Evaluating Language Models and Preprocessing Strategies for Detecting Psychosis-Related Language Disturbances

Amir Nikzad[1], Yan Cong[2], Sunghye Cho[3], Ryan Partlan[1], Jiefei Li[1], Sunny Tang[1]

1. Zucker Hillside Hospital, Northwell Health. 2. School of Languages and Cultures, Purdue University. 3. Department of Linguistics, University of Pennsylvania.

Introduction

- **Background:** Language models can detect psychosis-related language disturbances by measuring semantic similarity between speech units, but no standardized approach exists for model selection or transcript preprocessing.
- **Models Compared:** Static models (GloVe, LSA, word2vec) vs. contextual models (GPT-2, RoBERTa, LLaMA), which differ in their ability to capture semantic and grammatical context.
- **Preprocessing Levels:** Verbatim (Level 1), removal of disfluencies/repetitions (Level 2), and additional stop-word removal (Level 3).
- **Evaluation:** Models are assessed using established semantic similarity metrics and correlated with categorical SSD diagnosis and severity of three language disturbance dimensions (impaired expressivity, inefficient speech, incoherence).

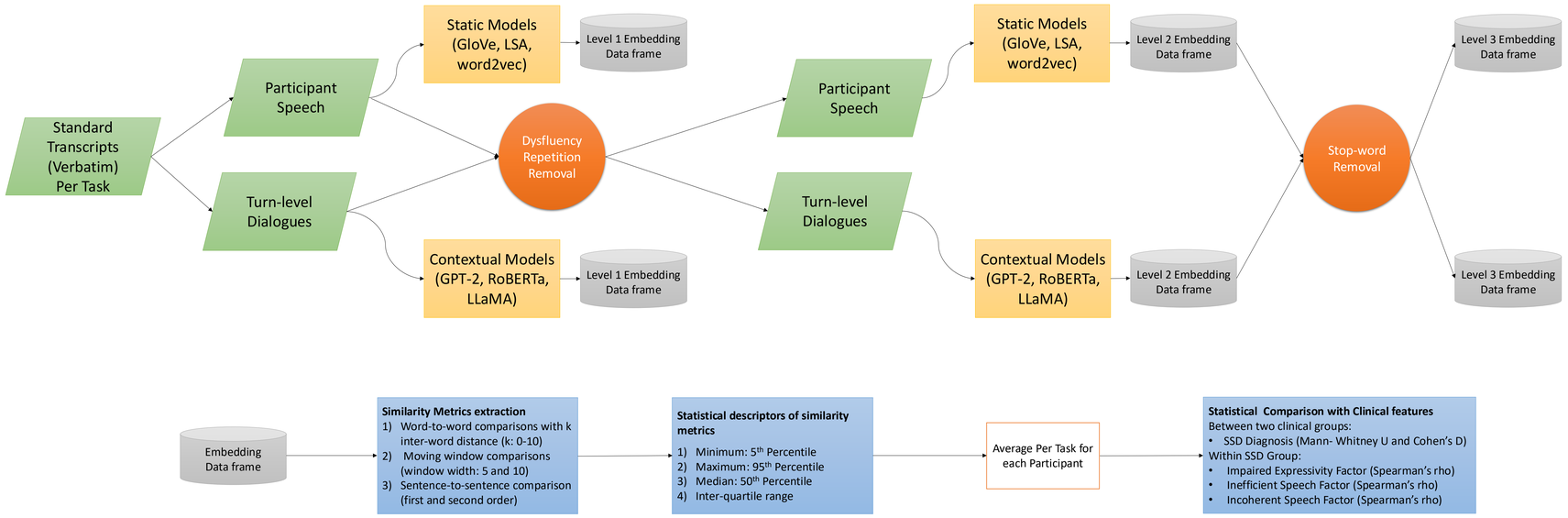
Participant Characteristics

	Healthy Control	SSD	p Value
n (%)	76 (33%)	152 (67%)	
Age mean (SD)	29 (7)	28 (7)	0.285
Sex			0.016*
Female (%)	41 (54%)	55 (36%)	
Male (%)	35 (46%)	97 (64%)	
Race			0.141
Asian	12 (16%)	19 (12%)	
Black	27 (36%)	58 (38%)	
White	30 (39%)	45 (30%)	
Other	7 (9%)	30 (20%)	
TLC Total Score	2 (3)	17 (14)	0.000***
Average Factor Scores			
Impaired Expressivity	-0.30	0.28	0.000***
Inefficiency	-0.58	0.59	0.000***
Incoherence	-0.38	0.53	0.000***

Feature Description and Nomenclature

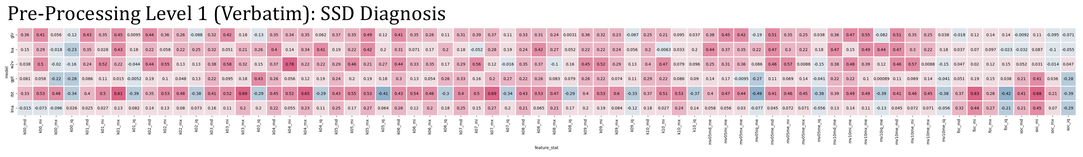
- **K-inter-word distances:** Compares word-to-word similarity at 0–10-word intervals. For example, k00 Compares each word to the next word (immediate neighbor, 0-word interval), whereas k01 compares each word to the word after the next (skipping one word, 1-word interval).
- **Moving window similarity:** Calculates average similarity of all word pairs within a fixed window size of 5 (mv05) or 10 (mv10).
- **Sentence similarity:** Compares sentence to sentence similarity between adjacent sentences (first–order coherence: foc) or between each sentence and the sentence after the next (second-order coherence: soc)
- **Statistical descriptors:** Comparisons across a sample generate a set of similarity metrics, which are summarized using minimum (mi: 5th percentile), median (md), maximum (mx: 95th percentile), and inter-quartile range (iq). For example, k03_mx represents the 95th percentile similarity value for inter-word comparisons at a 3-word interval.

Processing Pipeline

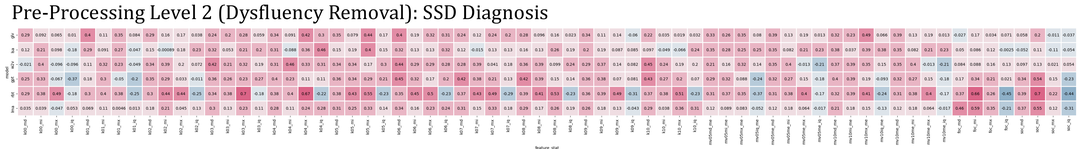


Effect Size Comparison

Pre-Processing Level 1 (Verbatim): SSD Diagnosis

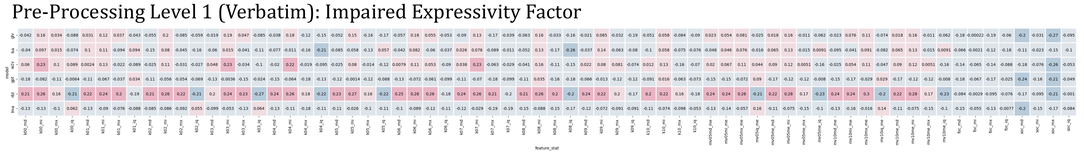


Pre-Processing Level 2 (Dysfluency Removal): SSD Diagnosis

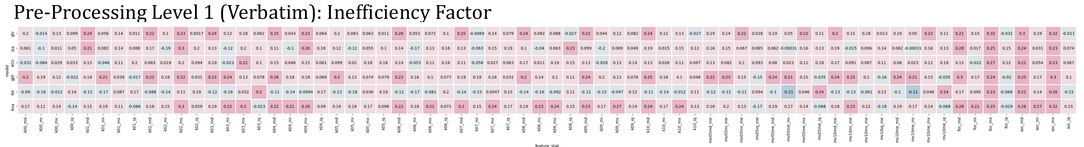


Pre-Processing Level 3 (Dysfluency + Stopword Removal): SSD Diagnosis

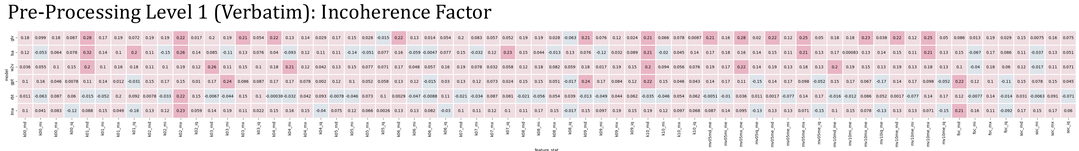
Pre-Processing Level 1 (Verbatim): Impaired Expressivity Factor



Pre-Processing Level 1 (Verbatim): Inefficiency Factor



Pre-Processing Level 1 (Verbatim): Incoherence Factor



Key Findings

- **Preprocessing Impact**: Levels 1, 2, and 3 produced 181, 173, and 131 significant associations ($p < 0.05$) between semantic similarity features and SSD diagnosis, with mean absolute CD values of 0.40, 0.36, and 0.39, respectively. Verbatim transcripts (Level 1) outperformed other preprocessing strategies.
- **Model Performance in SSD Classification**: RoBERTa generated the most significant features at Level 1 (67; CD=0.40), followed by word2vec (36; 0.41), GloVe (33; 0.39), LSA (23; 0.36), LLaMA (14; 0.28), and GPT-2 (8; 0.32). RoBERTa remained the top performer across preprocessing levels.
- **Language Disturbance Dimensions**: RoBERTa showed the strongest association with impaired expressivity (61 correlations) but performed poorly for incoherence and inefficient speech, where LLaMA and GloVe outperformed other models, respectively. Overall, models had similar performance across language disturbance dimensions, with max and mean absolute rho values of ~0.30 and ~0.21 for all three factors.
- **Main Takeaways**: Verbatim preprocessing (Level 1) and RoBERTa outperformed other approaches in SSD classification. No categorical advantage was observed for contextual or larger models, as static models performed comparably across multiple evaluation metrics.

References

For the three-factor model of cross-diagnostic language disturbances, see:
Sunny X Tang, Katrin Hänsel, Yan Cong, Amir H Nikzad, Aarush Mehta, Sunghye Cho, Sarah Berretta, Leily Behbehani, Sameer Pradhan, Majnu John, Mark Y Liberman, Latent Factors of Language Disturbance and Relationships to Quantitative Speech Features, *Schizophrenia Bulletin*, Volume 49, Issue Supplement_2, March 2023, Pages S93–S103, https://doi.org/10.1093/schbul/sbac145