Linguistic Features Competitive with Bert! Leveraging Speech for Detection of Mental Health in Paediatric Lupus

Jida Jaffan
Psychology Department, University of Toronto
Neurosciences and Mental Health Program, The Hospital for Sick Children
jida.jaffan@mail.utoronto.ca

Barend Beekhuizen
Department of Language Studies
University of Toronto Mississauga
barend.beekhuizen@utoronto.ca

Andrea Knight
Division of Rheumatology
The Hospital for Sick Children
University of Toronto
andrea.knight@sickkids.ca

Abstract

Neuropsychiatric lupus (NPSLE) is characterized by inflammation in the brain with common symptoms of depression and anxiety. Early detection is crucial as it may change the treatment regimen; however, current approaches are costly and resource intensive. Therefore, we propose that leveraging current work using linguistics in NLP detection of mental health symptoms can be advantageous in early detection of NPSLE. This study is a proof-of-concept using 20 interviews from $N = 20$ adolescents (10-17 years) diagnosed with Lupus. Our results suggest that linguistic feature-based models supported by Word2Vec embeddings offer an interpretable output compared with BERT models, while maintaining competitiveness in depression, and improvement over BERT in anxiety detection. This work may transform early screening methods in paediatric contexts and can be adapted to other clinical populations.

1 Introduction

Depression, anxiety, and cognitive impairment are common in children with lupus. Systemic lupus erythematosus (SLE) is a chronic autoimmune disease characterized by inflammation affecting 1 in 1000 Canadians (Aggarwal et al., 2024). Childhood-onset SLE (cSLE) is associated with aggressive disease resulting in multiple organ damage. Inflammation can target the brain leading to neuropsychiatric symptoms, such as depression in 60%, anxiety in 40% and cognitive impairment in 40% of the clinical population (Valdés Cabrera et al., 2024; El Tal et al., 2022) and may lead to a diagnoses of Neuropsychiatric lupus (NPSLE). cSLE typically onsets during adolescence, overlapping with important developmental milestones and a period of substantial brain maturation (Mendelsohn et al., 2021). These factors exacerbate neuropsychiatric symptoms with significant adverse impact on health-related quality of life, medication adherence, and transition to adult care (Quilter et al., 2021). Early identification and treatment of neuropsychiatric symptoms are therefore crucial; however, multiple challenges including assessment burden for healthcare personnel, patients and families prevent timely intervention (Barraclough et al., 2019).

Language functions provide a gateway for identifying brain involvement in cSLE. Linguistic measures offer quick, efficient proxies of underlying brain function which can inform individualized stepped-care intervention models. To produce speech, one must plan and execute an utterance; if a relevant brain area is impacted, the resulting speech will be as well. For example, aphasia caused by inflammation in Broca's area impacts grammar, while inflammation in Wernicke's area impairs semantic content (Gagliardi and Tamburini, 2022). Although language has been shown to be robust in detecting early symptoms of brain involvement in other populations such as Alzheimer, autism, depression, schizophrenia, and PTSD (Gagliardi and Tamburini, 2022; Kho et al., 2025), to date, very limited work has assessed linguistics in SLE (Ceccarelli et al., 2019). Language samples from depressed individuals consistently show increased first-person pronouns (*I*, *me*, *my*), reflecting self-focused rumination (Corbin et al., 2023). Anxiety, in contrast, is associated with reduced semantic coherence, as heightened cognitive load leads to less organized discourse (Kho et al., 2025). Deficits in inhibition (a domain of executive function) also reduce coherence, producing tangential speech or off-topic responses (de Boer et al., 2020). Finally, increased positive emotional content is associated with longevity, suggesting that positive framing and

Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), pages 249–256
March 28, 2026 ©2026 Association for Computational Linguistics

balanced emotional states increase resilience to adversity (Zaffino et al., 2025). Language therefore offers sensitive, downstream realizations of cognitive systems, brain structure, and mental health outcomes, yet has not been studied in cSLE.

Recent work has utilized NLP for mental health screening with approaches ranging across sentiment analysis, linguistic features extraction (acoustic, lexical, syntactic), as well as learning models using multi-layer neural networks (Teferra et al., 2024). Gumus et al. (2023) evaluated 290 linguistic features, and reported a strong association between sentiment scores and depression. Linguistic Inquiry and Word Count (LIWC), a text analysis software, has shown promise in detecting depression from online posts (Coppersmith et al., 2014) with deep learning approaches leveraging contextual embeddings of first-person pronouns to predict depression severity (Ren et al., 2023). Previous work in paediatric settings has primarily used LLM-based NLP methods to extract disease and mental health symptoms from electronic records, yet has not been widely utilized on adolescent speech in a clinical setting (Ignashina et al., 2025).

Nonetheless, promising results have emerged in NLP adult clinical work focused on spontaneous speech of patients with Alzheimer's Disease (AD). Taghibeyglou and Rudzicz (2023) demonstrated that lightweight word2vec and linguistic feature-based models can outperform Bidirectional Encoder Representations from Transformers(BERT) in AD classification tasks. This is important as integration of screening tools to clinical settings depend on their accessibility. Linguistic-based approaches are more cost effective and easier to implement. In addition, transformer models are often criticized for the opacity of their processes, whereas linguistic features are interpretable and may map onto underlying cognitive processes.

Leveraging linguistic analysis will aid screening and provide a cost-effective method for detecting NPSLE in this population, which could revolutionize diagnostic approaches and have direct clinical utility. Early identification of brain involvement in children with cSLE will provide the foundation for earlier intervention, which is crucial to optimizing long-term outcomes in health-related quality of life. Translating this to other clinic populations will be the natural next step in this innovative approach.

Here, we explore the performance of several classification techniques that have been found to be successful in other medical domains on our domain of depression and anxiety.

2 Methods

Participants and Speech samples: This cross-sectional approach utilizes visit data from ongoing studies at The Hospital for Sick Children in Toronto, Canada. Youth (10-17y) with cSLE ($N = 20$) completed mental health assessments for depression (Beck Depression Inventory-II) and anxiety (Screen for Child Anxiety Related Emotional Disorders) as part of routine clinic visits and completed 30 minute guided interviews as part of broader research initiatives exploring mental health. $N = 12$ participants reported elevated depression (BDI > 11) and $N = 12$ reported elevated anxiety (SCARED > 25). Transcripts were extracted from meeting links and checked for accuracy by two team members. All transcripts were de-identified by removing any occurrences of proper names of people and locations (e.g., doctors, schools, and hospitals). Self-referral of the participants own name during the interview was replaced by the corresponding participant study ID in the transcript.

2.1 Preprocessing

All interviewer speech was removed from transcripts such that only participant speech was used for analyses. Text was converted to lowercase to ensure case-insensitive handling across models, and non-lexical punctuation was stripped. Tokenization was performed using spaCy (Honnibal and Montani, 2017) for the three models in Section 2.2, whereas inbuilt tokenizers were employed for the transformer-based models (Sec. 2.3). The first 512 tokens of each resulting transcript were used so that document lengths were identical.

2.2 Feature Extraction and Models

Three feature-based models were tested:

1. **Linguistic-Based Features (LBF):** A Linguistic feature–based (LBF) pipeline extracted spaCy-derived lexical/syntactic features (e.g., POS proportions, Mean Length of Utterance (MLU), Type-Token ratio, Open/Closed-Class Ratios, First/Second/Third-Person Pronoun Ratios), VADER from nltk sentence-level sentiment (mean/SD), and demographic covariates (Age, Sex).

2. **W2V Embedding Features:** Here we follow Taghibeyglou and Rudzicz (2023) who report

good performance using aggregated static vectors for diagnosis in a different domain. We used the pretrained Wikipedia2Vec model (Yamada et al., 2018) containing 100-dimensional embeddings trained on full English Wikipedia, aggregating the vectors of all words in each transcript into a 100-dimensional median/standard-deviation–standardized embedding vector.

3. **Combined LBF + W2V Features:** All linguistic, demographic, and word-embedding dimensions were combined for this model.

To reduce redundancy and identify the most informative predictors, we applied an automated feature selection procedure based on the framework described by Taghibeyglou and Rudzicz (2023), using FeatureWiz package (AutoViML, 2020). We set a correlation threshold of 0.6 and repeated the selection 5 times with seeds [0,1,2,3,4] over all samples. We kept all features that were selected in at least 3 iterations for further analyses.

2.3 BERT Models

We compared the performance of the previous models to six pre-trained transformer models. Each transcript was passed through the pretrained model in inference mode, and we extracted the final hidden-state vector corresponding to the [CLS] token, yielding a 768-dimensional embedding representing the transcript as a whole.

In line with Taghibeyglou and Rudzicz (2023), we tested three variants of the uncased BERT-base architecture (Devlin et al., 2019). The first model, BERT1, consisted solely of the 768-dimensional final layer embedding of the [CLS] token connected to a binary, one-dimensional output layer ($768 \rightarrow 1$). The second model, BERT2, had the same setup but with one hidden layer ($768 \rightarrow 64 \rightarrow 1$). The third variant, BERT3, added two hidden layers ($768 \rightarrow 128 \rightarrow 16 \rightarrow 1$), serving in our study as a deeper nonlinear transformation of the [CLS] embedding. In addition to these BERT-base variants, we evaluated Bio-Clinical BERT (Alsentzer et al., 2019), DistilBERT (Sanh et al., 2019), and BioMed-RoBERTa (Gururangan et al., 2020), all used in inference mode only. We used Binary Cross Entropy and AdamW optimizer (`learning rate` $= 2 \times 10^{-5}$) with a linear warm-up schedule (Loshchilov and Hutter, 2017).

2.4 Leave-One-Subject-Out Evaluation

To evaluate and compare the models' generalizability and ensure independence between training and test samples, we used a leave-one-subject-out (LOSO) cross-validation procedure, in which each participant served once as the held-out test case while the remaining participants were used for training. For each held-out subject, we recorded model predictions and aggregated performance across all LOSO iterations. Evaluation metrics included accuracy, sensitivity, specificity, and F_1-score. For all feature-based models, including the linguistic, Word2Vec, and hybrid feature sets, we evaluated a suite of classical machine-learning classifiers to ensure robust comparison across representations. These included logistic regression, support vector machines, decision trees, linear and quadratic discriminant analysis, Gaussian naïve Bayes, random forest and extra-trees ensembles, AdaBoost, and XGBoost, allowing performance to be assessed across both linear and nonlinear decision boundaries. To ensure replicability, models and classifiers used `random_state`$=6$.

3 Results

3.1 Selected Features across models

The following features were selected by FeatureWiz procedure, per dependent variable and per model (Full lists in Appendix A).
For **Depression**:

- *LBF:* 8 linguistic features were selected: MLU, Open/Closed-Class Ratio, proportion of Adjectives, Adpositions, Auxiliaries, and Proper Nouns, Total Token Count, and the variability in Sentence-level Sentiment.
- *W2V:* 32 w2v dimensions were selected.
- *LBF+W2V:* 5 linguistic features were selected: Full Length, Total Token Count, proportion of adpositions, Second-person, and Third-person Pronouns, as well as 28 w2v dimensions, with 23 overlapping with the w2v-only model.

For **Anxiety:**

- *LBF:* 9 linguistic features were selected: Full Length, Total Token Count, Proportion of Adverbs, Determiners, Proper Nouns, and Verbs, Second-person pronouns, and Sentence-level Sentiment Average and Variability across sentences per participant.
- *W2V:* 35 w2v dimensions were selected.

		Depression					Anxiety				
Model		**Class.**	**Acc.**	**Sens.**	**Spec.**	F_1	**Class.**	**Acc.**	**Sens.**	**Spec.**	F_1
LBF	max	DT	0.70	0.75	0.63	**0.75**	LR	0.80	0.75	0.88	**0.82**
	avg		0.52	0.69	0.25	0.62		0.68	0.80	0.49	0.75
W2V	max	GNB	0.65	0.92	0.25	**0.76**	ExtraTrees	0.75	0.92	0.50	**0.81**
	avg		0.53	0.65	0.34	0.61		0.58	0.72	0.36	0.66
LBF + W2V	max	LDA	0.65	0.83	0.38	**0.74**	GNB	0.70	0.75	0.63	**0.75**
	avg		0.54	0.68	0.33	0.63		0.51	0.63	0.34	0.60
baseBERT1			0.55	0.92	0.00	0.71		0.55	0.92	0.00	0.71
baseBERT2			0.45	0.75	0.00	0.62		0.60	0.92	0.13	0.73
baseBERT3			0.50	0.58	0.38	0.58		0.65	0.75	0.50	0.72
BioClinicalBERT			0.55	0.92	0.00	0.71		0.60	1.00	0.00	**0.75**
BioMed-RoBERTa			0.55	0.92	0.00	0.71		0.60	0.92	0.00	**0.75**
DistilBERT			0.60	1.00	0.00	**0.75**		0.60	1.00	0.00	**0.75**

Table 1: Classification performance for Depression (left) and Anxiety (right), reporting max, average, and standard deviation across classifiers. Max score selection were based on F1, ties broken by Accuracy. Abbreviated column names are [Class]ifier, [Acc]uracy, [Sens]itivity, [Spec]ificity. (Full tables in Appendix B).

- *LBF+W2V:* 8 linguistic features were selected including Full Length, Idea Density, Type-Token ratio, Open/Closed-Class ratios, proportion of Adposition, Auxiliaries, Nouns, and Sentence-level Sentiment variability, as well as 28 w2v dimensions (19 overlapping with the w2v-only model).

3.2 Model performance

Performance metrics are summarized in Table 1 by the classifier with the Max F1 score, average of all classifiers, and standard deviation. Full results per classifier are presented in Appendix B.

Depression: Both the best and the average classifier perform comparably across models, meaning that a model based on interpretable (but automatically derivable) linguistic features (LBF) performs as well as the less (directly) interpretable W2V and BERT-based methods. The best-performing classifier for LBF (a decision tree) reaches its top score through a combination of a lower Sensitivity and a higher Specificity than the W2V and combined LBF+W2V model, meaning that it detects proportionally fewer True Positives, while also allowing fewer False Positives (i.e., displaying an overall lower Positive rate). However, the Sensitivity and Specificity rates were more even for the average classifier across the models, suggesting this may have been an outlier. The BERT models are characterized by high Sensitivity scores paired with low Specificity scores, which in turn suggests a high Positives (both False and True) rate.

Anxiety: Here, the LBF model outperformed W2V and the combined model on the max and average model, owing to a much higher Specificity (lower False Positive rate). The average LBF model performs comparable to the best BERT-based models, though again on grounds of LBF's high Specificity compared to the BERT-based models high Sensitivity.

4 Discussion

Our results show three consistent themes across depression and anxiety classification. First, linguistic features (LBF) perform competitively with distributional and transformer-based representations while remaining substantially more interpretable. Second, the Word2Vec approach is comparable to LBF for depression but falls behind for anxiety, suggesting that Part-of-Speech mix, Pronoun balance, Idea Density, and Sentiment Variability capture anxiety-relevant information that may be under-weighed by distributional vectors. Third, in the present configuration of feature extraction transformers, models achieve reasonable sensitivity, but specificity is weak and effectively over-diagnoses non-targets as depressed or anxious.

For depression, feature profiles related to utterance length and volume (Total Tokens, MLU), function-words (Adpositions/Auxiliaries), Propernoun use, and Sentiment Variability. Anxiety leaned more on lexical diversity and information density (Type–Token Ratio, Idea Density), together with POS (Verb/Adverb/Proper Noun) proportions, and Sentence-level Sentiment.

We explored the W2V model by studying top words associated with six FeatureWiz-selected dimensions (See Appendix C for detail). This exploration suggested that, for anxiety, predictive W2V dimensions reflected fluency markers (e.g., "umm"), and knowledge-based words ("what," "know") that may reflect internal questioning. Furthermore, medical and identity terms appeared across conditions, potentially suggesting disease salience in daily life.

In a paediatric rheumatology context where low-burden, interpretable screening is crucial, language is a sensitive, clinically meaningful window on brain involvement and mental health in cSLE, suitable for scalable, stepped-care screening.

5 Limitations and Future Directions

While common in rare-disease paediatric research, $N = 20$ limits the statistical power of the results and the ability to train deep learning models fairly. Further, all data comes from a single center, which may limit the linguistic generalization to other English-speaking regions or socioeconomic backgrounds. The study would benefit from an age/sex-matched healthy control group, as the current design makes it difficult to determine if the linguistic markers are specific to Lupus related outcomes or general adolescent development.

One approach future work may use to strengthen generalizability is to leverage transfer learning from broader paediatric depression datasets. Secondly, the inclusion of an age/sex-matched healthy control group could tease apart typical adolescent development from disease progression. Notably, the authors have received research ethics approval to expand to a control cohort. Further lines of research could expand and explore other autoimmune conditions.

Since the data are extracted from spoken interviews, adding acoustic features (pitch, jitter, shimmer, pauses) could capture further input than text alone. However, the strength of the current lexical based work is its feasibility in low-resource or remote clinical settings where high-quality recording equipment and controlled environments may not be available. Future studies should evaluate the incremental value of acoustic signals while continuing to identify which linguistic based features generalize most consistently across tasks, sites, and populations.

6 Ethics Statement

All work was approved through the Hospital of Sick Children Research Ethics Board (REB). This work has the potential to extend beyond the lupus population to other vulnerable paediatric populations. Given the sensitivity of the data and the elevated risk in these vulnerable groups, future knowledge translation and implementation should proceed under diligent multidisciplinary oversight, including physicians and cognitive scientists.

Acknowledgments

This work is supported by the Lupus Research Alliance, the US Department of Defense, SickKids Garry Hurvitz Centre for Brain and Mental Health Outcomes, and a Canadian Institutes for Health Research - Canada Research Chair Tier 2 in Mental Health and Chronic Disease of Childhood (to Dr Knight).

References

Amita Aggarwal, Taciana A. P. Fernandes, Angela Migowa, Eve M. D. Smith, Maria Hanif, Kate Webb, and Laura B. Lewandowski. 2024. Childhood-onset systemic lupus erythematosus (csle): An international perspective. *Current Allergy and Asthma Reports*, 24(10):559–569.

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. *Preprint*, arXiv:1904.03323.

AutoViML. 2020. featurewiz. https://github.com/AutoViML/featurewiz. GitHub repository.

Michelle Barraclough, Shane McKie, Ben Parker, Alan Jackson, Philip Pemberton, Rebecca Elliott, and Ian N. Bruce. 2019. Altered cognitive function in systemic lupus erythematosus and associations with inflammation and functional and structural brain changes. *Annals of the Rheumatic Diseases*, 78(7):934–940.

Fulvia Ceccarelli, Carmelo Pirone, Concetta Mina, Alfredo Mascolo, Carlo Perricone, Laura Massaro, Francesca Romana Spinelli, Cristiano Alessandri, Guido Valesini, and Fabrizio Conti. 2019. Pragmatic language dysfunction in systemic lupus erythematosus patients: Results from a single center italian study. *PLOS ONE*, 14(11):e0224437.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Lisette Corbin, Emily Griner, Salman Seyedi, Zifan Jiang, Kailey Roberts, Mina Boazak, Ali Bahrami Rad, Gari D. Clifford, and Robert O. Cotes. 2023. A comparison of linguistic patterns between individuals with current major depressive disorder, past major depressive disorder, and controls in a virtual, psychiatric research interview. *Journal of Affective Disorders Reports*, 14:100645.

Janna N. de Boer, Sanne G. Brederoo, Alban E. Voppel, and Iris E. C. Sommer. 2020. Anomalies in language as a biomarker for schizophrenia. *Current Opinion in Psychiatry*, 33(3):212–218.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

Tala El Tal, Santiago Arciniegas, Sarah Mossad, Ibrahim Mohamed, Victoria Lishak, Stephanie Fevrier, Lawrence Ng, Paris Moaf, Joanna Law, Ashley Danguecan, Linda Hiraki, Deborah Levy, and Andrea Knight. 2022. Poor executive function correlates with increased disease damage and impaired patient-reported outcomes in youth with childhood-onset lupus: A cross-sectional study. In *ACR Convergence 2022 Abstracts*, Philadelphia, PA. Poster/Abstract.

Gloria Gagliardi and Fabio Tamburini. 2022. The automatic extraction of linguistic biomarkers as a viable solution for the early diagnosis of mental disorders. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 5234–5242, Marseille, France. European Language Resources Association.

Melisa Gumus, Danielle D. DeSouza, Mengdan Xu, Celia Fidalgo, William Simpson, and Jessica Robin. 2023. Evaluating the utility of daily speech assessments for monitoring depression symptoms. *DIGITAL HEALTH*, 9:111.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Mariia Ignashina, Paulina Bondaronek, Dan Santel, John Pestian, and Julia Ive. 2025. Llm assistance for pediatric depression. *Preprint*, arXiv:2501.17510.

Jordon Junyang Kho, Shangzheng Song, Samuel Ming Xuan Tan, Nur Hikmah Fitriyah, Matheus Calvin Lokadjaja, Jie Yin Yee, Zixu Yang, Eric Yu Hai Chen, Jimmy Lee, and Wilson Wen Bin Goh. 2025. Leveraging computational linguistics and machine learning for detection of ultra-high risk of mental health disorders in youths. *Schizophrenia*, 11(1):98.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Sierra Mendelsohn, Lina Khoja, Sofia Alfred, Jennifer He, Melanie Anderson, Denise DuBois, Zahi Touma, and Lisa Engel. 2021. Cognitive impairment in systemic lupus erythematosus is negatively related to social role participation and quality of life: A systematic review. *Lupus*, 30(10):1617–1630.

Michelle Quilter, Linda Hiraki, Andrea M. Knight, Julie Couture, Deborah Levy, Earl D. Silverman, Ashley N. Danguecan, Lawrence Ng, Daniela Dominguez, Katherine T. Cost, Kate M. Neufeld, Reva Schachter, and Daphne J. Korczak. 2021. Evaluation of self-report screening measures in the detection of depressive and anxiety disorders among children and adolescents with systemic lupus erythematosus. *Lupus*, 30(8):1327–1337.

Xinyang Ren, Hannah A. Burkhardt, Patricia A. Areán, Thomas D. Hull, and Trevor Cohen. 2023. Deep representations of first-person pronouns for prediction of depression symptom severity. *Preprint*, arXiv:2310.03232. AMIA Annual Symposium 2023.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Behrad Taghibeyglou and Frank Rudzicz. 2023. Who needs context? classical techniques for alzheimer's disease detection. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 102–107, Toronto, Canada. Association for Computational Linguistics.

Bazen Gashaw Teferra, Alice Rueda, Hilary Pang, Richard Valenzano, Reza Samavi, Sridhar Krishnan, and Venkat Bhat. 2024. Screening for depression using natural language processing: Literature review. *Interactive Journal of Medical Research*, 13:e55067.

Diana Valdés Cabrera, Tala El Tal, Ibrahim Mohamed, Santiago E. Arciniegas, Stephanie Fevrier, Justine Ledochowski, and Andrea M. Knight. 2024. Effects of systemic lupus erythematosus on the brain: a systematic review of structural mri findings and their relationships with cognitive dysfunction. *Lupus Science & Medicine*, 11(2):e001214.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and

Yuji Matsumoto. 2018. Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia. *Preprint*, arXiv:1812.06280. EMNLP 2020 System Demonstration.

Isabella Zaffino, Louise Boulard, Joanna Law, Ashley Danguecan, Asha Jeyanathan, Lawrence Ng, Sandra Williams-Reid, Kiah Reid, Angela Cortes, Eugene Cortes, Deborah M. Levy, Linda T. Hiraki, and Andrea M. Knight. 2025. Understanding contributors of resilience in youth with childhood-onset systemic lupus erythematosus through a socioecological lens: A mixed-methods study. *Arthritis Care & Research*, 77(9):1112–1124.

A Full Feature Selection Results

W2V_SelectedFeatures_Depression:

w2v_1 w2v_13 w2v_16 w2v_27 w2v_29 w2v_30 w2v_34 w2v_36 w2v_38 w2v_40 w2v_45 w2v_46 w2v_47 w2v_48 w2v_49 w2v_52 w2v_58 w2v_60 w2v_61 w2v_66 w2v_68 w2v_71 w2v_72 w2v_73 w2v_76 w2v_79 w2v_80 w2v_82 w2v_86 w2v_91 w2v_94 w2v_99

LBF_W2V_SelectedFeatures_Depression:

Full_Length perc_ADP second_pron_ratio third_pron_ratio total_tokens w2v_1 w2v_13 w2v_14 w2v_16 w2v_19 w2v_27 w2v_29 w2v_30 w2v_45 w2v_46 w2v_47 w2v_48 w2v_49 w2v_52 w2v_58 w2v_60 w2v_61 w2v_66 w2v_69 w2v_71 w2v_72 w2v_73 w2v_78 w2v_79 w2v_80 w2v_82 w2v_86 w2v_94

W2V_SelectedFeatures_Anxiety: w2v_15 w2v_16 w2v_18 w2v_2 w2v_20 w2v_21 w2v_35 w2v_39 w2v_4 w2v_40 w2v_46 w2v_47 w2v_50 w2v_51 w2v_53 w2v_55 w2v_57 w2v_60 w2v_62 w2v_71 w2v_72 w2v_73 w2v_76 w2v_77 w2v_79 w2v_80 w2v_82 w2v_84 w2v_86 w2v_88 w2v_89 w2v_9 w2v_91 w2v_92 w2v_94

LBF_W2V_SelectedFeatures_Anxiety:

Full_Length Idea_Density open_closed_ratio perc_ADP perc_AUX perc_NOUN sentiment_sd_sentence type_token_ratio w2v_13 w2v_14 w2v_15 w2v_16 w2v_18 w2v_19 w2v_21 w2v_30 w2v_39 w2v_4 w2v_46 w2v_50 w2v_51 w2v_52 w2v_56 w2v_6 w2v_60 w2v_62 w2v_66 w2v_72 w2v_73 w2v_79 w2v_86 w2v_88 w2v_9 w2v_91 w2v_94 w2v_97

B Full LOSO results for Depression and Anxiety

C Top words per Word2Vec feature

		Depression				Anxiety			
Model	**Classifier**	**Acc.**	**Sens.**	**Spec.**	F_1	**Acc.**	**Sens.**	**Spec.**	F_1
LBF									
	LR	0.40	0.58	0.13	0.54	0.80	0.75	0.88	**0.82**
	ExtraTrees	0.45	0.67	0.13	0.59	0.75	0.92	0.50	0.81
	AdaBoost	0.30	0.33	0.25	0.36	0.75	0.83	0.63	0.80
	LDA	0.50	0.58	0.38	0.58	0.75	0.75	0.75	0.78
	Linear SVC	0.45	0.75	0.00	0.62	0.70	0.83	0.50	0.77
	QDA	0.60	1.00	0.00	0.75	0.60	1.00	0.00	0.75
	XGBoost	0.60	0.75	0.38	0.69	0.70	0.75	0.63	0.75
	DT	0.70	0.75	0.63	0.75	0.65	0.67	0.63	0.70
	GNB	0.55	0.83	0.13	0.69	0.60	0.75	0.38	0.69
	Nu-SVC	0.60	0.67	0.50	0.67	0.45	0.75	0.00	0.62
W2V									
	LR	0.65	0.67	0.63	0.70	0.60	0.58	0.63	0.64
	ExtraTrees	0.60	0.75	0.38	0.69	0.75	0.92	0.50	0.81
	AdaBoost	0.50	0.58	0.38	0.58	0.70	0.83	0.50	0.77
	LDA	0.60	0.58	0.63	0.64	0.65	0.75	0.50	0.72
	Linear SVC	0.50	0.83	0.00	0.67	0.40	0.67	0.00	0.57
	QDA	0.35	0.33	0.38	0.38	0.45	0.33	0.63	0.42
	XGBoost	0.50	0.58	0.38	0.58	0.65	0.83	0.38	0.74
	DT	0.35	0.33	0.38	0.38	0.40	0.50	0.25	0.50
	GNB	0.65	0.92	0.25	0.76	0.60	0.83	0.25	0.71
	Nu-SVC	0.55	0.92	0.00	0.71	0.55	0.92	0.00	0.71
LBF + W2V									
	LR	0.60	0.58	0.63	0.64	0.55	0.58	0.50	0.61
	ExtraTrees	0.55	0.75	0.25	0.67	0.65	0.75	0.50	0.72
	AdaBoost	0.55	0.67	0.38	0.64	0.45	0.50	0.38	0.52
	LDA	0.65	0.83	0.38	**0.74**	0.50	0.50	0.50	0.55
	Linear SVC	0.50	0.83	0.00	0.67	0.45	0.75	0.00	0.62
	QDA	0.45	0.50	0.38	0.52	0.50	0.67	0.25	0.62
	XGBoost	0.55	0.58	0.50	0.61	0.50	0.58	0.38	0.58
	DT	0.45	0.42	0.50	0.48	0.30	0.33	0.25	0.36
	GNB	0.60	0.83	0.25	0.71	0.70	0.75	0.63	**0.75**
	Nu-SVC	0.50	0.83	0.00	0.67	0.50	0.83	0.00	0.67
Transformer Models									
	baseBERT1	0.55	0.92	0.00	0.71	0.55	0.92	0.00	0.71
	baseBERT2	0.45	0.75	0.00	0.62	0.60	0.92	0.13	0.73
	baseBERT3	0.50	0.58	0.38	0.58	0.65	0.75	0.50	0.72
	BioClinicalBERT	0.55	0.92	0.00	0.71	0.60	1.00	0.00	**0.75**
	BioMed-RoBERTa	0.55	0.92	0.00	0.71	0.60	1.00	0.00	**0.75**
	DistilBERT	0.60	1.00	0.00	**0.75**	0.60	1.00	0.00	**0.75**

Table 2: Classification performance for Depression (left columns) and Anxiety (right columns), across all classifiers. Abbreviated column names are [Class]ifier, [Acc]uracy, [Sens]itivity, [Spec]ificity.

Table 3: Highest and Lowest *Ranked Words for Select Dimension across Participants

Dataset	**Dim.**	**Placement**	**Word 1**	**Word 2**	**Word 3**	N **Unique Tokens**
Selected for All	16	Top	*For*	*Lupus*	*"various ethnicities"*	86
		Bottom	*What*	*Different*	*Know*	36
	46	Top	*Make*	*Canada*	*Are*	56
		Bottom	*I*	*Mental*	*Hospital*	17
Selected for Depression	29	Top	*You*	*Guess*	*Yeah*	35
		Bottom	*Mental*	*Umm*	*Were*	40
	58	Top	*Do*	*Parents*	*Hospital*	29
		Bottom	*Kind*	*Canada*	*Figure*	76
Selected for Anxiety	15	Top	*Umm*	*Important*	*Place*	51
		Bottom	*Feel*	*Yeah*	*Every*	53
	50	Top	*Good*	*Eat*	*Enjoy*	62
		Bottom	*Is*	*Ohh*	*Wanna*	35

*Top and Bottom Five words were identified for each participant in each dimension.
Counts across the 100 token then identified the top 3 words and total unique tokens shown above.