Towards Intelligent Clinically-Informed Language Analyses of People with Bipolar Disorder and Schizophrenia

Ankit Aich[1], **Avery Quynh**[2], **Varsha Badal**[2], **Amy Pinkham**[3], **Philip Harvey**[4], **Colin Depp**[2], and **Natalie Parde**[1]

[1]Department of Computer Science, University of Illinois Chicago
{aaich2, parde}@uic.edu
[2]Department of Psychiatry, University of California San Diego
{akquynh, vbadal, cdepp}@health.ucsd.edu
[3]School of Behavioral and Brain Sciences, The University of Texas at Dallas
amy.pinkham@utdallas.edu
[4]University of Miami Miller School of Medicine
pharvey@miami.edu

Abstract

NLP offers a myriad of opportunities to support mental health research. However, prior work has almost exclusively focused on social media data, for which diagnoses are difficult or impossible to validate. We present a first-of-its-kind dataset of manually transcribed interactions with people clinically diagnosed with bipolar disorder and schizophrenia, as well as healthy controls. Data was collected through validated clinical tasks and paired with diagnostic measures. We extract 100+ temporal, sentiment, psycholinguistic, emotion, and lexical features from the data and establish classification validity using a variety of models to study language differences between diagnostic groups. Our models achieve strong classification performance (maximum F_1=0.93-0.96), and lead to the discovery of interesting associations between linguistic features and diagnostic class. It is our hope that this dataset will offer high value to clinical and NLP researchers, with potential for widespread broader impacts.

1 Introduction

Schizophrenia and bipolar disorder have been associated with observable language patterns in clinical and sociolinguistic studies (Kasanin, 1944; Elmore and Gorham, 1957; Perlini et al., 2012; Bambini et al., 2016). Some computational studies have sought to replicate these findings or derive novel clinical insights using automated analyses driven by natural language processing techniques (Ratana et al., 2019; Harvey et al., 2022). Effectively identifying features associated with these disorders or providing diagnostic aid offers substantial potential for real-world impact (Castro et al., 2015; Becker et al., 2018; Lovejoy, 2019). However, these studies to date have been constrained by limitations in dataset size and availability (Elvevåg et al., 2007; Bedi et al., 2015; Mota et al., 2012; Gutiérrez et al., 2017; Corcoran and Cecchi, 2020, $n \leq 51$ subjects), restricting the extent to which they can produce meaningful or generalizable conclusions.

We address this gap by introducing a new, large ($n = 644$ subjects) dataset of transcribed conversations between clinicians and people with *bipolar disorder* (BD), people with *schizophrenia* (SZ), and healthy control (HC) subjects. We also establish preliminary benchmarking models for automatically distinguishing between these groups using interpretable linguistic features, achieving promising proof-of-concept ranging from 70–96% accuracy in one-versus-one discrimination between subject groups. Finally, we conduct preliminary analyses across a large feature set to identify potential linguistic correlates with these groups. Our key contributions are as follows:

- We introduce a new, 644-subject (1288-transcript) dataset collected in clinically validated laboratory settings.
- Using this new dataset, we develop benchmarking models for the automated detection of bipolar and schizophrenia disorders in a one-versus-one classification setting, as a tool for facilitating analysis of language associated with members of these groups.
- Through these analyses, we identify potential linguistic correlates with diagnostic groups.

This research was jointly conducted by an interdisciplinary team of researchers from psychiatry and computer science departments to foster translational impact in both communities (Newman-Griffis et al., 2021). We hope that the data and

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2871 - 2887
December 7-11, 2022 ©2022 Association for Computational Linguistics

insights provided will pave the way for new research and subsequently exciting new clinical and computational findings in this domain.

2 Background

Social media data has dominated research at the intersection of computational linguistics and clinical psychology (Bucci et al., 2019). Its popularity owes partially to its size and availability (Perrin, 2015; Fuchs, 2015; Graham et al., 2015). High rates of social media usage are also evident in users who face mental health concerns (Gowen et al., 2012; Birnbaum et al., 2015), with associations observed between social media use and the occurrence of psychosis (Kalbitzer et al., 2014; Krishna et al., 2012; Nitzan et al., 2011), mood disorders (Lin et al., 2016; Pantic et al., 2012), personality disorders (Rosen et al., 2013), eating disorders (Mabe et al., 2014; Smith et al., 2013), and obsessive compulsive disorder (Lee et al., 2015).

The observed connections between social media and mental health have triggered meaningful work into leveraging this data for downstream tasks (Aich and Parde, 2022), such as the automated detection of depression (Morales et al., 2018; Husseini Orabi et al., 2018), schizophrenia and psychosis (Zomick et al., 2019; Bar et al., 2019), and suicide risk prediction (Zirikly et al., 2019a; Matero et al., 2019a). Reddit posts, a popular resource for this work due to their semi-anonymity and length (Zirikly et al., 2019b), have been used to identify stress (Turcan and McKeown, 2019a), eating disorders (Yan et al., 2019; Trifan and Oliveira, 2019), depression (Tadesse et al., 2019), and suicide (Zirikly et al., 2019b; Matero et al., 2019b), among others (Sekulic and Strube, 2019). Twitter has been used for the detection of depression and post-traumatic stress disorder (Amir et al., 2019; Kirinde Gamaarachchige and Inkpen, 2019), schizophrenia (Ernala et al., 2019), anti-social behavior (Singh et al., 2020), suicidal ideation (Wang et al., 2016a; Shahreen et al., 2018), and stress (Winata et al., 2018).

Most social media datasets for mental health tasks are annotated along binary or linear scales and label users based on analysis of a set number of posts. Annotations may be provided by trained human annotators (Wang et al., 2016b; Coppersmith et al., 2015), annotators with clearly referenced domain expertise (e.g., Birnbaum et al. (2017)'s work employing a clinical psychiatrist and a graduate student from Northwell Health's Early Treatment Program), user disclosures of mental health conditions (Coppersmith et al., 2015; Safa et al., 2022; Zhou et al., 2021), and crowdsourcing services (Turcan and McKeown, 2019b). Annotation schema for some mental health conditions can be subjective, causing varied inter-annotator agreement. For example, Birnbaum et al. (2017) reported a Cohen's kappa score of κ=0.81, whereas Turcan and McKeown (2019b) reported a much lower agreement of κ=0.47 for their dataset of stressed and unstressed social media users. Turcan and McKeown (2019b)'s dataset also offers an example of how fuzzy label boundaries can affect annotation quality—it is well established that stress is often temporary (Dhabhar, 2018); hence, post labels do not always equate to a user's mental state. Finally, independent decision-making when selecting sources may influence annotation outcomes.

Although social media data has been leveraged for a variety of mental health tasks, data accessibility remains an enormous challenge. In their analysis of more than 100 mental health datasets, Harrigian et al. (2020) found only three to be available without any restrictions. They found that $\geq 50\%$ of the data they analyzed was not readily available.[1] Of those that were described in some capacity (48), 13 were removed from public records or limitations made them unavailable. Out of the 35 that remained, 12 needed signed agreements or Institutional Review Board (IRB) approvals, 18 had instructions and APIs to reproduce them, 2 could be obtained directly by emailing the authors, and as mentioned, 3 were available without restrictions. These trends have also been observed on a broader scale with other healthcare data in NLP studies (Valizadeh and Parde, 2022).

Moreover, for publicly accessible data, the inherent subjectivity of many mental health annotation tasks and the frequent reliance on user self-disclosures means that many "gold standard" labels are imperfectly assigned. Most datasets fail to capture nuances of mental health (Arseniev-Koehler et al., 2018), and medical self-disclosures may be indirect (Valizadeh et al., 2021). For example, Birnbaum et al. (2017)'s dataset labels the following sample as *YES*, but provides little clarity regarding the user's diagnosis:

[1] Only 48 of 102 datasets were described to such an extent that they could be analyzed for availability, naturally suggesting that the others were fully inaccessible.

> *I have schizophrenia/depression. I am trying to become better by exercise and working I have a job xoxo I love Saturday xx*

Issues related to fairness, gender, balance, and representation of racial and ethnic biases in social media datasets have also been found (Aguirre et al., 2021). We seek to address many of these limitations by providing a publicly accessible dataset of manually transcribed interactions between individuals with clinically diagnosed mental health conditions and trained clinicians. We also provide dataset transparency regarding representational balance through validity of diagnoses and descriptive statistics.

3 Data

3.1 Task Selection

We collected data through a standardized performance-based test of social competence called the Social Skills Performance Assessment (Patterson et al., 2001, SSPA). The SSPA involves a prompted conversation between a confederate/examiner and a patient, wherein the patient's social abilities during the conversation are scored by a trained rater to provide an estimate of social skill. The SSPA is useful in clinical assessment because it provides a measure of social abilities that is free of biases associated with self-report or informants (Leifker et al., 2010). The SSPA has been used as an endpoint of clinical rehabilitation trials and is a predictor of social function (Miller et al., 2021).

The SSPA involves two scenarios administered by a trained rater in a laboratory setting, and the interaction is audiorecorded. The measure consists of two simulated interactions in which the rater plays the role of a conversation partner and the participant plays the role of themselves in the scene. The first scene is affiliative and involves meeting a new neighbor. The second scene is confrontational and asks the participants to complain to their landlord, after a prior notification about a leak had not been addressed. These scenarios last on average four minutes each. In Appendix A we provide sample texts for both scenes from people who are clinically diagnosed with schizophrenia.

3.2 Collection

Data was collected during three projects supported by the National Institute of Mental Health, each of which recruited outpatients with either schizophrenia/schizoaffective disorder or bipolar disorder or healthy controls. The inclusion criteria for these studies involved ability to provide informed written consent, diagnosis of either bipolar disorder or schizophrenia/schizoaffective disorder according to the Diagnostic and Statistical Manual of the American Psychiatry Association, and outpatient status at the time of assessment. Informed written consent was taken from participants for audiorecording and de-identified research data sharing for each of these projects. Psychiatric diagnoses were performed under supervision of medical researchers and practicing clinicians at the University of California San Diego, the University of Miami, and the University of Texas at Dallas. A total of 644[2] SSPAs were available across these studies (SZ/SC=247, BD=286, HC=110).

Category	Value
Mean Age	44.2
σ(Mean Age)	11.4
Females	58.4%
Males	41.3%
Unspecified	0.3%
African Americans	37.4%
American Indian/ Alaskan Native	0.5%
Asian	5.4%
White	48.3%
Multirace	7.0%
Hawaiian	0.6%
Unreported	0.6%

Table 1: Descriptive statistics for the participant pool. Age and its standard deviation are provided in years. Other demographic details are provided in frequency percentages.

3.3 Descriptive Statistics

We experiment in Section 6 with a random subset of 300 subjects divided equally between the SZ ($n = 100$), BD ($n = 100$), and HC ($n = 100$) groups. Each participant has two audio files (for the two tasks described in §3.1) for a total of 600 audio files. Descriptive statistics for all 644 participants in the full dataset are provided in Table 1.

3.4 Data Release

We release our data freely in two ways. Extracted features (described in §4.2) can be downloaded as

[2] One participant was later found ineligible.

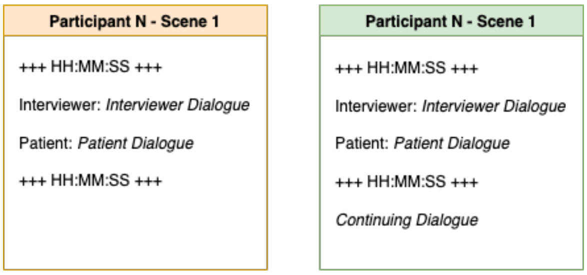


Figure 1: Transcription formats prior to preprocessing. The format at right was used when patient or interviewer utterances exceeded a given timestamp and continued onward into the next dialogue block.

CSV files from Github[3] without any special permission. The fully de-identified transcripts can be downloaded from the National Institute of Mental Health data archive[4] in adherence with National Institutes of Health reporting requirements and the corresponding research grant that funded this work. Users of our data will be responsible for their own statements, analysis, interpretation, and uses. We refer readers to the Ethical Considerations (end of paper) and Appendix C for a fuller understanding of how to use this dataset.

4 Methods

4.1 Preprocessing

Verbatim transcriptions of the audiorecordings for all participants were made by a trusted third-party service and then manually stripped of identifiable information. These were stored in *docx* format by the transcription service, using one of the two formats shown in Figure 1. We preprocessed these files to prepare them for further computational work using a series of steps determined through preliminary data analysis. These steps included the automated extraction of timestamps, separation of interviewer and participant dialogue, and (described in the next subsection) computation of linguistic features inspired by and extended from previously published work on other datasets.

We first converted the transcripts verbatim from *docx* to *txt* format to enable easier parsing using Python 3.7. We then applied a set of regular expressions to extract essential information:

- **Timestamps** were extracted by searching for strings in the format *HH:MM:SS* enclosed by + sign characters.

Algorithm 1 Utterance Speaker Labeling

```
$s_c \leftarrow$ ""                                   ▷ Initialized to empty.
$u_p = [\quad]$
$u_i = [\quad]$
while $l$ is not FALSE do
    $s_p \leftarrow s_c$
    $t_c \leftarrow$ GETTIME$(l)$
    if GETINTERVIEWER$(l)$ is not FALSE then
        Append $l$ to $u_i$
        $s_c \leftarrow$ Interviewer
    else if GETPATIENT$(l)$ is not FALSE then
        Append $l$ to $u_p$
        $s_c \leftarrow$ Patient
    else if $t_c$ is FALSE then                       ▷ No matches.
        if $s_p ==$ Interviewer then
            Append $l$ to $u_i$
            $s_c \leftarrow$ Interviewer
        else if $s_p ==$ Patient then
            Append $l$ to $u_p$
            $s_c \leftarrow$ Patient
        end if
    end if
end while
```

- **Interviewer dialogue** was extracted by searching for strings starting with *Interviewer:*.
- **Patient dialogue** was extracted by searching for strings starting with *Patient:*.

Transcripts following the second format in Figure 1 were more complex to initially parse, since the continuous dialogue extending beyond the initial timestamp was not matched effectively by these patterns. To address this, we applied a speaker labeling algorithm (Algorithm 1) to these cases. This algorithm processes strings using our regular expression patterns, repeatedly iterating through lines in the transcript until the end of the document is reached. The variable t_c holds the current timestamp for the speaker utterances, l holds the current line of text (set to FALSE if no more lines exist in the document), s_p holds the previous speaker label, s_c holds the current speaker label, u_p holds patient utterances, and u_i holds interviewer utterances.

The functions GETTIME$(\cdot)$, GETINTERVIEWER$(\cdot)$, and GETPATIENT$(\cdot)$ hold the regular expressions necessary to extract the timestamp, interviewer label, and patient label from a string, respectively, or otherwise return FALSE. Strings

[3] https://github.com/ankitaich09/EMNLP_2022

[4] https://nda.nih.gov

matched by GETINTERVIEWER(·) or GETPATIENT(·) are appended to u_i or u_p depending on the specified speaker, and strings not matched by any of the regular expression patterns (e.g., continued dialogue) are appended to the previous speaker's utterance list. The final, preprocessed lists of interviewer and patient utterances with extracted timestamps are converted to pandas[5] dataframes for feature extraction and further processing.

4.2 Features Extracted

To assess the importance and utility of linguistic features in the context of this new, large dataset, we extract varied features from the patient dialogue. These features can be broadly categorized as pertaining to time, sentiment, psycholinguistc attribute, emotion, and lexical diversity.

4.2.1 Temporal Features

We extracted two temporal features for each patient: the maximum time taken for a dialogue, and the mean time taken per dialogue. To do so, all timestamp strings were first converted to time objects in seconds, allowing for straightforward calculation of the difference between start and end times in a given dialogue. The maximum difference is labeled as the *max_time*. The mean is taken from this list of differences and is our other temporal feature *mean_time*. These numbers are stored in seconds.

4.2.2 Sentiment Features

We extracted sentiment features based on SentiWordNet (Baccianella et al., 2010) scores. We calculated a transcript-level *total_sentiment_score* by concatenating all patient utterances in the transcript, tokenizing the concatenated text, and computing token-level scores that were then used to increment positive, negative, or objective features across the full transcript. We then extract the *average_positive*, *average_negative*, and *average_objective* scores from this information.

4.2.3 Psycholinguistic Features

To compute psycholinguistic features, we used the 2022 Linguistic Inquiry and Word Count (LIWC) framework (Boyd et al., 2022), which offers key updates over existing versions of LIWC. Specifically, the processes for computing classical LIWC features such as *WC*, *Analytic*, *Clout*, *Authentic*, and *Tone* are changed to reflect shifts in culture and in social sciences, while still correlating with their previous implementations from the LIWC 2015 framework. We extract the full set of 118 LIWC 2022 features described by Boyd et al. (2022) for each transcript in our dataset.

Feature Name	Formula
Type Token Ratio (Chotlos, 1944; Templin, 1957)	TTR = $\frac{T}{W}$
Root Type Token Ratio (Pierre, 1959)	RTTR = $\frac{T}{\sqrt{W}}$
Corrected Type Token Ratio (Carol, 1964)	CTTR = $\frac{T}{\sqrt{2W}}$
Herdan's Lexical Diversity (Herdan, 1960)	HLD = $\frac{log(T)}{log(W)}$
Summer's Lexical Diversity (Somers, 1966)	SLD = $\frac{log(log(T))}{log(log(W))}$
Dugast's Lexical Diversity (Dugast, 1978)	DLD = $\frac{log(W)^2}{log(W) - log(T)}$
Maas' Lexical Diversity (Mass, 1972)	MLD = $\frac{log(W) - log(T)}{log(W)^2}$

Table 2: Lexical Diversity Features

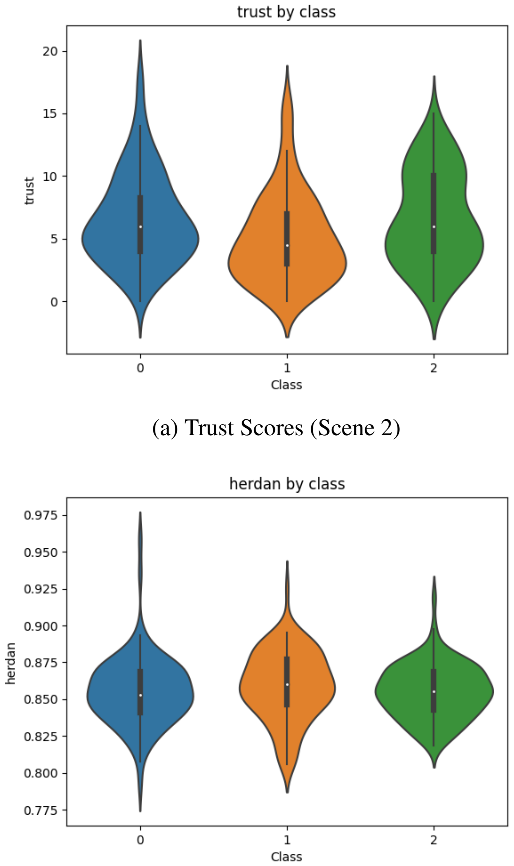
4.2.4 Emotion Features

We extracted emotion features based on the NRC Word-Emotion Lexicon (Mohammad and Turney, 2010, 2013). Specifically, for each transcript we compute the total number of words associated with *Anger*, *Anticipation*, *Disgust*, *Fear*, *Joy*, *Sadness*, *Surprise*, and *Trust* as denoted by the NRC lexicon. We assign a score of 0 for a given emotion if the transcript contains no words corresponding to that emotion in the NRC lexicon.

4.2.5 Lexical Diversity Features

Finally, to measure a transcript's linguistic variety and richness, we computed seven popular measures of lexical diversity at the transcript level. These measures are described in detail in Table 2. Lexical diversity indices have proven crucial in psychometric evaluation tasks (Kapantzoglou et al., 2019).

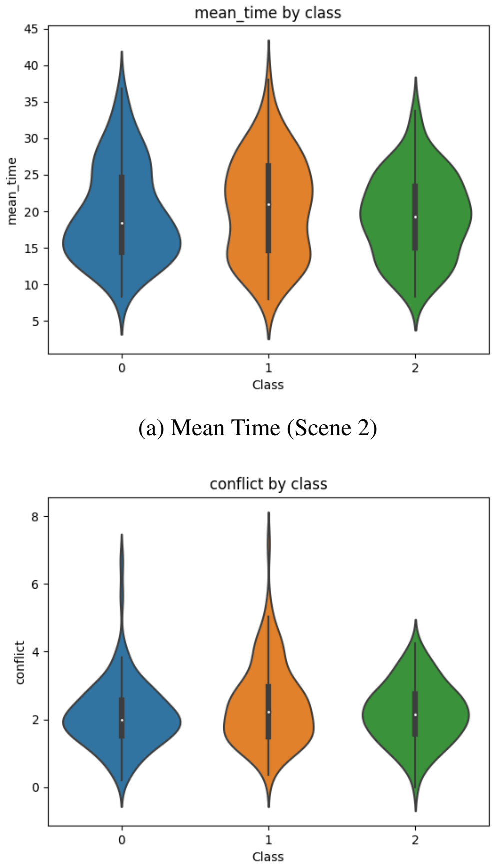
[5] https://pandas.pydata.org



(a) Trust Scores (Scene 2)

(b) Lexical Diversity Scores (Scene 2)

Figure 2: Blue represents healthy controls, orange represents schizophrenia, and green represents bipolar. Figure is best viewed in color. Figure shows violin plots with quartiles, medians, and interquartile ranges across classes *Healthy*, *Schizophrenic*, and *Bipolar*.



(a) Mean Time (Scene 2)

(b) Interpersonal Conflict (Scene 2)

Figure 3: Blue represents healthy controls, orange represents schizophrenia, and green represents bipolar. Figure is best viewed in color. Figure shows violin plots with quartiles, medians, and interquartile ranges across classes *Healthy*, *Schizophrenic*, and *Bipolar*.

5 Feature Analyses

Since we computed features across three subject pools (SZ, BD, and HC), we analyzed feature correlations, patterns, and trends across subject groups. This investigation provides a starting ground for the more detailed follow-up studies that our new dataset is designed to enable. We make our analysis and visualization scripts publicly available to lower the barrier for others to pursue these studies.[3]

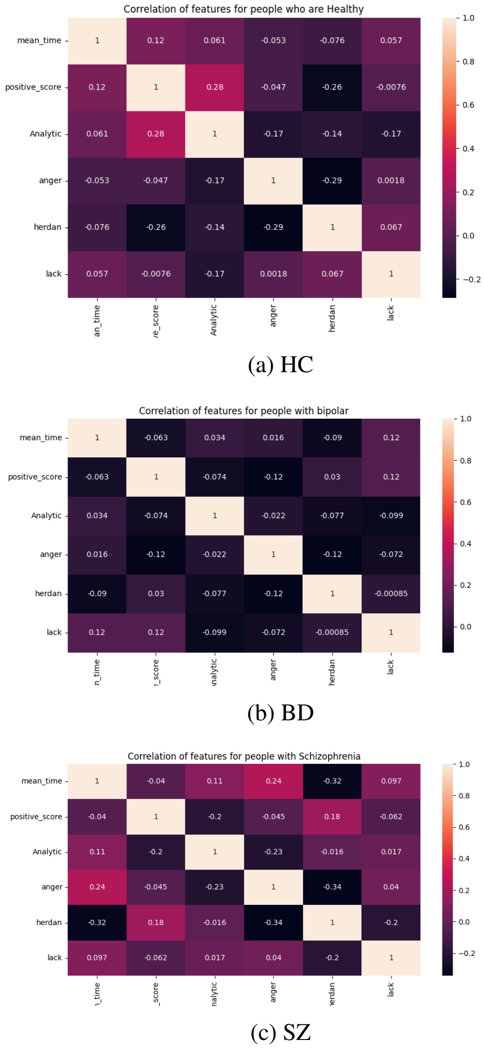
In Figures 2 and 3, we present violin plots illustrating score distributions across selected features from major feature groups described in §4.2. We examine *trust* emotion features (Figure 5a), *Herdan* measures of lexical diversity (Figure 5b and 2b), *mean time* per dialogue (Figure 6a and 3a), and *interpersonal conflict* features from LIWC 2022 (Figure 6b and 3b). Class labels are represented using the numeric signifiers *HC*=0, *SZ*=1, and *BD*=2, and colors blue, orange, and green, respectively. Due to space restrictions we present plots based on the Scene 2 transcripts here, and include plots representing the same features from Scene 1 as supplemental content in Appendix B (Figures 5 and 6).

We observe that HC subjects exhibit larger overall ranges of lexical diversity and *trust* language than SZ or BD subjects (Figure 2). SZ subjects exhibit lower *trust* scores, and BD subjects exhibit a bimodal score distribution with two large frequency centers (Figure 5a and Figure 2a). This differs from patterns associated with lexical diversity. We observe that BD subjects have a single concentrated distribution of mass slightly above a *Herdan* score of 0.85. SZ subjects exhibit a similar mean *Herdan* score, but with a wider score distribution.

When examining *mean time*, we observe that both HC and SZ subjects have slightly bimodal score distributions, with SZ subjects also having the widest score range (Figure 6a and 3a). BD subjects have a single frequency center and relatively consistent frequency spread from 10-30 seconds. Finally, we observe that *interpersonal conflict* features are concentrated near scores of 2 for all subjects, although SZ subjects show the largest score range with a relatively large share of subjects with scores of 4 or greater (Figure 3b and 6b).

In Figure 4, we present pairwise feature correlations among six selected features across our five broad feature categories: *mean time*, *positive sentiment*, *LIWC analytic score*, *anger score*, *Herdan lexical diversity*, and *LIWC lack score*.[6] We study and compare pairwise correlations between members of different subject groups, with feature correlations for HC, BD, and SZ subjects shown in Figures 4a, 4b, and 4c, respectively.

We observe weakly positive correlations between analytic scores and positive sentiment among HC subjects, but very weakly (BD) to weakly (SZ) negative correlations between this same feature pairing among subjects in other groups, suggesting a stronger relationship between logic and optimism in control subjects compared to subjects with bipolar disease or schizophrenia. Interestingly, we also observe stronger positive correlations between anger and mean time, as well as between lexical diversity and positive sentiment, in SZ subjects than in HC or BD subjects. HC subjects have weakly negative correlations between lexical diversity and positive sentiment.



(a) HC

(b) BD

(c) SZ

Figure 4: Heat maps show correlations between features in Scene 2 transcripts among different subject groups. Correlations range from weakly negative (darkest) to strongly positive (lightest).

6 Classification Task

To establish learning validity of our dataset, we designed a simple task to predict subject group membership. Specifically, we conduct binary classification experiments to discriminate between two classes from the set of *HC*, *SZ*, and *BD* subjects. This also creates an additional avenue through which group-level language behaviors can be analyzed (e.g., through learned feature weights). We experiment with both classical (§6.1) and Transformer-based (§6.2) models.

6.1 Classical Models

We experimented with five feature-based models that have demonstrated high efficiency for a variety of language tasks: *random forest* (Xu et al., 2012; Bouaziz et al., 2014; Jurka et al., 2013, RF), *K nearest neighbors* (Yong et al., 2009; Jodha et al., 2018; Trstenjak et al., 2014; Pranckevičius and Marcinkevičius, 2017, KNN), *logistic regression* (Pranckevičius and Marcinkevičius, 2017; Jurka, 2012; Genkin et al., 2007; Lee and Liu, 2003, Logistic), *ridge classifier* (Aseervatham et al., 2011; He et al., 2014, Ridge), and *support vector machine* (Joachims, 2002; Yang, 2001, SVM). We randomly separated our data for each class into 75%/25% train/test splits. Since we used the 300-subject sample defined in §3.3 for these experiments, this meant that the training data for a given scene, for a given subject group pair, included 150 transcripts. The corresponding test set for that scene/pair setting included 50 transcripts. We performed three

[6] We refer interested readers to the LIWC 2022 manual (Boyd et al., 2022) for full descriptions of all LIWC features.

	SCENE 1						SCENE 2					
	BD × *HC*		*BD* × *SZ*		*HC* × *SZ*		*BD* × *HC*		*BD* × *SZ*		*HC* × *SZ*	
Model	**A**	$\mathbf{F_1}$	**A**	$\mathbf{F_1}$	**A**	$\mathbf{F_1}$	**A**	$\mathbf{F_1}$	**A**	$\mathbf{F_1}$	**A**	$\mathbf{F_1}$
RF	**0.93**	0.87	**0.96**	0.84	**0.96**	**0.96**	**0.96**	0.94	**0.92**	**0.96**	**0.70**	**0.93**
KNN	0.58	0.64	0.51	0.59	0.82	0.75	0.37	0.62	0.71	0.69	0.66	0.48
LR	0.89	0.91	0.82	**0.90**	0.89	0.83	0.86	**0.97**	0.89	0.78	0.55	0.62
Ridge	0.89	**0.94**	0.86	0.70	0.93	0.72	0.93	**0.97**	0.78	0.78	**0.70**	0.70
SVM	0.89	0.91	0.86	0.67	0.93	0.72	0.89	**0.97**	0.89	0.79	0.60	0.75

Table 3: Performance comparisons between classifiers on Scene 1 and Scene 2 transcripts. Results show accuracy (A) and F_1 score for one-versus-one classification between BD, SZ, and HC subjects.

classification experiments (*BD* × *HC*, *BD* × *SZ*, and *SZ* × *HC*) for each model, for each of the two scenes. We trained each model on the full set of features described previously (§4.2).

We report our results for Scene 1 and Scene 2 in Table 3. We observe that the consistently highest-performing model across both scenes is the random forest classifier, achieving strong accuracies ranging from 0.93 (BD × HC) to 0.96 (SZ versus either) in Scene 1 and 0.70 (HC × SZ) to 0.96 (BD × HC) in Scene 2. Greater variation among top-performing classifiers was observed when comparing F_1, with the random forest classifier still achieving the highest performance most of the time. Interestingly, classification appeared to be more challenging when discriminating between HC and SZ in Scene 2 transcripts. Nonetheless, the overall strong classification performance across the board for Scenes 1 and 2 using feature-based classification models suggests high learning validity for both the dataset and the features extracted.

	BD × *HC*		*BD* × *SZ*		*HC* × *SZ*	
Mo.	**A**	$\mathbf{F_1}$	**A**	$\mathbf{F_1}$	**A**	$\mathbf{F_1}$
BB	0.42	0.52	0.33	0.5	0.33	0.5
MB	0.37	0.62	0.38	0.42	0.38	0.49
MR	0.48	0.63	0.62	0.60	0.60	0.27

Table 4: Performance comparisons between Transformers on Scene 2 transcripts. BB refers to *BERT base*, MB is *MentalBERT*, and MR is *MentalRoBERTa*.

6.2 Transformer-based Models

Applying pretrained Transformers to domain-specific tasks may produce more robust, dependable, and accurate models (Alsentzer et al., 2019). Since much recent success in NLP has been achieved using Transformer models, we also experiment with several using the same one-versus-one classification setting and data splits from our other experiments. We compare the performance of pretrained BERT base (Devlin et al., 2018), MentalBERT (Ji et al., 2022), and Mental-RoBERTa (Ji et al., 2022) models for our task. BERT base is a pretrained English model using a masked language modeling objective. It randomly masks a small percentage of words and learns to predict the masked samples. The model was trained for one million steps in batch sizes of 256 with fine-tuned hyperparameters set to: *optimizer*=Adam, *learning rate*=1e-4, β_1=0.9, β_2=0.999, and *decay*=0.01. MentalBERT and Mental-RoBERTa follow the same architecture but use dynamic masking and domain adaptive pretraining. The pretraining corpus includes depression, stress, and suicidal ideation data from Reddit. We passed subject utterances from our transcripts directly to these models for automated encoding of implicitly learned features.

We present the results for a sample of these experiments (Scene 2 classifications of HC vs. SZ subjects) in Table 4. We observe much lower performances than seen with feature-based classifiers. There may be many reasons for this, ranging from characteristics of the data used for pretraining to inefficiencies in implicitly learned features relative to features engineered based on known psycholinguistic attributes. Since we do not observe promising results using pre-trained Transformer models and these models also do not lend themselves as easily as tools to facilitate linguistic analyses, we leave further probing of this to future work.

7 Conclusion

Publishing language data collected in clinical settings that is paired with validated psychiatric diagnoses is an essential first step towards realizing more realistic, medically relevant NLP applications in the mental health domain. In this work, we take that step and describe our new corpus developed in close consultation between NLP and psychiatric researchers and clinicians. The corpus includes manually transcribed interactions between clinical interviewers and healthy control subjects or those with diagnosed schizophrenia and bipolar disorder. We describe all data collection procedures, extract a wide range of promising linguistic features from the data, and conduct an extensive first set of analyses to document trends in linguistic behavior among the SZ, BD, and HC subject groups. We show that linguistic diversity manifests itself in various ways across subject populations.

We hope that our work will diversify NLP research in the mental health domain beyond social media settings, and that it will open the door for more clinically valid studies of language behavior associated with diagnosed psychiatric conditions. All features extracted for this work are freely available on GitHub and can be downloaded without any further permission.[7] The de-identified transcripts can be downloaded from the National Institute of Mental Health data archive, in keeping with the terms of our NIH reporting requirements and the corresponding research grant that funded this work.[8] In the future, we plan to extend our study to also investigate spoken language and acoustic properties from the collected audiorecordings.

Limitations

This work is limited by a few factors. First, although our dataset is large by psychiatric standards, its size is still limited compared to datasets used for many other modern NLP tasks. This prevents us from being able to productively use complex models that have achieved state-of-the-art performance in other tasks, as documented in §6.2 with our experiments using fine-tuned versions of BERT, MentalBERT, and Mental-RoBERTa. We note that a disadvantage of deep learning models is that they are less interpretable than feature-based counterparts; thus, since classifier performance is not a central goal of our work, the poor performance observed with pre-trained Transformers is not a crucial shortcoming. Our primary interest in the classification experiments described in Section 6 was to establish learning validity for our dataset.

Second, although we explore a wide range of temporal, sentiment, psycholinguistic, emotion, and lexical diversity features in our experiments, our feature set does not comprehensively or conclusively cover all linguistic traits that may be of interest when analyzing the language behaviors of our target subject groups. Thus, our claims are limited by the boundaries of the conditions tested in our experiments—it may be that the most informative linguistic features are as yet undiscovered. We hope that this is indeed the case, and that future work develops new innovations that expand upon our findings.

Finally, our dataset is restricted to English conversations. The extent to which this research generalizes to other languages, including those vastly different from or substantially less-resourced than English, is unknown for now. The collection of complementary data in other languages, and especially those with different morphological typology, is a promising direction for future work.

Ethical Considerations

Several important ethical questions arise when working with data collected from human participants generally, and data dealing with mental health concerns specifically. We consider both questions here. We also point readers to our datasheet and other details regarding fair and inappropriate uses of our data in Appendix C.

Dataset Creation

In collecting this data, we followed all codes of ethics laid out by the Association for Computational Linguistics, the United States of America's National Institutes of Health, and the U.S. National Institute of Mental Health. All universities, laboratories, hospitals, and research centers involved in this project have secured ethics approval from their Institutional Review Boards before working with any data. Data was collected from outpatients recruited through studies supported by the National Institute of Mental Health. Inclusion criteria were ability to provide informed written consent, diagnosis of either bipolar disorder or schizophrenia/schizoaffective disorder according to the Diagnostic and Statistical Manual of the American Psychiatry Association, and outpatient status at the time of assessment. Informed written consent was taken from all participants for audiorecording and de-identified research data sharing.

Audiorecordings were professionally transcribed

[7] https://github.com/ankitaich09/EMNLP_2022

[8] https://nda.nih.gov

by a trusted third-party company. Any identifiable data was manually removed from the transcripts at the time of transcription, and transcripts were verified to be de-identified by members of the study team. No data that might point toward the identity of any person(s) was used in any way in this work, including for feature creation, modeling, or analysis, nor will it be shared at any time. Collected audiorecordings are stored securely and are not part of the data release (and are also inaccessible to some members of the study team).

De-identified transcripts are shared in full compliance with all governing bodies involved, through the National Institute of Mental Health's data archive following federally mandated grant reporting and data sharing requirements. All parties interested in accessing the data will be required to complete the NIMH Data Archive Data Use Certification, which outlines terms and conditions for data use, collaboration with shared data, compliance with human subjects and institutional research requirements, and other information.[9] The data use certification is non-transferable and recipients are not allowed to distribute, sell, or move data to other individuals, entities, or third-party systems unless they are authorized under a similar data use certification for the same permission group. The released transcripts include timestamps and de-identified utterances. Feature files (containing only the numeric feature vectors generated for each transcript using the procedures described in §4.2) are also available on GitHub at the link provided in this paper.

Intended Use

The intended use for this dataset is to enable discovery and analysis of the linguistic characteristics and language behaviors associated with members of three subject groups: people with schizophrenia, people with bipolar disorder, and healthy controls. Although we provide results from proof-of-concept experiments to classify transcripts into subject groups, these are intended merely to demonstrate evidence of data validity and learnability, and the experimental inferences are provided to showcase linguistic differences between groups. This in turn establishes feasibility of the dataset as a language analysis resource for the target populations. We *do not* condone use of this dataset to develop models to automatically diagnose individuals with mental health conditions, especially in the absence of feedback from trained professionals and psychiatric experts.

When used as intended and when functioning correctly, we anticipate that models developed and analyses performed using this dataset may be used to facilitate discovery of novel linguistic biomarkers of schizophrenia or bipolar disorder. This information could be used to support mental health research. When used as intended but giving incorrect results, researchers may place undue importance on irrelevant linguistic biomarkers. Since this dataset is not intended for diagnostic purposes, this is unlikely to lead to real-world harm, although it may slow the progress of some psychiatric research as researchers attempt to replicate and verify results.

Potential harms from misuse of the technology include the development of models to predict mental health status, and subsequent misprediction of serious mental health conditions. We reiterate that this dataset is not intended for diagnostic use, and that individuals seeking mental health care should always consult trained professionals. The National Institute of Mental Health's data archive includes a mechanism for logging research studies associated with the shared dataset. We will monitor this log and contact researchers who attempt to use the data for purposes outside its intended use.

Acknowledgements

We thank the anonymous reviewers for their insightful feedback. Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under award number R01MH116902. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2932–2949, Online. Association for Computational Linguistics.

Ankit Aich and Natalie Parde. 2022. Are you really okay? a transfer learning-based approach for identification of underlying mental illnesses. In *Proceedings of the Eighth Workshop on Computational Linguistics*

[9] https://nda.nih.gov/ndapublicweb/Documents/NDA+Data+Access+Request+DUC+FINAL.pdf

and Clinical Psychology, Seattle, WA. Association for Computational Linguistics.

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.

Silvio Amir, Mark Dredze, and John W. Ayers. 2019. Mental health surveillance over social media with digital cohorts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 114–120, Minneapolis, Minnesota. Association for Computational Linguistics.

Alina Arseniev-Koehler, Sharon Mozgai, and Stefan Scherer. 2018. What type of happiness are you looking for? - a closer look at detecting mental health from language. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 1–12, New Orleans, LA. Association for Computational Linguistics.

Sujeevan Aseervatham, Anestis Antoniadis, Eric Gaussier, Michel Burlet, and Yves Denneulin. 2011. A sparse version of the ridge logistic regression for large-scale text categorization. *Pattern Recognition Letters*, 32:101–106.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Valentina Bambini, Giorgio Arcara, Margherita Bechi, Mariachiara Buonocore, Roberto Cavallaro, and Marta Bosia. 2016. The communicative impairment as a core feature of schizophrenia: Frequency of pragmatic deficit, cognitive substrates, and relation with quality of life. *Comprehensive Psychiatry*, 71:106–120.

Kfir Bar, Vered Zilberstein, Ido Ziv, Heli Baram, Nachum Dershowitz, Samuel Itzikowitz, and Eiran Vadim Harel. 2019. Semantic characteristics of schizophrenic speech. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 84–93, Minneapolis, Minnesota. Association for Computational Linguistics.

Dennis Becker, Ward van Breda, Burkhardt Funk, Mark Hoogendoorn, Jeroen Ruwaard, and Heleen Riper. 2018. Predictive modeling in e-mental health: A common language framework. *Internet Interventions*, 12:57–67.

Gillinder Bedi, Facundo Carrillo, Guillermo A Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B Mota, Sidarta Ribeiro, Daniel C Javitt, Mauro Copelli, and Cheryl M Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1(1):15030.

Michael Birnbaum, Sindhu Kiranmai Ernala, Asra Rizvi, Munmun Choudhury, and John Kane. 2017. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of Medical Internet Research*, 19:e289.

Michael Birnbaum, Asra Rizvi, Christoph Correll, and John Kane. 2015. Role of social media and the internet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders. *Early intervention in psychiatry*, 11.

Ameni Bouaziz, Christel Dartigues-Pallez, Célia da Costa Pereira, Frédéric Precioso, and Patrick Lloret. 2014. Short text classification using semantic random forest. In *Data Warehousing and Knowledge Discovery*, pages 288–299, Cham. Springer International Publishing.

Ryan Boyd, Ashwini Ashokkumar, Sarah Seraj, and James Pennebaker. 2022. The development and psychometric properties of liwc-22.

Sandra Bucci, Matthias Schwannauer, and Natalie Berry. 2019. The digital revolution and its impact on mental health care. *Psychology and Psychotherapy: Theory, Research and Practice*, 92.

J. B Carol. 1964. Language and thought. *Francais Moderne*, 46:25–32.

Victor M. Castro, Jessica Minnier, Shawn N. Murphy, Isaac Kohane, Susanne E. Churchill, Vivian Gainer, Tianxi Cai, Alison G. Hoffnagle, Yael Dai, Stefanie Block, Sydney R. Weill, Mireya Nadal-Vicens, Alisha R. Pollastri, J. Niels Rosenquist, Sergey Goryachev, Dost Ongur, Pamela Sklar, Roy H. Perlis, Jordan W. Smoller, , Jordan W. Smoller, Roy H. Perlis, Phil Hyoun Lee, Victor M. Castro, Alison G. Hoffnagle, Pamela Sklar, Eli A. Stahl, Shaun M. Purcell, Douglas M. Ruderfer, Alexander W. Charney, Panos Roussos, Carlos Pato, Michele Pato, Helen Medeiros, Janet Sobel, Nick Craddock, Ian Jones, Liz Forty, Arianna DiFlorio, Elaine Green, Lisa Jones, Katherine Dunjewski, Mikael Landén, Christina Hultman, Anders Juréus, Sarah Bergen, Oscar Svantesson, Steven McCarroll, Jennifer Moran, Jordan W. Smoller, Kimberly Chambert, and Richard A. Belliveau. 2015. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *American Journal of Psychiatry*, 172(4):363–372. PMID: 25827034.

John W Chotlos. 1944. A statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 56:75–111.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Cheryl Mary Corcoran and Guillermo A. Cecchi. 2020. Using language processing and speech analysis for the identification of psychosis and other disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 5(8):770–779. Understanding the Nature and Treatment of Psychopathology: Letting the Data Guide the Way.

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.

Firdaus Dhabhar. 2018. The short-term stress response – mother nature's mechanism for enhancing protection and performance under conditions of threat, challenge, and opportunity. *Frontiers in Neuroendocrinology*, 49.

Daniel Dugast. 1978. On what is the notion of theoreticalextent of the vocabulary based. *Frenchçais (Le) Moderne Paris*, 46(1):25–32.

Clyde M. Elmore and Donald R. Gorham. 1957. Measuring the impairment of the abstracting function with the proverbs test. *Journal of Clinical Psychology*, 13(3):263–266.

Brita Elvevåg, Peter W. Foltz, Daniel R. Weinberger, and Terry E. Goldberg. 2007. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93(1):304–316.

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–16, New York, NY, USA. Association for Computing Machinery.

Christian Fuchs. 2015. *Culture and Economy in the Age of Social Media*. Taylor and Francis Group.

Alexander Genkin, David Lewis, and David Madigan. 2007. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49.

Kris Gowen, Matthew Deschaine, Darcy Gruttadara, and Dana Markey. 2012. Young adults with mental health conditions and social networking websites: Seeking tools to build community. *Psychiatric rehabilitation journal*, 35:245–50.

Melissa Graham, Elizabeth Avery, and Sejin Park. 2015. The role of social media in local government crisis communications. *Public Relations Review*, 41.

E. Darío Gutiérrez, Guillermo Cecchi, Cheryl Corcoran, and Philip Corlett. 2017. Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2923–2930, Copenhagen, Denmark. Association for Computational Linguistics.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. On the state of social media data for mental health research.

Daisy Harvey, Fiona Lobban, Paul Rayson, Aaron Warner, and Steven Jones. 2022. Natural language processing methods and bipolar disorder: Scoping review. *JMIR Ment Health*, 9(4):e35928.

Jinrong He, Lixin Ding, Lei Jiang, and Ling Ma. 2014. Kernel ridge regression classification. *Proceedings of the International Joint Conference on Neural Networks*, pages 2263–2267.

Herdan. 1960. Quantitative linguistics. *London, Butterworth.*

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of Twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, LA. Association for Computational Linguistics.

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

Thorsten Joachims. 2002. *Learning to classify text using support vector machines*, volume 668. Springer Science & Business Media.

Rajshree Jodha, BC Gaur Sanjay, KR Chowdhary, and Amit Mishra. 2018. Text classification using knn with different features selection methods. *Text Classification using KNN with different Features Selection Methods*, 8(1):8–8.

Timothy Jurka. 2012. maxent: An r package for low-memory multinomial logistic regression with support for semi-automated text classification. *The R Journal*, 4.

Timothy P Jurka, Loren Collingwood, Amber E Boydstun, Emiliano Grossman, et al. 2013. Rtexttools: A supervised learning package for text classification. *RJournal*, 5(1):6–12.

Jan Kalbitzer, Thomas Mell, Felix Bermpohl, Michael Rapp, and Andreas Heinz. 2014. Twitter psychosis a rare variation or a distinct syndrome? *The Journal of nervous and mental disease*, 202:623.

Maria Kapantzoglou, Gerasimos Fergadiotis, and Alejandra Auza. 2019. Psychometric evaluation of lexical diversity indices in spanish narrative samples from children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research*, 62:1–14.

J.S. Kasanin, editor. 1944. *Language and thought in schizophrenia*. University of California Press, Berkeley, CA, US. ID: 1944-01428-000.

Prasadith Kirinde Gamaarachchige and Diana Inkpen. 2019. Multi-task, multi-channel, multi-input learning for mental illness detection using social media text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 54–64, Hong Kong. Association for Computational Linguistics.

Nithin Krishna, Bernard Fischer, Moshe Miller, Kelly Register-Brown, Kathleen Patchan, and Ann Hackman. 2012. The role of social media networks in psychotic disorders: A case report. *General hospital psychiatry*, 35.

Soon Li Lee, Miriam Park, and Cai Lian Tam. 2015. The relationship between facebook attachment and obsessive-compulsive disorder severity. In *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, volume 9.

Wee Sun Lee and Bing Liu. 2003. Learning with positive and unlabeled examples using weighted logistic regression. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 448–455. AAAI Press.

Feea R. Leifker, Thomas L. Patterson, Christopher R. Bowie, Brent T. Mausbach, and Philip D. Harvey. 2010. Psychometric properties of performance-based measurements of functional capacity: Test–retest reliability, practice effects, and potential sensitivity to change. *Schizophrenia Research*, 119(1):246–252.

Liu Lin, Jaime Sidani, Ariel Shensa, Ana Radovic, Elizabeth Miller, Jason Colditz, Beth Hoffman, Leila Giles, and Brian Primack. 2016. Association between social media use and depression among u.s. young adults. *Depression and anxiety*, 33.

Christopher A. Lovejoy. 2019. Technology and mental health: The role of artificial intelligence. *European Psychiatry*, 55:1–3.

Annalise Mabe, Jean Forney, and Pamela Keel. 2014. Do you "like" my photo? facebook use maintains eating disorder risk. *International Journal of Eating Disorders*, 47.

Heinz-Dieter Mass. 1972. Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 2(8):73.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019a. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H. Andrew Schwartz. 2019b. Suicide risk assessment with multi-level dual-context language and BERT. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44, Minneapolis, Minnesota. Association for Computational Linguistics.

Michelle L. Miller, Martin T. Strassnig, Evelin Bromet, Colin A. Depp, Katherine Jonas, Wenxuan Lin, Raeanne C. Moore, Thomas L. Patterson, David L. Penn, Amy E. Pinkham, Roman A. Kotov, and Philip D. Harvey. 2021. Performance-based assessment of social skills in a large sample of participants with schizophrenia, bipolar disorder and healthy controls: Correlates of social competence and social appropriateness. *Schizophrenia Research*, 236:80–86.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Michelle Morales, Stefan Scherer, and Rivka Levitan. 2018. A linguistically-informed fusion approach for multimodal depression detection. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 13–24, New Orleans, LA. Association for Computational Linguistics.

Natalia B. Mota, Nivaldo A. P. Vasconcelos, Nathalia Lemos, Ana C. Pieretti, Osame Kinouchi, Guillermo A. Cecchi, Mauro Copelli, and Sidarta Ribeiro. 2012. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLOS ONE*, 7(4):1–9.

Denis Newman-Griffis, Jill Fain Lehman, Carolyn Rosé, and Harry Hochheiser. 2021. Translational NLP: A new paradigm and general principles for natural language processing research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4125–4138, Online. Association for Computational Linguistics.

Uri Nitzan, Efrat Shoshan, Shaul Lev-Ran, and Shmuel Fennig. 2011. Internet-related psychosis - a sign of the times? *The Israel journal of psychiatry and related sciences*, 48:207–11.

Igor Pantic, Aleksandar Damjanovic, Jovana Todorovic, Dubravka Topalovic, Dragana Bojovic Jovic, Sinisa Ristic, and Senka Pantic. 2012. Association between online social networking and depression in high school students: Behavioral physiology viewpoint. *Psychiatria Danubina*, 24:90–3.

Thomas L Patterson, Sherry Moscona, Christine L McKibbin, Kevin Davidson, and Dilip V Jeste. 2001. Social skills performance assessment among older patients with schizophrenia. *Schizophrenia Research*, 48(2):351–360.

C. Perlini, A. Marini, M. Garzitto, M. Isola, S. Cerruti, V. Marinelli, G. Rambaldelli, A. Ferro, L. Tomelleri, N. Dusi, M. Bellani, M. Tansella, F. Fabbro, and P. Brambilla. 2012. Linguistic production and syntactic comprehension in schizophrenia and bipolar disorder. *Acta Psychiatrica Scandinavica*, 126(5):363–376.

Andrew Perrin. 2015. Social media usage. *Pew research center*, pages 52–68.

GUIRAUD Pierre. 1959. *Problegravemes et meacute methodes de la statistique linguistique*. Payol.

Tomas Pranckevičius and Virginijus Marcinkevičius. 2017. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2):221.

Randall Ratana, Hamid Sharifzadeh, Jamuna Krishnan, and Shaoning Pang. 2019. A comprehensive review of computational methods for automatic prediction of schizophrenia with insight into indigenous populations. *Frontiers in Psychiatry*, 10.

Larry Rosen, Kelly Whaling, S Rab, Mark Carrier, and Nancy Cheever. 2013. Is facebook creating "idisorders"? the link between clinical symptoms of psychiatric disorders and technology use, attitudes and anxiety. *Computers in Human Behavior*, 29:1243–1254.

Ramin Safa, Peyman Bayat, and Leila Moghtader. 2022. Automatic detection of depression symptoms in twitter using multimodal analysis. *The Journal of Supercomputing*, 78.

Ivan Sekulic and Michael Strube. 2019. Adapting deep learning methods for mental health prediction on social media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 322–327, Hong Kong, China. Association for Computational Linguistics.

Nabia Shahreen, Mahfuze Subhani, and Md Mahfuzur Rahman. 2018. Suicidal trend analysis of twitter using machine learning and neural network. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–5.

Ravinder Singh, Jiahua Du, Yanchun Zhang, Hua Wang, Yuan Miao, Omid Sianaki, and Anwaar Ulhaq. 2020. *A Framework for Early Detection of Antisocial Behavior on Twitter Using Natural Language Processing*, pages 484–495. Springer Link.

April Smith, Jennifer Hames, and Thomas Joiner. 2013. Status update: Maladaptive facebook usage predicts increases in body dissatisfaction and bulimic symptoms. In *Journal of affective disorders*, volume 149.

HH Somers. 1966. Statistical methods in literary analysis. *The computer and literary style*, 128:140.

Michael M. Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Mildred Templin. 1957. *Certain Language Skills in Children: Their Development and Interrelationships*. University of Minnesota Press.

Alina Trifan and José Luís Oliveira. 2019. Bioinfo@uavr at erisk 2019: delving into social media texts for the early detection of mental and food disorders. In *CLEF*.

Bruno Trstenjak, Sasa Mikac, and Dzenana Donko. 2014. Knn with tf-idf based framework for text categorization. *Procedia Engineering*, 69:1356 – 1364. 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013.

Elsbeth Turcan and Kathy McKeown. 2019a. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong. Association for Computational Linguistics.

Elsbeth Turcan and Kathy McKeown. 2019b. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong. Association for Computational Linguistics.

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.

Yufei Wang, Stephen Wan, and Cécile Paris. 2016a. The role of features and context on suicide ideation detection. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 94–102, Melbourne, Australia.

Yufei Wang, Stephen Wan, and Cécile Paris. 2016b. The role of features and context on suicide ideation detection. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 94–102, Melbourne, Australia.

Genta Indra Winata, Onno Pepijn Kampman, and Pascale Fung. 2018. Attention-based lstm for psychological stress detection from spoken language using distant supervision. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6204–6208.

Baoxun Xu, Xiufeng Guo, Yunming Ye, and Jiefeng Cheng. 2012. An improved random forest classifier for text categorization. *Journal of Computers*, 7.

Hao Yan, Ellen Fitzsimmons-Craft, Micah Goodman, Melissa Krauss, Sanmay Das, and Patty Cavazos-Rehg. 2019. Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention. *International Journal of Eating Disorders*, 52.

Yiming Yang. 2001. A study on thresholding strategies for text categorization. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*.

Zhou Yong, Li Youwen, and Xia Shixiong. 2009. An improved knn text classification algorithm based on clustering. *Journal of Computers*, 4.

Jianlong Zhou, Hamad Zogan, Shuiqiao Yang, Shoaib Jameel, Guandong Xu, and Fang Chen. 2021. Detecting community depression dynamics due to covid-19 pandemic in australia. *IEEE Transactions on Computational Social Systems*, PP:1–10.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019a. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019b. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Zomick, Sarah Ita Levitan, and Mark Serper. 2019. Linguistic analysis of schizophrenia in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 74–83, Minneapolis, Minnesota. Association for Computational Linguistics.

A Appendix A: Sample Transcripts

A.1 Scene 1: Introducing Yourself

***Examiner**: Can you tell me, are the residents in this building friendly?*

Participant: I don't really know because I keep to myself. I don't really socialize with other residents to find out what they're really like. Everyone is really nice, definitely knock on their door to see what they're doing or not. Introduce yourself and find out you know what their place is like, or you know, who they live with, all that stuff - kind of what goes on in your apartment.

***Examiner**: I see.*

Participant: Their apartment, not your apartment. Um, if you have a car, you can park in the resident parking. It talks about having maintenance having stuff done at your place and all that.

A.2 Scene 2: Confronting Your Landlord

Participant: - Do they have a key to my place to unlock it? Or do I need to be there in the apartment for them to get inside and look at the leak? Or do I need a key? Or do they need a key? Not me. Do they need me physically there in the apartment to see the leak? Or, two, do they need a key from me to get inside the apartment to do the leak, if that case I need to get on my errands by then.

***Examiner**: Um, so I have a list, and you're on the list. But there are other problems that are more serious.*

Participant: Okay, but this leak is getting worse, and I would like for you to try and get back to me in the next possible days to let me know what's going on with the leak. Or I might have to threaten to move out because this is unright and you are not being justice with this. And, um, I think it's unfair that you're putting other people that are higher ahead and their problems ahead of mine. I think if I'm paying your rent and your deposit, and if I had a pet or whatever and I paid the deposit for that too.

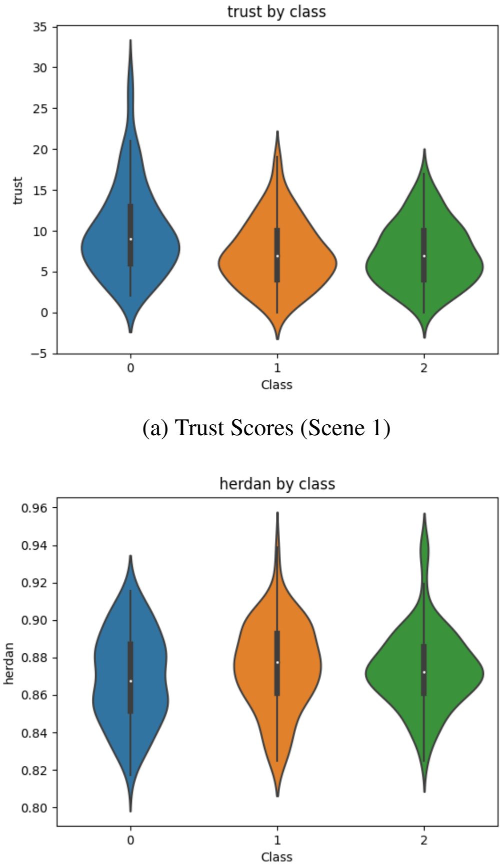
B Appendix B: Extended Visualization

Figures 5 and 6 visualize the feature distributions that complement those provided in the main paper (Figures 2 and 3). The figures provided in the main paper correspond to Scene 2 from our dataset, whereas the figures from this section correspond to Scene 1.

C Appendix C: Datasheet and Fair and Inappropriate Usage

C.1 Data Collection and Creation

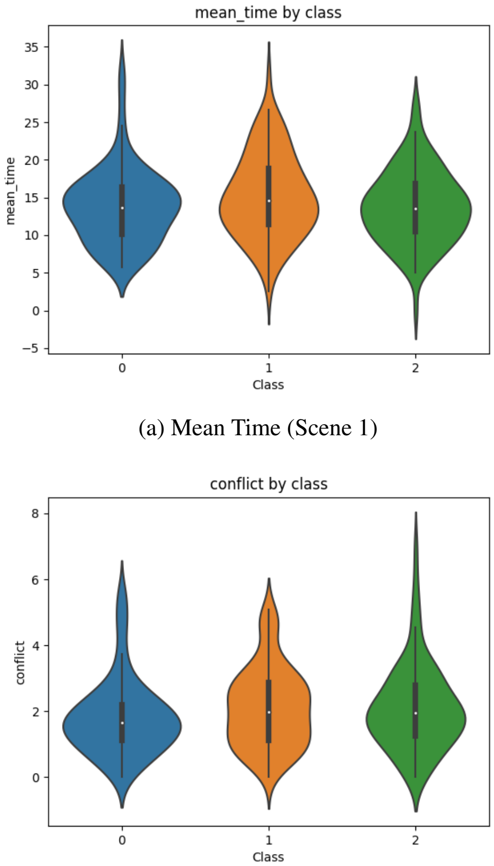
Data in the form of audiorecordings was collected at the University of California San Diego, the Uni-



(a) Trust Scores (Scene 1)

(b) Lexical Diversity Scores (Scene 1)

Figure 5: Blue represents healthy controls, orange represents schizophrenia, and green represents bipolar. Figure is best viewed in color. Figure shows violin plots with quartiles, medians, and interquartile ranges across classes *Healthy*, *Schizophrenic*, and *Bipolar*.



(a) Mean Time (Scene 1)

(b) Interpersonal Conflict (Scene 1)

Figure 6: Blue represents healthy controls, orange represents schizophrenia, and green represents bipolar. Figure is best viewed in color. Figure shows violin plots with quartiles, medians, and interquartile ranges across classes *Healthy*, *Schizophrenic*, and *Bipolar*.

versity of Miami, and the University of Texas at Dallas. Audiorecordings were then sent to a professional third-party service for transcription. Deidentification was performed by the transcription service, and verified on site by the study teams. The de-identified data was processed by the study team at the University of Illinois Chicago.

Participants provided written informed consent. No identifying information such as name or birth date was collected. Demographic information such as biological sex and race were collected to help in future studies, but this information is not released publicly and will not be shared with others. Descriptive statistics of the participant demographics are provided in Section 3.3.

C.2 Intended Audience

The intended audience for this dataset includes psychiatric and computer science researchers, and others interested in understanding language patterns common in people with diagnosed mental health concerns. The intended use for this data is to enable discovery and analysis of the linguistic characteristics and language behaviors associated with people with schizophrenia, people with bipolar disorder, and healthy controls. We do not intend for this dataset to be used for automated diagnostic purposes, and we do not encourage others to attempt to replace psychological or psychiatric treatment with classification or deep learning methods.

C.3 Validity of Diagnoses

Recruited subjects were clinically diagnosed as having a DSM-IV diagnosis of schizophrenia or schizoaffective disorder, and being medicated for the same. Subjects with bipolar disorder met the conditions defined in the APA's DSM-5. Healthy controls did not have a clinical diagnosis for either disorder.

The data was collected at the University of California San Diego, the University of Miami, and the University of Texas at Dallas under clinical supervision with medical experts on scene. All labels are clinically valid. Changing them for any reason after acquiring the data is a violation of ethical code.

C.4 Fair Uses

Fair usage of this dataset includes performing data analyses and developing methods to understand emotions, speech variations, feature validity, and language differences among people with schizophrenia or bipolar disorder, or healthy subjects. Data was collected under controlled experimental settings and underwent rigorous deidentification processes. By using this data you agree to participate only in experiments that do not undermine the validity of clinical diagnoses provided by the original labels. Visualization of data patterns, user distribution, language changes, and emotional changes across populations are all fair uses of the data.

C.5 Interpreting the Paper

The paper introduces a novel, clinically valid dataset that enables the study of language in the context of diagnosed mental health conditions. In addition to describing the data, we provide a detailed analysis of temporal, sentiment, psycholinguistc, emotion, and lexical diversity features extracted from the data. We also show visual aids to facilitate understanding of this analysis. We provide classification results for a task designed to categorize transcripts into groups only to offer evidence of the dataset's validity for automated analysis problems. We do not intend to suggest that a machine learning model can accurately predict an individual's mental health from 3-4 minutes of transcribed conversation.

C.6 Inappropriate Uses

The data can only be downloaded directly from the National Institute of Mental Health's data archive. Privately distributing the data is an inappropriate usage. Any attempt to try to identify the subjects is also an inappropriate use. Other inappropriate uses may include but are not limited to:

- Augmenting the data for machine learning or deep learning purposes
- Annotating (or re-annotating) the data on your own
- Running speech classifiers to try to predict speaker identities
- Sharing the data with others on your own
- Stating that a person's mental health condition can be accurately predicted based on their speech transcript

We hope that this data will diversify NLP research in the mental health domain and open new opportunities for interdisciplinary research. We remind all readers and users of this dataset to respect the fairness and ethical codes laid out by the National Institutes of Health, the Association for Computational Linguistics, and the National Institute for Mental Health.