Identifying Medical Self-Disclosure in Online Communities

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

Abstract

Self-disclosure in online health conversations may offer a host of benefits, including earlier detection and treatment of medical issues that may have otherwise gone unaddressed. However, research analyzing medical self-disclosure in online communities is limited. We address this shortcoming by introducing a new dataset of health-related posts collected from online social platforms, categorized into three groups (NO SELF-DISCLOSURE, POSSIBLE SELF-DISCLOSURE, and CLEAR SELF-DISCLOSURE) with high inter-annotator agreement ($\kappa = 0.88$). We make this data available to the research community. We also release a predictive model trained on this dataset that achieves an accuracy of 81.02%, establishing a strong performance benchmark for this task.

1 Introduction

Self-disclosure is a communicative act that helps people develop close relationships (Altman and Taylor, 1973) through reciprocal sharing of personal information, promoting maintenance of trust and security (Bruss and Hill, 2010). It is defined as the "process of making the self known to others" (Joinson and Paine, 2007), often by sharing one's personal thoughts, opinions, or experiences. For example:

- *When I was 19 years old, I met a man on the internet. He was 21 years old, 2 years older than me.*
- *My name is Amy and I live in Australia.*
- *I have suffered from migraines for three years.*

In addition to facilitating social bonds, self-disclosure in general produces a wide variety of health benefits and plays a critical role in successful treatment of many physical and psychological health issues (Ellis and Cromby, 2012). The revelation of private and sensitive information is more widespread online than in face-to-face interactions (Joinson, 2001; Tidwell and Walther, 2002; Wang et al., 2016), perhaps due to the anonymity that online platforms provide, or the ability to avoid the face-to-face stigma of some uncomfortable topics. The benefits of *medical self-disclosure* (i.e., disclosing symptoms, diagnoses, or other information specifically related to mental or physical health issues) in online settings may be particularly valuable from a clinical perspective, enabling earlier detection and treatment of medical issues that may have otherwise gone unaddressed (Pennebaker and Chung, 2007; Joinson, 2001). However, medical self-disclosure has been under-explored in prior computational work. We set out to address that limitation, making several key contributions.

First, we establish the novel task of medical self-disclosure detection, and create a 6,639-instance dataset comprised of public online social posts covering a wide range of mental and physical health issues, annotated with graded (NO SELF-DISCLOSURE, POSSIBLE SELF-DISCLOSURE, and CLEAR SELF-DISCLOSURE) labels. We release this dataset to the research community to facilitate easy replication of our work, as well as rapid entry to this new task by others. Next, we compare a suite of classical machine learning and neural network approaches (including LSTM-, CNN-, and Transformer-based models) for this task, finding that neural approaches typically outperform classical machine learning models. Our highest-performing model, a BERT-based model fine-tuned for the medical self-disclosure task, achieves an accuracy of 81.02%, establishing a strong performance benchmark for this novel task.

Finally, we find that our highest-performing model outperforms the best existing (general) categorical self-disclosure model (Balani and De Choudhury, 2015), retrained on our new med-

*Authors contributed equally.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4398–4408
June 6–11, 2021. ©2021 Association for Computational Linguistics

ical self-disclosure dataset and fine-tuned for this task, by relative percentage increases of 41.81%, 32.63%, 66.60%, and 49.76% for accuracy, precision, recall, and F1-measure, respectively. This provides empirical support that detecting medical self-disclosure is a distinct task with unique linguistic nuances, making it impractical to simply apply existing non-medical self-disclosure models to the medical domain with expectations of similarly high performance. In the long term, it is our hope that high-performing medical self-disclosure models can be deployed in clinical settings to support overburdened healthcare workers in understanding, diagnosing, and treating patients' health issues.

2 Related Work

Self-disclosure detection has been the focus of prior work in psychology (Meleshko and Alden, 1993; Bridges, 2001; Meissner, 2002) and computer science (Bak et al., 2012; Walton and Rice, 2013; Balani and De Choudhury, 2015). However, research examining self-disclosure in online health discourse specifically has been limited. Existing work in this domain shows that detecting self-disclosure in the areas of health and wellness can be beneficial (Pennebaker and Chung, 2007), with patients often preferring to engage in interviews with computers rather than humans and also providing more candid and honest answers to computers (Joinson, 2001). Thus, detecting illness may be an easier process when taking into account patients' virtual disclosures (Ferriter, 1993; Greist et al., 1973). In fact, Coppersmith et al. (2015) relied on self-reported diagnosis when examining linguistic trends in a wide range of mental health conditions on Twitter.[1]

Most computational work on self-disclosure detection has taken place in the general domain, and specifically on tweets. Bak et al. (2012) presented a computational framework for automatically detecting self-disclosure using text mining techniques applied to Twitter conversations, and Walton and Rice (2013) investigated the roles of gender and social identities and their influences on self-disclosure on Twitter by adult users. Outside of Twitter, Umar et al. (2019) also focused on detecting self-disclosure in news commentaries using dependency parsing and named entity recognition. While these studies involve social posts, they do not specifically focus on health.

Balani and De Choudhury (2015) presented a simple neural network with three classes (NO SD, LOW SD, and HIGH SD) to predict self-disclosure of mental wellness in Reddit[2] posts. Their highest-performing approach, a perceptron-based model, achieved an accuracy of 78.4%. Balani and De Choudhury's work is the closest existing work to ours; however, although mental wellness may be a significant interest when identifying self-disclosure in health domains, limiting work to this precludes other critical health concerns such as psychosomatic (Karasu, 1979; Kellner, 1975) or physical ailments.

We address the limitations of prior work in automated self-disclosure detection by including an extensive range of mental and physical health concerns in our dataset. Like Balani and De Choudhury (2015), we consider three self-disclosure categories (in contrast to, e.g., the two classes employed by Umar et al. (2019)). This facilitates a more precise prediction, and focusing on medical self-disclosure specifically helps to (a) validate the distinction between medical and other types of self-disclosure when building automated models for the task, and (b) develop techniques attuned to the latter.

3 Data

3.1 Data Collection

There are currently no publicly-available medical self-disclosure datasets; thus, a key contribution in this work lies in the creation of such a resource. We downloaded publicly-available English-language posts from randomly-selected forums on *patient.info*,[3] as well as a random selection of public posts from other popular online platforms (Reddit, Twitter, and Facebook[4]) to avoid overfitting models to site-specific stylistic trends rather than characteristics more closely linked to the presence of medical self-disclosure.[5] We selected *patient.info* as our primary data source since it is a popular online forum that is well-respected among users from different backgrounds (Lewy, 2013), and it offers publicly available posts on a myriad of

[1] https://twitter.com

[2] https://www.reddit.com

[3] https://patient.info, an online resource that provides information on health, disease, and other medical topics.

[4] https://www.facebook.com

[5] As the focus in this work is on detecting self-disclosure in health-related posts, most instances in our dataset (88.1%) are from *patient.info*. The rest of the instances are approximately distributed as follows: 7.1% from Reddit, 3.3% from Twitter, and 1.5% from Facebook.

	No SD	Possible SD	Clear SD
Kappa (κ)	0.9780	0.6829	0.9840

Table 1: Averaged per-class kappa scores.

general and specific mental and physical health concerns. We randomly sampled these posts to avoid learning too strong of a reliance on disease-specific characteristics (e.g., disclosures about COVID-19 specifically). For posts not from *patient.info*, we scraped data using keywords and hashtags corresponding to frequent unigrams in the *patient.info* posts that were indicative of medical concerns (e.g., "depression," "sick," and "nausea") and purposely included expressions pertaining to both medical and non-medical senses of those words.[6] This discouraged subsequent models from blindly associating certain keywords with medical self-disclosure. For the Reddit data, no specific subreddits were targeted.

We define instances, or *posts*, as complete written utterances submitted by users of the respective data sources. In longer source samples, such as those spanning multiple paragraphs on Reddit, Facebook, or *patient.info*, paragraphs were considered complete utterances. Long samples were thus segmented at the paragraph level, resulting in posts that were approximately equivalent in length to tweets (segmented posts had an average length of 41 tokens, or 214 characters) and thereby avoiding introducing biases associated with post length into the dataset. This resulted in 6,639 instances, each of which were annotated individually. As stipulated by our IRB protocol, we make the dataset available upon request from the authors.

3.2 Data Annotation

Three trained annotators (computer science graduate and undergraduate students; a mixture of fluent L2 and native English speakers) were provided with guidelines describing different levels of medical self-disclosure, or the absence thereof, ranging from 0-5. They were told to label posts without considering prior or future context. Annotators were compensated for their work as part of assistantships or course credit, and were briefed on annotation procedures and best practices prior to starting the annotation process.

[6]For example, "depression" is in isolation most often a medical term, whereas "the great depression" is not.

The guidelines instructed annotators to label posts as containing high self-disclosure (label=5) if they contained clear indications that the poster: (a) had been diagnosed with a specific illness by a medical professional; (b) was taking a specific medication; (c) had undergone a surgery, or was undoubtedly about to have one; (d) had visited a doctor, or was undoubtedly about to see one; or other cases disclosing clear, specific medical variables or events. The guidelines directed annotators to assign labels of "4" when the poster indicated specific symptoms they had but did not further specify an illness, medication, or other diagnosis; and labels ranging from 1-3 to instances with very low (ambiguous hinting of possible, non-specific medical concerns) to moderate (clear reference to non-specific medical concerns) self-disclosure. Finally, the guidelines instructed annotators to assign labels of "0" to instances clearly containing no medical self-disclosure at all.

Each instance was labeled by all three annotators. Annotations were then averaged across all annotators for each instance, and the individual distance between each annotator's label and the average for a given instance was computed. For instances for which the distance between one or more individual annotators and the average was greater than 1.0, the instance was forwarded to a third-party, native English-speaking adjudicator, who determined the gold standard value based on the three annotations and the instance itself. For all other instances, the average label was accepted as the gold standard. These averaged scores were then discretized into the three classes: NO SELF-DISCLOSURE, POSSIBLE SELF-DISCLOSURE, and CLEAR SELF-DISCLOSURE.

We measured inter-annotator agreement using averaged pairwise Cohen's kappa, as well as by calculating the percentage of instances that did *not* require adjudication (91.29%). Averaged pairwise Cohen's kappa across the entire dataset was $\kappa = 0.88$, suggesting high agreement (Landis and Koch, 1977). Table 1 shows the averaged pairwise kappa score among annotators for each class. Agreement for the NO SELF-DISCLOSURE and CLEAR SELF-DISCLOSURE classes was extremely high, whereas agreement for POSSIBLE SELF-DISCLOSURE was lower, although still fair (Landis and Koch, 1977). In Table 2 we provide the raw count and percentage distribution across binned gold standard score ranges of $\{[0-1], (1-4), [4-5]\}$.

Score Ranges	Raw Count	% Distribution
[0-1]	2651	39.93%
(1-4)	1019	15.34%
[4-5]	2969	44.72%

Table 2: Raw count and percentage distribution across binned score ranges.

3.3 Categorical Class Labels

Self-disclosure naturally occurs along a spectrum rather than only at two extremes (Farber, 2006), as is evidenced by the distribution in Table 2, which guided our decision to collect annotations along a continuum. Researchers may be able to leverage these continuous annotations directly in future work. However, work to date has framed the problem as a classification rather than regression task (Balani and De Choudhury, 2015; Umar et al., 2019). In following earlier precedent (Balani and De Choudhury, 2015), we frame our self-disclosure task as a multi-class classification problem, facilitating comparison with prior computational work. We binned our score ranges as follows to produce three classes: [0-1] NO SELF-DISCLOSURE, (1-4) POSSIBLE SELF-DISCLOSURE, and [4-5] CLEAR SELF-DISCLOSURE.

Examples from each class are shown in Table 3. We leave the development of true regression models for predicting continuous medical self-disclosure scores to future work,[7] and release both the averaged (and thus continuous) scores and our discretized class labels with our dataset.

Class	Example	Description
NO SD	1. *I wish you all the strength x.* 2. *Cheers and happy new year!*	No disclosure of medical issue.
POSSIBLE SD	1. *I'm not angry, I'm not even sad as such, I'm just tired...* 2. *I do think my rib pain is from bad posture. I have worked at a computer for years.*	General, non-specific mention of or allusion to medical issue.
CLEAR SD	1. *Metoprolol gave me the most horrendous headaches, so I had my doctor take me off.* 2. *I did the ultrasound a couple times now since this started 2 yrs ago I'd like to find a good ortho doc.*	Clear disclosure of specific symptom, diagnosis, and/or treatment.

Table 3: Medical self-disclosure class descriptions and corresponding examples.

3.4 Data Privacy and Permissions

To preserve user privacy, we did not download usernames or other metadata during our data collection process. We further manually reviewed all posts and replaced any names appearing directly within the text with a generic NAME_TOKEN. The *patient.info* terms and conditions maintain public accessibility of forum posts, and allow use of content in non-commercial contexts.[8] Public Facebook posts may be freely downloaded, accessed, and reshared both on and off the platform,[9] and the same applies to public Reddit posts.[10] Twitter's data policy stipulates that only tweet IDs, not fully hydrated tweets, be shared with third parties.[11] Thus, for Twitter data we provide tweet IDs and corresponding labels, and encourage interested individuals to download the tweet text for their own research use.

4 Methods

To demonstrate efficacy and learnability of our dataset, we created a suite of classification models for comparative analysis. This offered the parallel

[7] Our early pilot experiments suggest that this is a challenging task, due in part to an uneven distribution of labels at that level of granularity for which straightforward solutions (e.g., data augmentation techniques) yield somewhat diminished prediction quality.

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

[9] https://www.facebook.com/policy.php

[10] https://www.redditinc.com/policies/privacy-policy

[11] https://developer.twitter.com/en/developer-terms/policy

opportunity to identify a strong performance benchmark for this task. We describe our preprocessing techniques and modeling algorithms below.

4.1 Data Preprocessing

Prior to training our models, we applied the following preprocessing steps to our data:

1. **DeEmojifying:** Emojis are often used to express emotion on online platforms, and since emotional content may provide valuable clues to the presence of self-disclosure (Eisner et al., 2016; Felbo et al., 2017; Coppersmith et al., 2016), we retained emojis and converted them to text. Each emoji is represented as its CLDR short name.[12] For example, a happy face with a Unicode of U+1F600 would be converted to *[grinning face]*.

2. **Number Replacement:** The presence of numbers may likewise be indicative of medical content in a post (e.g., *I've always started on 20mg (albeit with side effects for the first few weeks)*). However, we hypothesized that retaining value specificity (e.g., "**20**mg") may produce too much noise to yield high value. We thus replaced all numbers with a single NUMBER_TOKEN.

3. **Stopword Removal:** We removed stopwords using a modified version of the NLTK (Bird, 2006) English stopwords list. Since some words, such as personal pronouns, may signify the presence ([*I*, *my*, *myself*, *me*, *mine*]) or absence ([*you*, *your*, *yours*, *yourself*, *yourselves*, *he*, *his*, *him*, *himself*, *she*, *her*, *hers*, *herself*]) of self-disclosure, we retained them. Likewise, auxiliary verbs may not have significant individual importance, but could switch self-disclosure class. For example, *I have depression* has higher self-disclosure than *I might have depression*.

4. **Punctuation Removal:** Since most punctuation marks are unimportant to our task, we removed them, retaining only sentence boundary markers ([*!*, *.*, *?*]). Question marks in particular could change high self-disclosure to a lower category. For example, *I have depression* could be interpreted quite differently from *I have depression?*

[12] https://unicode.org/emoji/charts/full-emoji-list.html

Technique	Accuracy
Base Model (No Preprocessing)	78.62%
Base + DeEmojifying	80.01%
Base + Number Replacement	80.82%
Base + Stopword Removal	80.79%
Base + Punctuation Removal	79.81%
Base + Spelling Correction	75.62%

Table 4: Model performance in accuracy (%) before and after applying each preprocessing technique. *Base model* refers to our highest-performing model (§5.2).

We initially experimented with spelling correction as an additional preprocessing step, but ultimately abandoned it since it reduced performance. Inaccurate corrections (e.g., *dr* → *dry*) led to considerable, and often detrimental, changes in predicted class values. We present an empirical analysis of these preprocessing steps in Table 4 to illustrate their relative merits.

4.2 Model

We experimented with multiple supervised machine learning methods for our task. We considered the following classification models:

- **Support Vector Machine (SVM):** SVM is a classical machine learning model that has achieved a very high success rate in text classification (Forman, 2008). We applied a linear kernel and kept the penalty parameter C at a default value of 1.0.

- **Naive Bayes (NB):** Naive Bayes is another classical machine learning method that has proven to be useful for a wide range of text classification tasks (Kim et al., 2006).

- **Long Short Term Memory (LSTM):** Neural networks are capable of achieving strong performance in many text classification problems, with LSTM models being particularly adept at tasks relying on sequential data (Gers et al., 2000). We used the following fine-tuned hyperparameters: *learning rate* = 0.001, *batch size* = 64, *dropout* = 0.5, *max sequence length* = 286, and *optimizer* = Adam.

- **Bidirectional LSTM (BLSTM):** BLSTMs are an extension of traditional LSTMs that

consider both prior and forthcoming information in a sequence, allowing them to improve sequential text classification performance (Wöllmer et al., 2010). We used the following fine-tuned hyperparameters: *learning rate* = 0.0003, *batch size* = 64, *dropout* = 0.2, *max sequence length* = 286, and *optimizer* = Adam.

- **1D-Convolutional Neural Network (1D-CNN):** Convolutional neural networks have achieved exceptional performance for many text classification problems (Kim, 2014). We used the following fine-tuned hyperparameters: *learning rate* = 0.0002, *batch size* = 32, *dropout* = 0.3, *max sequence length* = 286, and *optimizer* = Adam.

- **DistilBERT:** DistilBERT (Sanh et al., 2019) is a lightweight Transformer-based model. It was designed as a variation of BERT (Devlin et al., 2018) that is well-suited for tasks utilizing smaller datasets. We used the following fine-tuned hyperparameters: *learning rate* = 0.003, *batch size* = 32, and *epochs* = 40.

We also compare these models to two additional approaches:

- **Baseline:** Predicts a constant label (CLEAR SD, the highest frequency label in the dataset) for every record. This allowed us to validate that our models were able to learn to predict medical self-disclosure using our novel dataset at a rate higher than chance.

- **Balani and De Choudhury (2015):** Our reimplementation of Balani and De Choudhury's best-performing self-disclosure model, fine-tuned for our dataset and task. This allowed us to compare our model performance directly with a high-performing existing model for self-disclosure detection, and subsequently provide empirical justification that detecting self-disclosure within our task domain carries its own uniquely challenging, subtle complexities.

We applied sequence padding for all deep learning models, padding sentences with zeroes to normalize length. The maximum sequence length (maximum number of tokens) of the instances in our dataset is 286, and thus we padded all shorter instances to reach that length. We used TF-IDF vectors with a vocabulary size of 5000 words (Zhang et al., 2011) for the classical machine learning models, optimizing the vocabulary size on a held-out validation set and retaining the 5000 most-frequent words. We used 100-dimensional GloVe (Pennington et al., 2014) word embeddings pretrained on Wikipedia 2014 and Gigaword 5 for the deep learning models.[13]

We randomly split the data into training (80%), validation (10%), and test (10%) subsets, training the models on the training data and fine-tuning them on the validation set to optimize hyperparameters. Since weights for our deep learning models were randomly initialized, we repeated this process multiple times for each model, performing five-fold Monte Carlo cross-validation (Xu and Liang, 2001) and reporting the averaged results. We optimized hyperparameters using grid search.

4.3 Classification Settings

In addition to experimenting with a variety of statistical and neural classification models, we experimented with two classification settings: (1) a binary classification setting, and (2) our target multinomial classification setting. We did so in light of our observation that POSSIBLE SELF-DISCLOSURE exhibited noticeably lower inter-annotator agreement than the two classes at the respective ends of the self-disclosure spectrum (see Table 1). We anticipated that automated self-disclosure models would similarly struggle more with this class.

In the binary setting, we only trained and evaluated our models using data from the NO SELF-DISCLOSURE and CLEAR SELF-DISCLOSURE classes. This had the effect of simplifying the task greatly, but it was also less realistic—in the real world, as shown in the class distribution for our dataset, many instances may be more ambiguous and fall somewhere between the two endpoints of the self-disclosure spectrum. In our more challenging multinomial setting (the setting upon which we placed our primary focus) we retained all three classes: NO SELF-DISCLOSURE, POSSIBLE SELF-DISCLOSURE, and CLEAR SELF-DISCLOSURE. We applied the same hyperparameters specified in §4.2 (fine-tuned under the multinomial classification setting) to models in both settings.

[13] Word embeddings represent words as n-dimensional feature vectors and capture latent patterns in meaning, semantic relationships, and the context in which words are used (Collobert et al., 2011).

Model	Acc.	Precision	Recall	F_1
Binary	84.75	0.8938	0.8486	0.8623
Multi	**81.02**	**0.8084**	**0.8102**	**0.8089**

Table 5: Comparison between binary and multiclass DistilBERT models. Accuracy shown as a percentage (%).

5 Evaluation

We evaluated the performance of all models using accuracy, precision, recall, and F1-measure, following prior work on self-disclosure detection (Balani and De Choudhury, 2015; Umar et al., 2019). We provide the results from three separate experiments in the following subsections. In §5.1, we compare performance between the binary and multinomial classification settings. In §5.2, we compare performance between our SVM, NB, LSTM, BLSTM, 1D-CNN, and DistilBERT models for the multiclass setting. Finally, in §5.3, we provide external validation for our highest-performing multinomial model by comparing it to the baseline and Balani and De Choudhury's highest-performing model.

5.1 Binary vs. Multinomial Self-Disclosure Classification

We compare the performance of our binary and multiclass DistilBERT models (the highest-performing models for binary and multinomial classification) in Table 5. Unsurprisingly, the binary DistilBERT model outperforms its multiclass counterpart; as predicted, the model was able to learn to distinguish between NO SELF-DISCLOSURE and CLEAR SELF-DISCLOSURE with relatively little trouble, much like human annotators. The multiclass DistilBERT model struggled slightly more but nonetheless still exhibited strong overall performance, dropping only 3.73% in absolute accuracy compared to the binary classification setting. We demonstrate later (see Table 8) that a much larger relative percentage of instances from the POSSIBLE SELF-DISCLOSURE class were misclassified than were instances from the other two classes, suggesting ample room for future work that disentangles the nuances of these more ambiguous cases.

5.2 Model Comparison

We present the results of our model comparison for the multinomial classification setting in Table 6. DistilBERT achieved the best performance overall

Model	Acc.	Precision	Recall	F_1
SVM	71.18	0.5945	0.5963	0.5632
NB	67.22	0.4510	0.5197	0.4803
LSTM	74.40	0.7937	0.6582	0.7179
BLSTM	72.89	0.7565	0.6621	0.7052
Distil-BERT	**81.02**	**0.8084**	**0.8102**	**0.8089**
1D-CNN	71.29	0.7493	0.6592	0.7003

Table 6: Model comparison for the multinomial classification setting. Accuracy shown as a percentage (%).

with an accuracy of 81.02%, precision of 0.8084, recall of 0.8189, and F_1-score of 0.8089. In general, the deep learning models outperformed the standard classification models for this task, with DistilBERT outperforming the highest-performing standard classification model (SVM) by relative percent increases in accuracy, precision, recall, and F1-measure by 13.82%, 35.97%, 35.87%, and 43.62%, respectively.

5.3 External Validation

As mentioned earlier, Balani and De Choudhury (2015) detected three grades of self-disclosure in Reddit posts. Their task has similarities with ours, with ours focusing on medical self-disclosure specifically and theirs targeting more general disclosure of mental wellness. Although we were unable to directly acquire their data or source code, we reimplemented their best model and fine-tuned it such that it maximized performance on our dataset and task. Our motivation in performing this experiment was to establish that models designed for general self-disclosure do not necessarily generalize to the additional subtle complexities of medical self-disclosure, and correspondingly that different forms of self-disclosure should be managed differently in automated systems. In Table 7 we compare the results achieved by (1) the most frequent class baseline, (2) our best-performing multinomial model, and (3) our reimplementation of Balani and De Choudhury's best-performing model. Our model outperforms both the baseline and Balani and De Choudhury's model by a wide margin, with relative percentage increases of 41.84%, 32.63%, 66.60%, and 49.76% for accuracy, precision, recall,

Model	Acc.	Precision	Recall	F_1
Baseline	45.38	0.4471	0.4489	0.4479
Distil-BERT	**81.02**	**0.8084**	**0.8102**	**0.8089**
Balani and De Choudhury	57.12	0.6095	0.4863	0.5401

Table 7: Comparison between the baseline, our best (multinomial) model performance, and our replication of Balani and De Choudhury's model (2015). Accuracy is shown as a percentage (%).

Class	# Test Samples	# TP	Accuracy
NO SD	398	349	87.68
POSSIBLE SD	153	68	44.44
CLEAR SD	445	390	87.64

Table 8: Total number of test samples per class, number of true positives per class, and overall class accuracy. Accuracy is shown as a percentage (%).

and F1-measure, respectively, over Balani and De Choudhury's model.

6 Discussion

Although Balani and De Choudhury's model worked well for their setting, we found that it did not transfer well to our task. It may be that detecting medical self-disclosure inherently carries extra levels of complexity. For example, identifying first-person pronouns could be a decisive indicator of general self-disclosure, whereas for medical self-disclosure, self-identifiers would also need to be accompanied by medical terms, some of which may be obscure (Meystre et al., 2008).

To further disentangle the performance of our highest-performing model, we computed the number of true positives for each class separately, shown alongside per-class accuracy in Table 8. We found that model performance was lowest when predicting POSSIBLE SELF-DISCLOSURE. This was anticipated due to the difficulty of agreeing upon labels for this class even among trained annotators (refer to Table 1 for per-class agreement statistics); in many cases, only one annotator may have felt that an instance clearly disclosed a medical issue, with others being less certain. Performance was high for NO SELF-DISCLOSURE and CLEAR SELF-DISCLOSURE, with accuracies of 87.68% and 87.64%, respectively. Since cases of POSSIBLE SELF-DISCLOSURE may comprise a sizeable contingent of data instances (slightly over 15% of the dataset in our case), we recommend that this subset of data is examined more closely in follow-up work. Downstream applications may need to handle these more ambiguous cases differently from incidences in which symptoms, diagnoses, or treatments clearly are (or clearly are not) being disclosed.

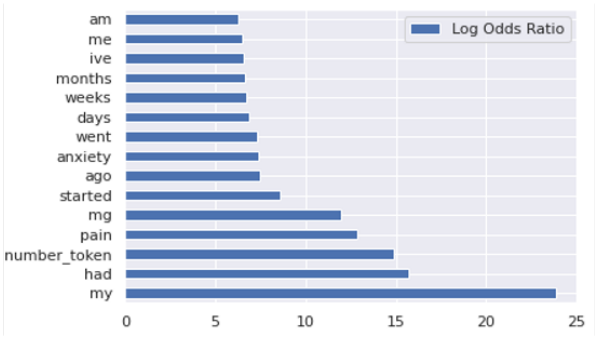

Figure 1: Words most closely associated with CLEAR SELF-DISCLOSURE. The x-axis shows the log odds ratio.

Figure 2: Words most closely associated with POSSIBLE SELF-DISCLOSURE. The x-axis shows the log odds ratio.

To develop a further understanding of the linguistic patterns associated with CLEAR SELF-DISCLOSURE, POSSIBLE SELF-DISCLOSURE, and NO SELF-DISCLOSURE instances, we computed the log odds ratio with an informative Dirichlet prior (Monroe et al., 2008; Hessel, 2016) for words in these classes to assess which words were most strongly correlated with each, and plot them in Figures 1, 2, and 3. The plots support our hypotheses. The words most closely associated with POSSIBLE SELF-DISCLOSURE have much lower

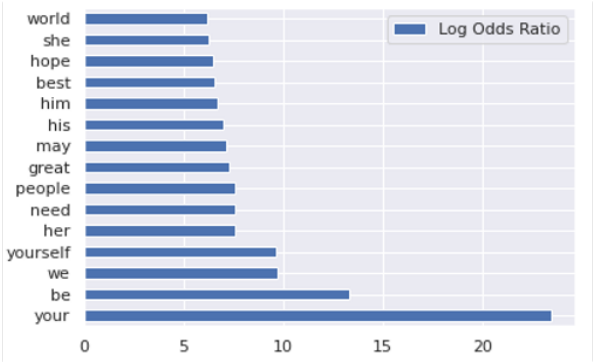

Figure 3: Words most closely associated with NO SELF-DISCLOSURE. The x-axis shows the log odds ratio.

ratios in general than the words most closely associated with CLEAR SELF-DISCLOSURE or NO SELF-DISCLOSURE, suggesting that this class is characterized by fewer strong cues indicating membership. Furthermore, while the words closely associated with CLEAR SELF-DISCLOSURE are a mix of personal pronouns, medical terms, duration, and narrative descriptors, the words most closely associated with POSSIBLE SELF-DISCLOSURE are mostly about others and family, or being scared and in search of hope and support. Words closely associated with NO SELF-DISCLOSURE are less personal or narrative, and more indicative of support or general health interest.

7 Conclusion

In this work, we introduced a novel medical self-disclosure dataset containing 6,639 instances collected from public online social platforms. Instances in this dataset are triple-annotated with high inter-annotator agreement (κ=0.88) for NO SELF-DISCLOSURE, POSSIBLE SELF-DISCLOSURE, and CLEAR SELF-DISCLOSURE. We evaluated a wide range of classical machine learning and neural classifiers (including LSTM-, CNN-, and Transformer-based models) to assess their efficacy at learning to predict medical self-disclosure. We examined both a simpler binary classification setting and a more challenging multinomial setting, finding that the highest-performing model in both cases was a fine-tuned DistilBERT model.

We compared our best-performing model to the best existing categorical model for self-disclosure detection (Balani and De Choudhury, 2015), finding that our model outperformed that model by a wide margin for the task of detecting medical self-disclosure (relative percent increases of 41.84% and 49.76% for accuracy and F1-measure, respectively). Our findings pave the way for subsequent experiments with other models, moving the dial a necessary step forward by establishing a strong performance benchmark. In the future, we hope to explore medical self-disclosure in the context of goal-oriented dialogue systems, resulting in downstream benefits for both physicians and patients. We make our dataset available to interested researchers to foster further progress on this emerging research task.

8 Ethical Considerations

This research was approved by the Institutional Review Board at the University of Illinois at Chicago. All data was collected in a manner consistent with the terms and conditions of the respective data sources, as outlined in §3.4. In particular, since Twitter's data policy prohibits direct sharing of tweet text, we release only tweet IDs and corresponding annotations for that subset of the data. Annotations were collected using the process described in §3.2, and annotators were compensated for their work through assistantships and course (independent study) credit. Additional characteristics of the data are provided in §3.1 and §3.3. Instances have been anonymized, with any usernames or other personal names found in the text replaced with a generic NAME_TOKEN, to further promote privacy of content creators when possible (this is not possible with the tweets since they are provided as stand-off annotations). Data is available upon request by emailing the authors, and posts known or assumed to be deleted at the time of request will be removed prior to sharing. We will communicate further data use guidelines as outlined in our IRB protocol directly when sharing the data.

Acknowledgements

We thank Yasmin Isa for her role in creating the dataset and in participating in helpful research discussions along the way. We also thank the anonymous reviewers for their insightful suggestions, which further strengthened this work. This work was supported in part by a startup grant from the University of Illinois at Chicago.

References

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

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

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.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Nancy Bridges. 2001. Therapist's self-disclosure: Expanding the comfort zone. *Psychotherapy: Theory, Research, Practice, Training*, 38:21–30.

Olivia E Bruss and Jennifer M Hill. 2010. Tell me more: Online versus face-to-face communication and self-disclosure. *Psi Chi Journal of Undergraduate Research*, 15(1).

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 106–117, San Diego, CA, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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.

Darren Ellis and John Cromby. 2012. Emotional inhibition: A discourse analysis of disclosure. *Psychology & health*, 27(5):515–532.

Barry A. Farber. 2006. *Self-disclosure in psychotherapy*. Guilford Press.

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.

Michael Ferriter. 1993. Computer aided interviewing and the psychiatric social history. *Social Work and Social Sciences Review*, 4(3):255–263.

George Forman. 2008. Bns feature scaling: an improved representation over tf-idf for svm text classification. In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 263–270, New York, NY, USA. ACM.

Felix A. Gers, Jürgen Schmidhuber, and Fred A. Cummins. 2000. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471.

John H. Greist, Marjorie H. Klein, and L J Van Cura. 1973. A computer interview for psychiatric patient target symptoms. *Archives of general psychiatry*, 29 2:247–53.

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

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.

Adam N Joinson and Carina B Paine. 2007. Self-disclosure, privacy and the internet. *The Oxford handbook of Internet psychology*, 2374252.

Toksoz B Karasu. 1979. Psychotherapy of the psychosomatic patient. *American journal of psychotherapy*, 33(3):354–364.

Robert Kellner. 1975. Psychotherapy in psychosomatic disorders: a survey of controlled studies. *Archives of General Psychiatry*, 32(8):1021–1028.

Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung-Hyon Myaeng. 2006. Some effective techniques for naive bayes text classification. *IEEE Trans. Knowl. Data Eng.*, 18(11):1457–1466.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

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

Ruth Lewy. 2013. 50 top websites you can't live without. The Times. Accessed: 2021-03-31.

W.W. Meissner. 2002. The problem of self-disclosure in psychoanalysis. *Journal of the American Psychoanalytic Association*, 50:827–67.

Kenneth George Andrew Meleshko and Lynn E. Alden. 1993. Anxiety and self-disclosure: toward a motivational model. *Journal of personality and social psychology*, 64 6:1000–9.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 17(01):128–144.

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.

James W. Pennebaker and Cindy K. Chung. 2007. Expressive writing, emotional upheavals, and health. In *Foundations of Health Psychology*, pages 263–284. Oxford University Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. 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.

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.

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

S. Courtney Walton and Ronald E. Rice. 2013. Mediated disclosure on twitter: The roles of gender and identity in boundary impermeability, valence, disclosure, and stage. *Comput. Hum. Behav.*, 29(4):1465–1474.

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.

Martin Wöllmer, Florian Eyben, Alex Graves, Björn W. Schuller, and Gerhard Rigoll. 2010. Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework. *Cogn. Comput.*, 2(3):180–190.

Qingsong Xu and Yi-Zeng Liang. 2001. Monte carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56:1–11.

Wen Zhang, Taketoshi Yoshida, and Xijin Tang. 2011. A comparative study of tf*idf, LSI and multi-words for text classification. *Expert Syst. Appl.*, 38(3):2758–2765.