NIMH
National Institute
of Mental Health

A New Dataset for Language Analysis in Serious Mental Illness

AI/ML Readiness for NIH Data Supplement

Supplement to: NIMH R01 1169032 (PI: Depp)

UC San Diego

UIC

Motivations

- Language abnormalities have been linked to psychosis for over 100 years

- Feature Engineering and Feature Exploration have proven to be useful for NLP and Psycholinguistic Tasks (Aich and Parde 2022).

- Yet, translating Psycholinguistics to clinical use has been slow with barriers including reproducibility, bias, and limited outcome variables (Cohen et al. 2022)

Key gap: Major datasets in Computational Psycholinguistics lack deep clinical validation data (Coppersmith et al 2016).

Data Source

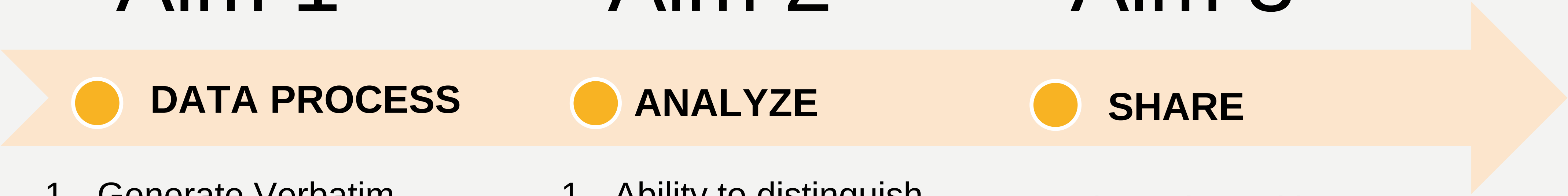
- Patient and healthy control audio recordings of the Social Skills Performance Assessment (SSPA; Patterson et al., 2001)
- The SSPA is a standardized observer-rated measure of Social Abilities
- It involves two open-ended interactions with a trained rater/confederate:
  - Meeting a new neighbor
  - Complaining to a landlord about a leak
- Performance is manually scored by reliable raters on dimensions like coherence, affect, etc.
- SSPA scores are predictive of functional outcome (Miller et al., 2021)

Project Timeline

Aim 1

Aim 2

Aim 3



DATA PROCESS

ANALYZE

SHARE

Aim 1

1. Generate Verbatim Transcripts from SSPA recordings
2. De-Identify Transcripts
3. Create harmonized Linked Clinical Dataset
4. Process Transcripts for NLP
5. Feature Extraction

Aim 2

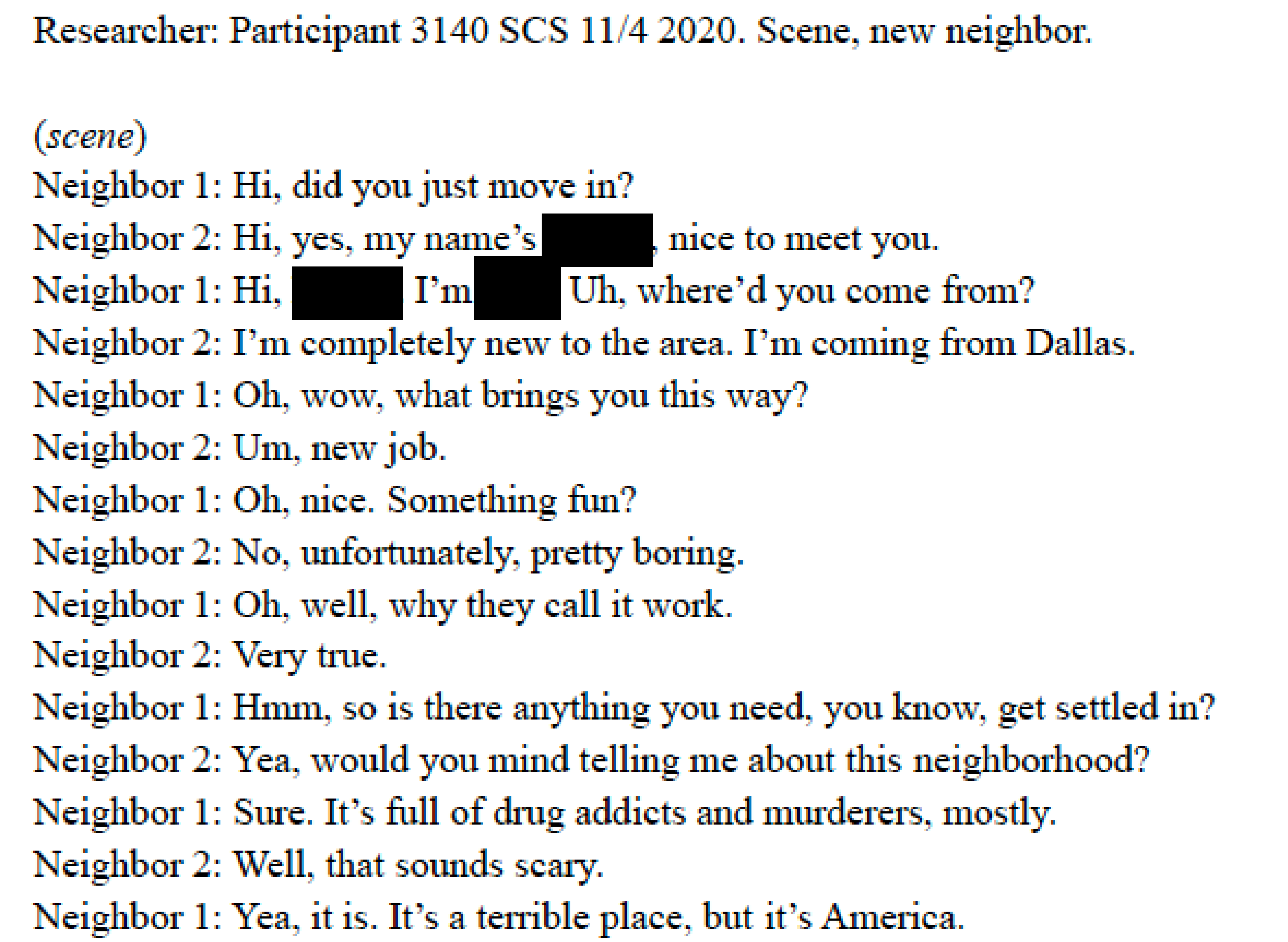
1. Ability to distinguish diagnoses
2. Bias of prediction by demographic variation
3. Ability to predict SSPA score and other clinical data
4. Feature domains most predictive of above

Aim 3

Share data with NIMH National Data Archive

Create tools for investigators

Sample Transcript Excerpt



Researcher: Participant 3140 SCS 11/4 2020. Scene, new neighbor.

(*scene*)
Neighbor 1: Hi, did you just move in?
Neighbor 2: Hi, yes, my name's [redacted], nice to meet you.
Neighbor 1: Hi, [redacted] I'm [redacted] Uh, where'd you come from?
Neighbor 2: I'm completely new to the area. I'm coming from Dallas.
Neighbor 1: Oh, wow, what brings you this way?
Neighbor 2: Um, new job.
Neighbor 1: Oh, nice. Something fun?
Neighbor 2: No, unfortunately, pretty boring.
Neighbor 1: Oh, well, why they call it work.
Neighbor 2: Very true.
Neighbor 1: Hmm, so is there anything you need, you know, get settled in?
Neighbor 2: Yea, would you mind telling me about this neighborhood?
Neighbor 1: Sure. It's full of drug addicts and murderers, mostly.
Neighbor 2: Well, that sounds scary.
Neighbor 1: Yea, it is. It's a terrible place, but it's America.

Data Available

	Total (n = 558)	Schizophrenia (n = 229)	Bipolar Disorder (n = 228)	Healthy Control (n = 101)
Age	41.3 ± 12.0, range 18-65	43.0 ± 11.6	39.6 ± 12.0	41.7 ± 12.4
Gender (% Female)	54.5%	51.5%	59.6%	49.5%
Years of Education	13.7 ± 2.5	13.3 ± 2.7	14.2 +2.4	13.2 ± 2.0
Race (% Afr-American)	36%	52.4%	18.4%	38.6%
Ethnicity (% Hisp/Latino)	23.2%	21.8%	27.8%	15.8%

<u>Data derived from</u>: **Parent Study (n=300) as well as SCOPE and IA studies (n=500+; PI: Amy Pinkham)**

<u>Repeat (longitudinal) data available</u>: n=188 ; <u>Harmonized</u> measures of symptoms, cognition, and many more

Pre-processing steps - Deep learning

- Regex and Timestamp Extraction.
- Grouping of patient dialogues.
- Vectorization and Tokenization of dialogues
- Hyper-parameter setting for Transformer models.

Feature Extraction

- Temporal
- Sentiment
- Psycholinguistic / Affective
- Lexically informed emotions
- Diversity of speech
- Human features such as race, diagnosis, age etc.

Results: Do SSPA NLP features Differentiate Diagnoses?

- KNN, RF, SVM, Logistic, and Ridge Classification.
- One v One classification between diagnoses

Accepted: Ankit Aich, Avery Quynh, Varsha Badal, Amy Pinkham, Philip Harvey, Colin Depp, and Natalie Parde in Findings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022) - December 7-11 Abu Dhabi, UAE

Results - READ AS [ACCURACY | F1] (SC1 VS SC2)

Model	BD v SZ	BD v HC	HC v SZ	BD v SZ	BD v HC	HC v SZ
RF	.93 | .87	.96 | .84	.96 | .96	.96 | .94	.92 | .96	.70 | .93
KNN	.58 | .64	.51 | .59	.82 | .75	.37 | .62	.71 | .69	.66 | .48
LR	.89 | .91	.82 | .90	.89 | .83	.86 | .97	.89 | .78	.55 | .62
Ridge	.89 | .94	.86 | .70	.93 | .72	.93 | .97	.78 | .78	.70 | .70
SVM	.89 | .91	.86 | .67	.93 | .72	.89 | .97	.89 | .79	.60 | .75

Accepted: Ankit Aich, Avery Quynh, Varsha Badal, Amy Pinkham, Philip Harvey, Colin Depp, and Natalie Parde in Findings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022) - December 7-11 Abu Dhabi, UAE

Next Steps for Project

- Evaluating bias, feature specificity, and clinical symptom effects
- Explore potential for automated scoring of the SSPA
- Perform speech analysis and create corresponding feature extracted dataset
- Share transcript and feature data through NIMH Data Archive and create tools to support AI/ML researchers using the data

THANK YOU!

Please contact us if interested in collaborating:

- **Colin Depp (UCSD): cdepp@ucsd.edu**
- **Natlie Parde (UIC): parde@uic.edu**
- **Ankit Aich (UIC): aaich2@uic.edu**