Approximating the semantic space: word embedding techniques in psychiatric speech analysis

C. Palominos [1] K. Fröhlich [1] F. Stein [2] T. Kircher [2] W. Hinzen [1 3]

[1]Universitat Pompeu Fabra [2]Philipps Universität Marburg [3]Institució Catalana de Recerca i Estudis Avançats (ICREA)

Universitat Pompeu Fabra Barcelona

TRUSTING

Funded by the European Union

Word embedding techniques in psychiatric speech analysis

- LLMs are utilized in recent research to measure semantic distances and explore changes in psychosis, shedding light on cognitive mechanisms but yielding mixed results on coherence in psychotic speech.
- Semantic similarity metrics in psychosis research are complex and influenced by various factors, with a suggested shift towards understanding semantic space navigation, revealing patterns of cognitive processing differences in psychosis.
- This study extends previous research to German, comparing word embedding techniques and introducing dynamic metrics across chronic schizophrenia, schizoaffective (SSD), and major depressive disorders (MDD), aiming to explore specificity in these conditions.

Methods

- **Data collection** 129 German speakers: 43 SSD, 43 MDD, 44 healthy controls (HC). Speech samples from four pictures descriptions (3 minutes each), from Thematic Apperception Test (TAT) ([3]).
- **Word and sentence embeddings** Used fastText ([2]) and BERT ([1]) for words, and SentenceTransformers ([4]) for sentences.
- **Sentence embedding centroids** Averaged dimensions to distinguish groups/pictures.
- **Semantic similarity** Analyzed via mean, max, min, slope sign change (SSC), mean crossing and autocorrelation of item pairs derived from the wave function of semantic similarity values.
- **Convex hull and dimensionality reduction** Samples as hyper-polyhedrons from embeddings, volume and area measured after t-Distributed Stochastic Neighbor Embedding (t-SNE) ([5]).
- **Statistical analysis**
  - k-nearest neighbors algorithm (kNN) applied post-dimensionality reduction using t-SNE for picture and group classification.
  - Mixed-effects models for group semantic differences, controlling for picture and speech sample length.

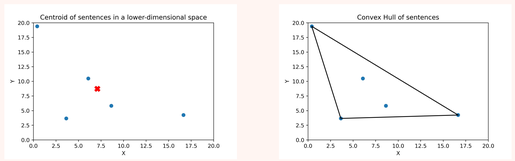

Figure 1. Centroid embedding of sentences (left), and convex hull (right).

Results: 1

- **Baseline content analysis**: From group classification based on centroids of speech, we can imply that any variations we observe in semantic variables could be attributed to how individuals navigate the same semantic space.
- **Picture effect**: Different pictures have a significant effect on semantic similarity variables. In addition to controlling this effect, further research is relevant in this regard.

Results: 2

- **Static semantic similarity variables**: We report lack of significant differences between groups in the mean of semantic similarity, but we found significance increase in maximum semantic similarity in SSD for BERT model.
- **Dynamic semantic similarity variables**: We found significant differences in dynamic variables (see Table 3)
- **Displacement**: Larger displacement in SSD relative to HC, despite the unchanged centroids and mean semantic similarities.
- **Dispersion**: Larger dispersion of sentence embeddings in MDD relative to HC.
- **Convex hull volume**: Significant increase in the volume of the convex hull in SSD compared to HC.

Classification of text centroids

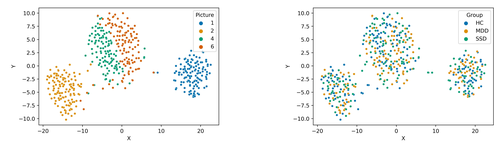

Figure 2. Classification of text centroids from sentence embeddings in 2D. By picture (left), and group (right).

Static and dynamic semantic variables

Table 1. Summary of groups effects on semantic similarity variables.

	FastText		BERT	
Variable	MDD	SSD	MDD	SSD
mean semsim	n.s.	n.s.	n.s.	n.s.
max semsim	n.s.	n.s.	n.s.	Positive
min semsim	n.s.	n.s.	n.s.	n.s.
average ssc	n.s.	n.s.	n.s.	Negative
average crossing	Negative	n.s.	n.s.	n.s.
autocorrelation	Positive	n.s.	n.s.	n.s.

Displacement

Table 2. Summary of Mixed Linear Model Regression results for Cumulative Euclidean distance

	Coefficient	Std. Error	z	$P > \|z\|$	[0.025	0.975]
Intercept	-16.469	4.813	-3.422	0.001	-25.902	7.036
SSD	18.649	4.516	4.516	0.000	9.798	27.501
Content words	1.199	0.036	33.655	0.000	1.129	1.268
Av sentence length	-0.407	0.130	-3.139	0.002	-0.661	-0.153

Dispersion

Table 3. Summary of Mixed Linear Model Regression results for dispersion

	Coefficient	Std. Error	z	$P > \|z\|$	[0.025	0.975]
Intercept	0.707	0.010	68.545	0.000	0.687	0.728
MDD	0.015	0.005	2.817	0.005	0.005	0.026
SSD	0.008	0.005	1.536	0.124	-0.002	0.019
Picture 2	-0.022	0.006	-3.703	0.000	-0.033	-0.010
Picture 4	0.013	0.006	2.147	0.032	0.001	0.024
Picture 6	-0.003	0.006	-0.465	0.642	-0.014	0.009
N of sentences	-0.003	0.000	-9.423	0.000	-0.003	-0.002
Av sentence length	-0.001	0.000	-1.782	0.075	-0.001	0.000

Volume of convex hull

Table 4. Summary of Mixed Linear Model Regression results for volume

	Coefficient	Std. Error	z	$P > \|z\|$	[0.025	0.975]
Intercept	0.153	0.664	0.230	0.818	-1.148	1.453
MDD	0.433	0.860	0.503	0.615	-1.253	2.118
SSD	1.737	0.867	2.004	0.045	0.038	3.436
above median	3.018	0.537	5.623	0.000	1.966	4.070

Discussion

- This study aimed to refine semantic analysis in SSD using LLM embedding, focusing on the 'shrinking' semantic space.
- Despite null results, accounting for factors like sample and sentence length, reveals diverging trajectories across semantic space in clinical groups, with MDD exhibiting higher autocorrelation and less crossing of the average line, and SSD showing less slope sign changes, consistent with a more restricted semantic space.
- Semantic distances/similarities collapse high-dimensional spaces, while Euclidean distances preserve geometrical relationships.
- Despite shared semantic domains, clinical groups exhibit divergent trajectories around centroids, indicating different navigation patterns.
- Sensitivity to SSD with BERT embeddings and to MDD with fastText embeddings suggests varying patterns at contextual vs. lexical-conceptual semantic levels in clinical speech.

Funding

Funded by the European Union (GA 101080251 - TRUSTING). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Agency. Neither the European Union nor the granting authority can be held responsible for them.

References