, 53(3):421–422.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Avijit Mitra, Bhanu Pratap Singh Rawat, David D McManus, Hong Yu, et al. 2021. Relation classification for bleeding events from electronic health records using deep learning systems: an empirical study. *JMIR medical informatics*, 9(7):e27527.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *CoRR*, abs/2005.10200.

Florin Oprescu, Shelly Campo, John Lowe, Julie Andsager, and Jose A Morcuende. 2013. Online information exchanges for parents of children with a rare health condition: key findings from an online support community. *Journal of medical Internet research*, 15(1):e2423.

Dharini Ramachandran and R Parvathi. 2021. A novel domain and event adaptive tweet augmentation approach for enhancing the classification of crisis related tweets. *Data & Knowledge Engineering*, 135:101913.

Ann-Katrin Reuel, Sebastian Peralta, João Sedoc, Garrick Sherman, and Lyle Ungar. 2022. Measuring the language of self-disclosure across corpora. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1035–1047, Dublin, Ireland. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

George Stricker. 2003. The many faces of self-disclosure. *Journal of Clinical Psychology*, 59(5):623–630.

Shabnam Tafreshi, João Sedoc, Anna Rogers, Aleksandr Drozd, Anna Rumshisky, and Arjun Akula, editors. 2022. *Proceedings of the Third Workshop on Insights from Negative Results in NLP*. Association for Computational Linguistics, Dublin, Ireland.

Lisa Collins Tidwell and Joseph B Walther. 2002. Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations: Getting to know one another a bit at a time. *Human communication research*, 28(3):317–348.

Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660, Dublin, Ireland. Association for Computational Linguistics.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4398–4408, Online. Association for Computational Linguistics.

Clifford H Wagner. 1982. Simpson's paradox in real life. *The American Statistician*, 36(1):46–48.

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*, CSCW '16, page 74–85, New York, NY, USA. Association for Computing Machinery.

Suzanne Weisband and Sara Kiesler. 1996. Self disclosure on computer forms: Meta-analysis and implications. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3–10.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer.

Diyi Yang, Zheng Yao, and Robert Kraut. 2017. Self-disclosure and channel difference in online health support groups. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 704–707.

Yu Zhu, Lishuang Li, Hongbin Lu, Anqiao Zhou, and Xueyang Qin. 2020. Extracting drug-drug interactions from texts with biobert and multiple entity-aware attentions. *Journal of biomedical informatics*, 106:103451.

A Data Collection

Our data collection and annotation process spanned four months. Similarly to our earlier work (Valizadeh et al., 2021), we covered numerous medical topics, ranging from mental to physical health issues. In Tables 7 and 8, we provide example instances from our dataset regarding some of these topics. We note that although we categorize these instances to facilitate presentation here, labels indicating specific health conditions are not included in our dataset.

B Data Refinement

After our manual data refinement process during which we removed the instances that required adjudication, we ran the benchmark model (*DistilBERT*) established in prior work (Valizadeh et al., 2021) on the final dataset versus the merged version. Table 9 presents the results of this experiment,

PHYSICAL HEALTH CONDITION
Brain & Nerves
1. *I get horrific migraines. The dizziness/feeling faint and spaced out feeling is the worst for me.*
2. *I have nerve testing next week and am going to have my thyroid retested. I also want an MRI and a brain scan. I feel hopeless.*
Chest & Lungs
1. *For the past year, I've had issues with taking in deep breaths, have had an X-Ray, CT Scan on my lungs and they couldn't find anything.*
2. *There is a mass in my right lung and I need to go for a bronchoscopy this coming Monday to determine what exactly it is.*
Bones, Joints, & Muscles
1. *Finally after a month of waiting for results, I was told I do not have Rheumatoid Arthritis. I do wish that they would take my Osteoarthritis more seriously though. I'm glad you had your test and all seemed ok.*
2. *I am nearly 2 years total left knee replacement and was a massive diet coke drinker before surgery. However, since surgery on March 16, I haven't been able to drink the stuff, makes me sick.*
Gut, Bowel, & Stomach
1. *I recently had appendicitis and got my appendix removed about a month ago.*
2. *I went to several doctors for three and a half months with these symptoms to no avail after various negative tests. I was eventually diagnosed with IBS.*

Table 7: Some of the physical health-related topics covered in our dataset.

MENTAL HEALTH CONDITION
Anxiety Disorders
1. *I have Panic Disorder and Generalized Anxiety Disorder. It is well maintained and I have not had a panic attack in years.*
2. *I start PT on Monday and I'm hoping I get a therapist that is understanding. I don't need to cry more than I already have or have anymore anxiety attacks.*
Post Traumatic Stress Disorder (PTSD)
1. *I also suffer from C-PTSD so everything is accentuated and I can't wait till this transition is over for all of us.*
2. *I did not sleep for 6 months due to PTSD maybe micro sleeps. I just laid in bed wide awake or with eyes closed.*
Depression
1. *Since my depression is genetic, it's been following me around for a while. It's always hung around me, but in the past couple of months it hit me again and I was struggling for the longest time.*
2. *I suffered from depression for many years but it is nowhere near as bad as it was because I am on ad's and had counselling. Now I can live with it and have a good life which means something to me.*
Bipolar Disorder
1. *I've only been diagnosed Bipolar just over a month ago. I am very new to it, but looking back over many years I realise I've had it for a while.*
2. *One of the meds I am taking for Bipolar is Seroquel at night. And after eliminating everything else, it seems that the Seroquel might be the cause of it.*

Table 8: Some of the mental health-related topics covered in our dataset.

showing relative percentage increases of 3.37%, 3.14%, 1.79%, and 2.41% for accuracy, precision, recall, and F_1, respectively. We also ran our highest performing model (*BERTweet*) established in Section §5.2 on the final dataset versus the merged version. Table 10 presents the results of this experiment, showing relative percentage increases of 3.16%, 1.87%, 3.47%, and 2.22% for accuracy, precision, recall, and F_1, respectively.

Dataset	Acc.	Precision	Recall	F_1
Merged	78.27	0.7593	0.7672	0.7632
Final	80.91	0.7832	0.7810	0.7816

Table 9: Model performance before and after data refinement. The model used for this experiment is the baseline, a *DistilBERT* model.

Dataset	Acc.	Precision	Recall	F_1
Merged	84.98	0.8371	0.8264	0.8351
Final	87.67	0.8528	0.8551	0.8537

Table 10: Model performance before and after data refinement. The model used for this experiment is our highest performing model, a *BERTweet* model (§5.2).

C Data Preprocessing

To validate the utility of the preprocessing techniques described in §4.1, we ran our *DistilBERT* benchmark and our highest performing model *BERTweet* on the final dataset without any preprocessing steps and with each step individually. We report our findings in Tables 11 and 12, demonstrating that each step results in small but measurable performance improvements.

Technique	Accuracy
Base Model (No Preprocessing)	77.21%
Base + DeEmojifying	78.07%
Base + Number Replacement	78.02%

Table 11: Model performance in accuracy (%) before and after applying each preprocessing technique. *Base model* refers to our baseline *DistilBERT* model, trained on the final dataset (§5.1).

Technique	Accuracy
BERTweet (No Preprocessing)	86.54%
BERTweet + DeEmojifying	87.43%
BERTweet + Number Replacement	87.32%

Table 12: Model performance in accuracy (%) before and after applying each preprocessing technique. *BERTweet* refers to our highest performing model, trained on the final dataset (§5.2).

D Experimental Settings

We optimized model hyperparameters using grid search. Table 13 shows the final hyperparameters we used when training our models.

Model	Learning Rate	Batch Size	Epochs
DistilBERT	2e-6	16	3
RoBERTa	2e-5	16	3
BioBERT	2e-5	8	4
Bio-ClinicalBERT	2e-5	8	4
Bio-RedditBERT	2e-5	16	3
MentalBERT	2e-5	32	3
BERTweet	2e-5	8	3

Table 13: Models' final hyperparameters.

E Data Augmentation

We generated augmented data for the *Possible SD* and *No SD* classes using each of the techniques described in §4.3 as a separate condition. We fine-tune the augmentation ratio, a hyperparameter indicating the multiplicative factor by which the dataset size is increased, setting it to 2 for *Possible SD* and 1.5 for *No SD*. Figure 5 demonstrates that the DA process resulted in a relatively balanced training set (3,488 *Clear SD*, 3,258 *Possible SD*, and 3,488 *No SD* instances).

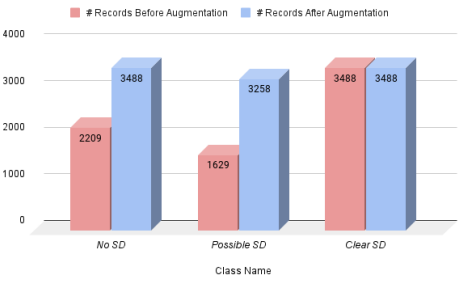

Figure 5: Distribution of training instances before and after data augmentation technique.

F Per-Class Performance for Data Augmentation Techniques

Our experiments did not provide conclusive evidence that data augmentation can be effectively leveraged for the task of detecting medical self-disclosure. This is somewhat unsurprising since the task is known to be challenging due to its reliance on often-subtle language patterns (Valizadeh et al., 2021; Reuel et al., 2022). We further examined the per-class performance of our best-performing model (*BERTweet* fine-tuned on various augmented datasets) compared to the *no augmentation* condition to develop a deeper understanding of opportunities for future improvement, and present the results in Table 14.

Dataset	NO SD	POSSIBLE SD	CLEAR SD
MLM-R	90.95	64.64	92.53
MLM-D	89.79	63.64	92.40
TA	89.95	68.47	89.92
WS	93.03	60.19	90.67
BT	93.72	70.94	92.27
NA	89.38	78.66	91.45

Table 14: Comparison of per-class accuracy for the BERTweet models trained on the augmented and non-augmented (NA) datasets. Accuracy is shown as a percentage (%).

G Clinical Validation

After our clinical validation process, we observed high agreement between the clinical expert and our gold standard: in total, 92.37% of the expert's labels (1353 of the 1465 instances) matched those in the gold standard. However, there were some instances for which our annotators and the clinical expert did not agree. Table 15 presents some of these instances, along with commentary provided by our clinical expert to offer additional insight into their thought process and rationale.

Instance	Clinical Label	Gold Standard	Expert's Note
Although the madness part of King George I might be close to. Stephen Fry is an idol of mine, and someone I watch closely for obvious link in the field of mental health but if he's Premier League I'd barely count as a part time player!	Clear SD	Possible SD	Although a specific diagnosis is not referenced, when viewed from a clinical perspective this is a clear disclosure of a mental health problem.
I don't want to die but I can't afford to live with pain of emotions that I am worth less and no one needs me!	Clear SD	Possible SD	Similar to above, clinical expertise affords additional sensitivity to mental health disclosures.
I have had a perfectly awful past year....the worst ever in my life. I have been in perimenopause for about 4-6 years now (hard to know for sure), I have been on here begging for advice on what I can do to ease my suffering and I was struck with a major realization today!	Possible SD	Clear SD	Menopause itself is not a medical problem since it is a natural circumstance; thus, focus should instead be placed on the suffering that is referenced but unspecific.
I'm also going through hormonal changes so its been rough trying to figure out whats causing what.	Possible SD	Clear SD	Again, hormonal changes are a natural circumstance rather than a healthcare problem; focus should be placed on the unspecified issues.
I'm honestly aware that it's extremely unlikely for a 16 year old to end up with colon cancer, but it does happen in rare occasions so it's hard for me to completely let go of the idea.	No SD	Possible SD	It is unclear whether this disclosure is in reference to the poster.
We have to tell our drs that we are not stupid nor irresponsible with test results. It is your body/health and you need to know and understand what is happening and not kept in the dark.	No SD	Possible SD	Advocating for rights or expertise is not a disclosure of a medical concern in itself.

Table 15: Instances with disagreements in the clinical validation process.