Towards Comprehensive Language Analysis for Clinically Enriched Spontaneous Dialogue

Baris Karacan[1], **Ankit Aich**[1, 5], **Avery Quynh**[2], **Amy Pinkham**[3], **Philip Harvey**[4], **Colin Depp**[2], and **Natalie Parde**[1]

[1]Department of Computer Science, University of Illinois Chicago
{bkarac3, aaich2, parde}@uic.edu

[2]Department of Psychiatry, University of California San Diego
{akquynh, 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

[5]National Institutes of Health, National Institute on Drug Abuse
ankit.aich@nih.gov

ABSTRACT

Contemporary NLP has rapidly progressed from feature-based classification to fine-tuning and prompt-based techniques leveraging large language models. Many of these techniques remain understudied in real-world, clinically enriched spontaneous dialogue. We fill this gap by systematically testing the efficacy and performance of varied NLP techniques on transcribed speech collected from patients with bipolar disorder, schizophrenia, and healthy controls taking a focused, clinically-validated language test. We observe impressive utility of feature-based and language modeling techniques, finding that these approaches may provide a plethora of information capable of upholding clinical truths about these subjects. Building upon this, we establish pathways for future research directions in automated detection and understanding of psychiatric conditions.

1. Introduction

The use of NLP to support mental healthcare has gained prominence recently with research focusing on a variety of conditions including schizophrenia (Kalbitzer et al., 2014; Krishna et al., 2012), mood disorders (Lin et al., 2016; Pantic et al., 2012), personality disorders (Rosen et al., 2013), eating disorders (Mabe et al., 2014) and others (Turcan and McKeown, 2019a; Tadesse et al., 2019; Zirikly et al., 2019a; Morales et al., 2018). The promising capabilities of these approaches have been demonstrated using a range of techniques—for instance, Singhal et al. (2022) showed that LLMs with sizes up to 540B parameters can encode clinical and medical knowledge. Most NLP work towards mental health support has thus far focused on social media (e.g., Twitter (Kang et al., 2016) or Reddit (Yan et al., 2019)) rather than clinically-enriched data, extracting and annotating user data based on social features deemed relevant by technical researchers but not necessarily validated by clinicians. Consequently, the question of whether these reported NLP techniques hold relevance for clinical data still requires thorough investigation. In this paper, we comprehensively and empirically investigate this question using a clinically enriched dataset drawn from actual patient dialogues and their performance on standardized clinical tests.

Our dataset includes patient-psychologist conversations between 644 participants categorized based on their status as healthy control (HC) participants or participants with schizophrenia (SZ) or bipolar disorder (BD). Each participant engaged in a focused test during which they conversed for approximately four minutes across two scenarios. A previous paper introducing this dataset (Aich et al., 2022) established a feature-based performance benchmark, but did not provide details about feature importance or participant demographic trends, nor did it apply more contemporary LLM-based models to the task. Our contributions are:

- We systematically and comprehensively analyze these features and their significance across age and gender demographics.
- We assess feature-based statistical significance and importance when differentiating be-

LREC-COLING 2024, pages 16457–16472

tween feature groups.

- We use an encoder-based topic model to extract relevant topics from participant dialogue to visualize and confirm clinical observations.
- We show that large language model (LLM) settings can find patterns in subject dialogue.

Taken holistically, our work demonstrates how a range of more traditional and recent NLP methods can be leveraged to understand and work with clinically enriched spontaneous dialogue.

2. Data

We collected our data by audiorecording 644 participants as they took a standardized clinical test known as the social skills performance assessment (SSPA) (Patterson et al., 2001b). The SSPA is a conversational test designed to delineate social skills across multiple dimensions. Our participants were recruited based on their confirmed diagnoses of specific clinical conditions and medical histories, with the exception of those categorized as healthy controls (Aich et al., 2022). Participants were thus grouped into three categories: people with schizophrenia (referred to as SZ hereon; n=247), people with bipolar disorder (BD; n=286), and healthy controls (HC; n=110).

The SSPA has proven to be a useful and bias-free assessment and a strong predictor of social performance, and it has served as the basis of clinical rehabilitation-based work (Leifker et al., 2010; Miller et al., 2021). It includes two improvisational scenes, each of which involve a participant conversing with an interviewer (a trained psychologist). The scenes probe for specific but different information. The conversations were audiorecorded and later transcribed, and the transcribed dialogues were annotated by clinical professionals across different social skills dimensions corresponding to content relayed in the conversation.

The first scene seeks to facilitate a *friendly* interaction, to assess the social appropriateness with which the participants introduce themselves and engage in conversation. The participant is asked to imagine that they have just moved into a new neighborhood and must introduce themselves to a new neighbor. The second scene seeks to facilitate a *confrontational* interaction. Participants are given a focused, defined objective: They are told to imagine that they have a leaky pipe in their apartment which has not been fixed for a while, and they need to complain to their landlord and get it fixed. Dual annotation by clinical professionals across the SSPA skills dimensions achieved strong ($\kappa > 0.8$) inter-annotator agreement scores.

In our earlier work (Aich et al., 2022), we established dataset validity using a binary classification benchmark to discriminate between pairs of subject groups. The classifier was trained using 138 extracted linguistic features. We only ran experiments using half of the data (n=300 subjects), and since our focus was on data validation and establishing proof of concept, we did not study individual feature importance or significance. Here, we perform a more thorough set of experiments across the full SSPA dataset to deepen our insights regarding this task. We incorporate metadata pertaining to age and gender demographics, and experiment with LLM-based methods to demonstrate the ability of contemporary NLP approaches to reveal clinical patterns in a rich dataset.

Group A	**Group B**
Positive Score	Personal pronouns
Preposition	Authentic
Drive	Diction
Achieve	Linguistic
Cognition	Function
Cause	i-pron
Discrepancy	Friend
Non Fluency	Quantity
Filler Words	Certitude
All Punctuations	Sad Emotion
	Death
	Moral
	Adjectives

Table 1: Features showing statistically significant differences between age groups, indicating the group with higher feature values.

3. Analysis of Linguistic Differences Between Participants

We sought to discern and differentiate the linguistic patterns among participants across *Age*, *Gender*, and *Diagnoses*. To do so, we extracted our original 138 linguistic features (Aich et al., 2022) and studied their group-level differences. Briefly, these features include temporal, sentiment, psycholinguistic, lexical, and emotional characteristics. Examples of specific features within these groups include the recorded time taken to complete a task or utterance, the overall sentiment of a conversation, features derived from specific lexicons or tools such as the Linguistic Inquiry and Word Count (Boyd et al., 2022a, LIWC), and various measures of lexical diversity (Mass, 1972).

3.1. Age

For age-based analysis, we grouped participants into two categories depending on whether they were greater than or equal to 50 years old: Group

Female	Male
Numbers	Universal Quantifiers
Negative Tone	Insight
Positive Tone	Discrepancy
Swear Words	Focus Present
Time	Focus Future
Social Behavior	Netspeak
Conflict	Non Fluency
Exclamation	Punctuation
TTR	Maas Lexical Diversity
Summer	Female
Anticipation	Male
Conjunctions	
Herdan	
Dugast	

Table 2: Features showing statistically significant differences between gender groups, indicating the group with higher feature values.

A (age < 50) and Group B (age ≥ 50).[1] We computed standardized t-tests to find features with significantly different values between the two groups, allowing us to determine the association between linguistic traits and age and more fully understand how this demographic dimension may influence predictive models trained on the SSPA data. In Table 1, we highlight features that were found to have significantly different values between groups, indicating the group for which the feature value was higher. We observe that participants in Group A used more cognition-related words than those in Group B, aligning with Koch et al. (2022) which demonstrates a negative correlation between age and causation.

3.2. Gender

For gender-based analyses, we divided participants into two groups: male and female.[2] We tested feature correlation with both genders, and features with the highest correlation were compared using t-tests to assess group-level statistical significance. We show the dominant groups for statistically significant features in Table 2. We observe that males used more universal quantifiers (e.g., "all" or "nothing" words), present tense and future tense words, exclamations, and anticipatory speech. We observed that females had higher insight scores, discrepancy in speech, and sentiment and affect in speech. These findings support clinical observations such as that of Fast and Horvitz (2016), which shows that women verbalize more cognition and can more easily characterize non-dogmatic language.

3.3. Diagnoses

To analyze features across diagnostic groups (*BD*, *SZ* and *HC*), we first computed mean values for each feature, for each group, and then extracted features that exhibited statistically significant differences between their group-based means. We elaborate on this process below.

3.3.1. Determining Feature Significance

We extracted 138 linguistic attributes from the SSPA dataset, all of which were represented as normally distributed continuous values. Ross and Willson (2017) suggest that having a sample size greater than or equal to 30 decreases the chance of making a Type 2 error, and in our case each diagnostic group had > 30 samples. Each of the three groups was also independent of the others, thus satisfying all assumptions for the analysis of variance (ANOVA) test (Parab and Bhalerao, 2010). We perform one-way ANOVA tests to analyze the differences between groups of participants with different mental health diagnoses.

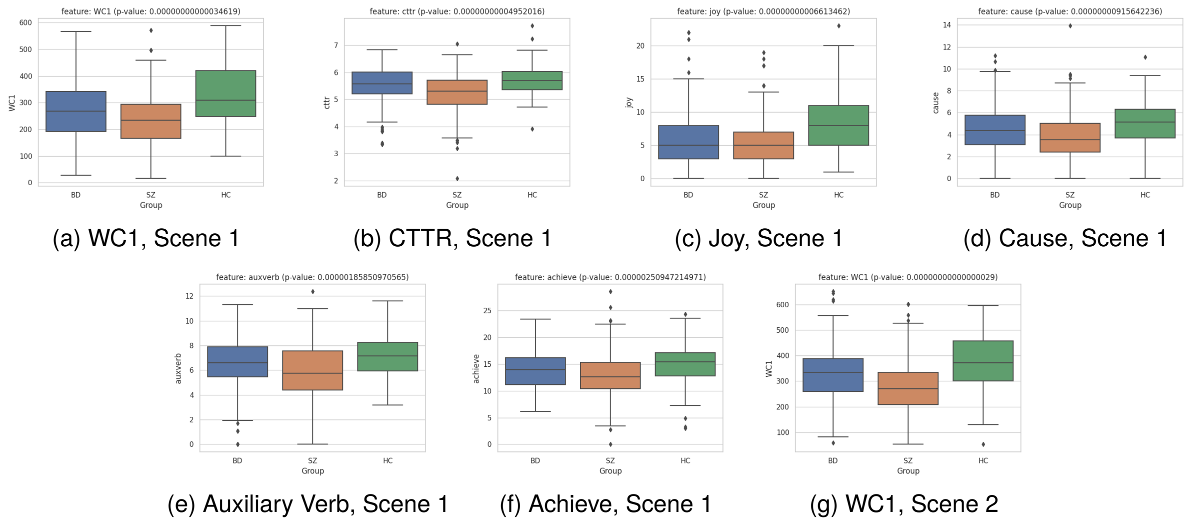
In ANOVA, a large F-value suggests that the group means are more spread out, indicating that at least one group might be significantly different from the others; conversely, with small F-values the data points within each group are more dispersed, making it harder to detect significant differences. The ANOVA test derives a p-value by comparing the resulting F-value with the F-distribution using the appropriate degrees of freedom (Bewick et al., 2004). If the p-value is less than a specified threshold (typically < 0.05), the null hypothesis is rejected, suggesting a statistically significant difference between at least one pair of the group means. We used the `statsmodels` module in Python to implement the ANOVA test.

3.3.2. Differentiating Diagnostic Groups

While one-way ANOVA identifies whether there are any significant differences among the group means, it does not specify which groups differ from each other. To pinpoint the groups with differing means, we conducted a post-hoc Tukey's HSD (Honestly Significant Difference) on the significant features since it shares the same assumptions as ANOVA. Tukey's HSD evaluates all possible

[1] This division reflects that used to originally define elderly SSPA participants (Patterson et al., 2001a).

[2] Participants may hold diverse gender identities. We asked participants to self-report their own gender, and all except one reported their gender as *Male* or *Female*. The remaining participant wrote "3," and we excluded this participant's data from analyses reported in this subsection since a single data point is an insufficient basis from which to draw conclusion.



(a) WC1, Scene 1 (b) CTTR, Scene 1 (c) Joy, Scene 1 (d) Cause, Scene 1

(e) Auxiliary Verb, Scene 1 (f) Achieve, Scene 1 (g) WC1, Scene 2

Figure 1: Distribution of features across diagnostic groups.

pairs of group means, determining which specific groups' means differ significantly (Glen, 2016).

3.3.3. Feature-Related Hypotheses

Several studies have investigated linguistic features for diagnosis of schizophrenia and bipolar disorder, giving rise to our hypotheses as follows:

- **H1:** Voleti et al. (2019) identified language features related to the SSPA that could successfully distinguish members of a clinical group that participated in the task (AUC=0.96), and also between subjects within the clinical group with SZ and BD (AUC=0.83). This motivates our hypothesis that further investigation into linguistic features could help uncover underlying characteristics of SZ and BD.
- **H2:** Park et al. (2018) examined lexical diversity of six active communities on Reddit. Three were related to mental health, including SZ and BD (r/depression, r/schizophrenia, r/bipolar), while the others were selected as controls and focused on unrelated topics (r/happy, r/loseit, r/bodybuilding). They found that members of r/bipolar and r/schizophrenia communities obtained poorer lexical diversity scores compared to the other communities, but they did not observe a significant difference between r/bipolar and r/schizophrenia. We hypothesized that lexical diversity of the BD and SZ groups could be better characterized and differentiated through an analysis of corrected type-token ratios (CTTR). CTTR offers a standardized measure of lexical diversity that is less influenced by the overall text length, achieved through the application of the square root of twice the number of tokens (Torruella and Capsada, 2013).
- **H3:** Chrobak et al. (2022) performed verbal fluency tests (VFT) on BD, SZ and HC groups, and analysis of Semantic VFT revealed that the SZ group showed lower word count than the HC group. Similarly, we hypothesized that the SZ group uses fewer words compared to other groups during both scenes.
- **H4:** Deng et al. (2018) observed that both the BD and SZ groups scored lower than HC in cognitive tests of verbal comprehension, executive functioning, and working memory, with SZ performing worst. Accordingly, we hypothesized that individuals in the BD and SZ groups would face challenges in causal reasoning, especially in Scene 2 when they confront the landlord.

Together, these studies motivate our feature comparison and highlight an advantage of using NLP for this purpose: although observational evidence may suggest an important conclusion, computationally extracting features representing this phenomenon and studying them at scale allows researchers to empirically ground these findings.

3.3.4. Results and Inferences

After conducting ANOVA and Tukey tests, we investigated feature significance across all pairs of the *BD*, *SZ* and *HC* groups to accept or reject our hypotheses. We also visualized the score distributions for each feature within each group, facilitating conclusions about language usage patterns within those groups. Table 3 describes features found to be significantly different across all groups, and Figures 1a-1g depict distributions of the corresponding features using box plots. Extra ANOVA details are in Table 4

Feature	Description	Category	Source	Scene
WC	Total word count	Psycholinguistic	Boyd et al. (2022b)	1, 2
cttr	Corrected type-token ratio	Lexical Diversity	Carroll (1964)	1
joy	Words associated with the emotion '*joy*'	Emotion	Mohammad and Turney (2013b)	1
cause	Causal words signifying a cause-and-effect relationship	Psycholinguistic	Boyd et al. (2022b)	1
auxverb	Number of auxiliary verbs	Psycholinguistic	Boyd et al. (2022b)	1
achieve	Words that reflect accomplishment	Psycholinguistic	Boyd et al. (2022b)	1

Table 3: Details of features that are significantly different across all the diagnostic groups.

Six features differentiated all three diagnostic pairs (BD vs. HC, BD vs. SZ and HC vs. SZ) for Scene 1 , and only one feature distinguished these pairs for Scene 2. When examining the feature score distributions, we observed that:

- *HC* exhibited a more diverse vocabulary in their speech compared to both *BD* and *SZ*. Among *BD* and *SZ*, *BD* demonstrated greater lexical richness.
- *HC* used a higher word count than both *BD* and *SZ*. When comparing *BD* and *SZ*, *BD* used more words.
- *HC* expressed higher levels of joy than the other groups. Between *BD* and *SZ*, *SZ* used less words associated with joy.
- *SZ* used fewer words related to causality and reasoning compared to both *BD* and *HC*. Within *BD* and *HC*, *HC* used causal words more frequently.
- *HC* used more auxiliary verbs than *BD*, and *BD* used more auxiliary verbs than *SZ*.
- *HC* displayed a higher sense of achievement in their language. Among *BD* and *SZ*, *SZ* used fewer words related to success and achievement.

Table 5 summarizes the diagnostic feature analyses across both scenes. Statistical tests identified a larger number of significant features in Scene 1 than Scene 2. Six Scene 1 features showed differences across all group pairs, whereas only one feature (WC.1) was significantly different across all pairs in Scene 2. Despite this, a similar number of features discerned either one and two significant group pairs in both scenes, and differentiating between *BD* and *HC* always proved more challenging than distinguishing *SZ*. Weiner et al. (2019) suggests that the link between mood states and linguistic capabilities in BD is intricate, with certain BD phases exhibiting linguistic traits akin to HC. Consequently, BD individuals not in acute mood episodes may retain similar linguistic functions to HC, leading to comparable language patterns. Notably, in Scene 2, the linguistic behavior of *SZ* deviated significantly from that of either *BD* or *HC* in over 90% of the significant features.

Feature	F	df	p
Achieve	13	2	0.000003
Auxverb	3.02	2	0.004
Anticipation	18.4	2	1.15e-08
All punctuation	6.18	2	0.0002
Big Words	3.95	2	0.001
Cause	19.0	2	9e-09
Drives	6.56	2	0.0001
Joy	24	2	6.1e-11
Max time	11.8	2	0.000009
Surprise	7.5	2	0.00002
Trust	16.4	2	1e-7
WPS	6.3	2	0.001

Table 4: More Features from ANOVA

4. Impact of Clinical Features

With the rise in use cases for LLMs, we also perform experiments showcasing their use with our clinically enriched data. We conduct topic modeling experiments using an encoder-only BERT model (Devlin et al., 2018), BERTopic (Grootendorst, 2022), and an encoder-decoder based Flan Unifying Language Learning (`flan-ul2`) 20B-parameter model (Tay et al., 2023).

4.1. Topic Modeling

We selected BERTopic as the backbone for our topic modeling experiments. BERTopic (Grootendorst, 2022) uses five independent sub-models to generate topic representations, giving the user flexibility to modify sub-models according to their requirements. To create topics from SSPA transcripts, we first separated the transcripts based on all possible group $\times$ scene combinations (HC

	Scene 1	Scene 2
Total features	144	144
Significant features (p<0.05)	**61**	52
Features that differentiate no pairs	2	2
Features that differentiate 1 pair	**28**	24
(BD vs. HC)	**2**	1
(BD vs. SZ)	15	15
(HC vs. SZ)	**11**	8
Features that differentiate 2 pairs	25	25
(BD vs. HC, BD vs. SZ)	**2**	1
(BD vs. SZ, HC vs. SZ)	11	**24**
(BD vs. HC, HC vs. SZ)	**12**	0
Features that differentiate all 3 pairs	**6**	1

Table 5: Comparative analysis of diagnostic features between scene 1 and scene 2.

Scene 1 (SC1), HC Scene 2 (SC2), BD SC1, BD SC2, SZ SC1, and SZ SC2). For each transcript, we extracted patient utterances using regular expressions and combined these utterances in lists to create patient dialogue subsets.

4.1.1. Building Topic Models

After constructing the subsets, we converted utterances to numerical representations (embeddings) using the SentenceTransformers (Reimers and Gurevych, 2019) framework as it is optimized for semantic similarity at the document (in our case, utterance) level. We specifically used the `all-mpnet-base-v2` model available on the HuggingFace model hub. Next, we performed dimensionality reduction on the document embeddings using UMAP (McInnes et al., 2018), a technique that retains both local and global features of the dataset while reducing its dimensions.

With our dimensionality-reduced document embeddings, we used HDBSCAN (McInnes et al., 2017) to cluster our data. We selected HDBSCAN based on its capability to detect clusters of varying shapes and outliers when applicable. In our case, this ensures that utterances from the same transcript are not compelled to be grouped within the same cluster. Given the varying degrees of density and shapes exhibited by HDBSCAN clusters, centroid-based topic representations are not necessarily anticipated; thus, to create topic representations that do not rely heavily on cluster structure assumptions, we employed a bag-of-words approach. All words within a cluster were aggregated into a single document (Grootendorst, 2022), and from that bag-of-words representation we sought to learn what distinguished one cluster from another. We used a class-based TF-IDF (cTF-IDF) approach for this, focusing on topics rather than individual documents or words (Grootendorst, 2022). Finally, we trained our topic models based on these pipelined sub-models.

4.1.2. Experiments and Results

We interpreted generated topics using the BERTopic *visualize_topics()* and *visualize_documents()* functions. *visualize_topics()* is inspired by the LDAVis method, which represents topics in two-dimensional space using circles to represent topics and the distance between them to represent topic similarity (Sievert and Shirley, 2014). Figures 2a–2f depict topic distributions for each diagnostic group $\times$ scene.

In Scene 1, we observed that HC discussed fewer topics when introducing themselves compared to BD and SZ. Most of the less frequent topics were closely related to more common topics, forming large clusters that were distant from each other. This suggests that individuals in the HC group tend to stay focused and on-topic during their conversations. In contrast, BD and SZ exhibited a different pattern, being less direct in their communication and often diverging from the main topic. On average, the BD group discussed a greater number of topics, and the distances between these topics were larger. We can infer that BD patients tend to become more easily distracted during their conversations and convey a wider range of information compared to other groups.

In Scene 2, patients were specifically asked to confront their landlords. This task focus reduced the number of discovered topics across all groups. We observed that participants in the HC group predominantly explained their concerns to the landlord by adhering to a single topic, whereas people in the BD and SZ groups often strayed to different topics and lost focus while addressing their issues.

4.2. Theme Identification

The ability of LLMs to generalize and pick up information in context has improved rapidly (Brown et al., 2020), and this process removes the need to backpropagate and update weights like in supervised fine-tuned settings. This saves training time and allows us to leverage compute resources for direct inference. We identify topical themes by prompting an encoder-decoder architecture. The flan model which extends t5, `flan-ul2`, has been proven useful and reliable for summarization tasks (Raffel et al., 2023). This model also provides a

(a) BD Subjects, Scene 1
(b) BD Subjects, Scene 2
(c) HC Subjects, Scene 1
(d) HC Subjects, Scene 2
(e) SZ Subjects, Scene 1
(f) SZ Subjects, Scene 2

Figure 2: Topic visualizations with intertopic distance maps.

large receptive field of 2048 which makes it ideal for our zero-shot task setting.

We frame theme identification similarly to text summarization, with the inputs being the topics produced by BERTopic and the output being theme titles. We use the following prompt and provide seven phrases (topic words identified by the topic model) as demonstrated below:

```
For these given phrases identify a
theme that captures all the phrases.

Phrase 1
Phrase 2
.
.
Phrase 7
```

We then expect the model to return a theme consisting of a word or short phrase capturing the essence of the input phrases. We use `flan-ul2` with PEFT-LoRA (Mangrulkar et al., 2022) and 8-bit quantization to account for GPU limitations. Inference is run on 7 T4 GPUs. The results are shown in Table 6.

From the identified themes, we observe that for Scene 1, the HC group quickly understands the task at hand, generally referring to a *new neighborhood*. We also observe that members of the BD group appear to discuss topics that are closely related but tangential to introductions, such as their cat or landlords. For members of the SZ group we observe a clear difference from the other groups. The first theme is *Don't know what to say*, which may align with catatonic behavior observed in people with schizophrenia (Jain and Mitra, 2020).

For Scene 2, we observe a similar pattern. Members of the HC group quickly discuss *Tenant rights*, whereas members of the BD and SZ groups reach the same theme later. We observe that HC participants can consistently maintain focus as opposed to other groups, as demonstrated through both the visualizations and the identified topic themes.

5. Discussion and Conclusion

In this paper we systematically investigated the reliable and trustworthy use of NLP methods for clinically enriched data. We studied patients' ages, genders, and clinical diagnoses in concert with their transcribed speech in a clinically validated spontaneous speech task. Through a study of 138 language features, we assessed feature importance and found that certain features are more associated with certain demographic traits. We also conducted multi-faceted statistical tests to dis-

Subject + Scene	Topic	Theme
BD Scene 1	Topic 0 Topic 1 Topic 2	My Cat Landlord Welcome
HC Scene 1	Topic 0 Topic 1 Topic 2	New Neighborhood Okay I am
SZ Scene 1	Topic 0 Topic 1 Topic 2	Don't know what to say Landlord Nice to meet you
BD Scene 2	Topic 0 Topic 1 Topic 2	how long will it take Tenant Rights leak
HC Scene 2	Topic 0 Topic 1 Topic 2	Tenant Rights Okay Thank You
SZ Scene 2	Topic 0 Topic 1 Topic 2	Water Tenant Rights Leak

Table 6: FLAN-UL2 theme identification.

cover which features reliably differentiate between diagnostic groups. We also demonstrate that the original set of 144 features can be reduced to 25 without performance reduction, helping us know which features are noisy and which are relevant.

Later, we showed that unsupervised topic modeling using encoder-based LLMs reveals clinically supported patterns. For instance, we see that members of the HC group exhibit better focus and more conciseness, relation, and close clustering among discussion points as opposed to members of the BD or SZ groups, as supported by detailed visuals and analyses. Finally, we prompted the `flan-ul2` model to identify themes from conversations in each subject group. Across both scenes and all groups we observed that HC participants arrived at desired topics more quickly and remained focused on them over the long term. Members of the other groups, and especially SZ, seemed less on-target, with many participants seeming unsure of what to say or how to start a friendly scene. This was also replicated in the more confrontational Scene 2.

An overarching outcome of this study was the observation that feature engineering and language modeling approaches carry separate but complementary advantages when analyzing this data. For instance, protected data such as ours cannot be run on remote servers (e.g., those used to serve OpenAI APIs) which record data logs. However, engineering many features and then intelligently reducing that feature set to statistically significant subgroups shows us which characteristics best discriminate between groups. While some features are better for understanding demographic splits of age or gender, others are better for understanding diagnostic labels or for (importantly) upholding known clinical truths. NLP in healthcare has often been plagued by explainability issues, but we observed that modern and older methods are able to beautifully and visually showcase patterns in data that have been previously suggested in clinical studies. Even completely unsupervised approaches such as our theme identification technique show how properly used LLMs can provide us with useful insights. While we still question whether some LLM predictions can be trusted, we can trust clinically-grounded insights for which LLMs validate previously hypothesized patterns in latent spaces of rich data. We conclude by hoping this leads to future work towards informed NLP use in clinical spaces, advancing progress toward explainable and reasonable conclusions.

6. Limitations

In this paper, we studied how NLP may be leveraged to analyze clinically enriched spontaneous speech. Our participant size, although large compared to contemporary relevant studies, was limited compared to that seen in many NLP task domains. We reported results over all participants, but note that a larger sample size would enable additional conclusions; it may also lead to slightly different performance distributions.

We did not add any new features beyond those introduced in our prior work (Aich et al., 2022), and preserved the full dataset from that prior work. Due to resource constraints, we used a language model with 20B parameters for our prompting-based theme identification, although it is known that models less than 40B-65B do not always perform optimally in prompting settings. Finally, we did not use any models that required running an API on a remote server since this would violate user privacy by relaying sensitive data and phrases to a third-party source. All models were run and experiments conducted locally.

7. Ethical Considerations

This paper uses real human data from generous participants. We do not intend for this paper to be interpreted as understanding the medical complexities and nuances of lifelong psychiatric illnesses such as schizophrenia or bipolar disorder. Our findings merely show how NLP techniques can provide a new perspective to the understanding and interpretation of the effects of these illnesses.

Data is stored in secure servers on laboratory computers with multi-factor authenticated security systems. At any point, only approved entities have access to the data. This data was originally collected under an approved Institutional Review Board (IRB) protocol at the University of California San Diego, and all uses of the data in this paper are in keeping with the data use provisions of that protocol. We refer readers to Aich et al. (2022) for a detailed description of the data collection and preservation procedures.

Acknowledgements

We thank the anonymous reviewers for their insightful feedback, and the participants in this dataset for contributing to the advancement of our research. Research reported in this publication was partially supported by the National Institute of Mental Health of the National Institutes of Health under award number R01MH116902. A. Aich was also supported as a research assistant during the development of this work by the National Institute of Nursing Research of the National Institutes of Health under award number 1R41NR020667-01. A. Aich was a doctoral student at the University of Illinois Chicago at the time of paper submission and a postdoctoral trainee at the National Institutes of Health at the time of paper acceptance. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

8. 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, Avery Quynh, Varsha Badal, Amy Pinkham, Philip Harvey, Colin Depp, and Natalie Parde. 2022. Towards intelligent clinically-informed language analyses of people with bipolar disorder and schizophrenia. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2871–2887, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ayo Akinyelu and Aderemi Adewumi. 2014. Classification of phishing email using random forest machine learning technique. *Journal of Applied Mathematics*, 2014.

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).

Yi Ji Bae, Midan Shim, and Won Hee Lee. 2021. Schizophrenia detection using machine learning approach from social media content. *Sensors*, 21(17):5924.

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.

Paul Best, Roger Manktelow, and Brian Taylor. 2014. Online communication, social media and adolescent wellbeing: A systematic narrative review. volume 41.

Viv Bewick, Liz Cheek, and Jonathan Ball. 2004. Statistics review 9: one-way analysis of variance. *Critical care*, 8:1–7.

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. 2022a. The development and psychometric properties of liwc-22.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022b. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

John B Carroll. 1964. Language and thought. *Reading Improvement*, 2(1):80.

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.

Adrian Andrzej Chrobak, Aleksander Turek, Karolina Machalska, Aleksandra Arciszewska-Leszczuk, Anna Starowicz-Filip, Anna Julia Krupa, Dominika Dudek, and Marcin Siwek. 2022. Graph analysis of verbal fluency tests in schizophrenia and bipolar disorder. *Brain Sciences*, 12(2):166.

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.

Mengjie Deng, Yunzhi Pan, Li Zhou, Xudong Chen, Chang Liu, Xiaojun Huang, Haojuan Tao, Weidan Pu, Guowei Wu, Xinran Hu, et al. 2018. Resilience and cognitive function in patients with schizophrenia and bipolar disorder, and healthy controls. *Frontiers in psychiatry*, 9:279.

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.

Ethan Fast and Eric Horvitz. 2016. Identifying dogmatism in social media: Signals and models.

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

Alexander Genkin, David Lewis, and David Madigan. 2005. Sparse logistic regression for text categorization. *DIMACS Working Group on Monitoring Message Streams Project Report*.

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

Stephanie Glen. 2016. Tukey test/tukey procedure/honest significant difference. *StatisticsHowTo. com: Elementary Statistics for the rest of us*.

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.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

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*.

David Hughes, Moss Rowe, Mark Batey, and Andrew Lee. 2012. A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage. volume 28, pages 561–569.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018a. 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.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018b. 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.

Georgiana Ifrim, Gökhan Bakir, and Gerhard Weikum. 2008. Fast logistic regression for text categorization with variable-length n-grams. *Proceedings of the 14th ACM KDD International Conference on Knowledge Discovery & Data Mining, ACM, 354-362 (2008)*.

Ankit Jain and Paroma Mitra. 2020. Catatonic schizophrenia.

Lauren Jelenchick, Jens Eickhoff, and Megan Moreno. 2013. "facebook depression?" social networking site use and depression in older adolescents. volume 52, pages 128–30.

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

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Machine Learning: ECML-98*, pages 137–142, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

Keumhee Kang, Chanhee Yoon, and Eun Yi Kim. 2016. Identifying depressive users in twitter using multimodal analysis. In *2016 International Conference on Big Data and Smart Computing (BigComp)*, pages 231–238, Los Alamitos, CA, USA. IEEE Computer Society.

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.

Timo K Koch, Peter Romero, and Clemens Stachl. 2022. Age and gender in language, emoji, and emoticon usage in instant messages. *Computers in Human Behavior*, 126:106990.

Paul Komarek and Andrew Moore. 2004. Fast logistic regression for data mining, text classification and link detection. In *Proceedings of NeurIPS*.

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.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

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.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

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. 2013a. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Saif M Mohammad and Peter D Turney. 2013b. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

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.

Shraddha Parab and Supriya Bhalerao. 2010. Choosing statistical test. *International journal of Ayurveda research*, 1(3):187.

Albert Park, Mike Conway, et al. 2018. Harnessing reddit to understand the written-communication challenges experienced by individuals with mental health disorders: analysis of texts from mental health communities. *Journal of medical Internet research*, 20(4):e8219.

Thomas Patterson, Sherry Moscona, Christine Mckibbin, Kevin Davidson, and Dilip Jeste. 2001a. Social skills performance assessment among older patients with schizophrenia. *Schizophrenia research*, 48:351–60.

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

Thomas L Patterson, Sherry Moscona, Christine L McKibbin, Kevin Davidson, and Dilip V Jeste. 2001c. Social skills performance assessment among older patients with schizophrenia. *Schizophrenia research*, 48(2-3):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. 2016. Application of logistic regression with part-of-the-speech tagging for multi-class text classification. In *2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, pages 1–5. IEEE.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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.

Amanda Ross and Victor L Willson. 2017. One-way anova. In *Basic and advanced statistical tests*, pages 21–24. Brill.

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

Elizabeth Seabrook, Margaret Kern, and Nikki Rickard. 2016. Social networking sites, depression, and anxiety: A systematic review. volume 3, page e50.

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.

S. T. Selvi, P. Karthikeyan, A. Vincent, V. Abinaya, G. Neeraja, and R. Deepika. 2017. Text categorization using rocchio algorithm and random forest algorithm. In *2016 Eighth International Conference on Advanced Computing (ICoAC)*, pages 7–12.

Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah. 2020. A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augmented Human Research*, 5.

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.

Carson Sievert and Kenneth Shirley. 2014. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70.

Teague Simoncic, Kate Kuhlman, Ivan Vargas, Sean Houchins, and Nestor Lopez-Duran. 2014. Facebook use and depressive symptomatology: Investigating the role of neuroticism and extraversion in youth. volume 40, page 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.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge.

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.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms.

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

Joan Torruella and Ramón Capsada. 2013. Lexical statistics and tipological structures: a measure of lexical richness. *Procedia-Social and Behavioral Sciences*, 95:447–454.

Umar Toseeb and Becky Inkster. 2015. Online social networking sites and mental health research. volume 6.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Pavel Veretilo and Stephen Billick. 2012. Psychiatric illness and facebook: A case report. volume 83, pages 385–9.

Rohit Voleti, Stephanie Woolridge, Julie M. Liss, Melissa Milanovic, Christopher R. Bowie, and Visar Berisha. 2019. Objective Assessment of Social Skills Using Automated Language Analysis for Identification of Schizophrenia and Bipolar Disorder. In *Proc. Interspeech 2019*, pages 1433–1437.

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.

Luisa Weiner, Nadège Doignon-Camus, Gilles Bertschy, and Anne Giersch. 2019. Thought and language disturbance in bipolar disorder quantified via process-oriented verbal fluency measures. *Scientific reports*, 9(1):14282.

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.

Qingyao Wu, Yunming Ye, Haijun Zhang, Michael Ng, and shen shyang ho. 2014. Forestexter: An efficient random forest algorithm for imbalanced text categorization. *Knowledge-Based Systems*, 67.

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)*.

Show-Jane Yen, Yue-Shi Lee, Jia-Ching Ying, and Yu-Chieh Wu. 2011. A logistic regression-based smoothing method for chinese text categorization. *Expert Syst. Appl.*, 38:11581–11590.

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. 2019a. Linguistic analysis of schizophrenia in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 74–83.

Jonathan Zomick, Sarah Ita Levitan, and Mark Serper. 2019b. 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.