

Multilingual Lexical Feature Analysis of Spoken Language for Predicting Major Depression Symptom Severity

Anastasiia Tokareva¹, Judith Dineley¹, Zoë Firth¹, Pauline Conde¹, Faith Matcham², Sara Siddi^{3,4}, Femke Lamers⁵, Ewan Carr¹, Carolin Oetzmann¹, Daniel Leightley⁶, Yuezhou Zhang¹, Amos A. Folarin^{1,7}, Josep Maria Haro⁴, Brenda W.J.H. Penninx⁵, Raquel Bailon^{8,9}, Srinivasan Vairavan¹⁰, Til Wykes^{1,7}, Richard J.B. Dobson^{1,11}, Vaibhav A. Narayan¹², Matthew Hotopf^{1,7}, Nicholas Cummins¹, The RADAR-CNS Consortium¹³

¹. Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

². School of Psychology, University of Sussex, Falmer, UK

³. Impact and Prevention of Mental Disorders Research Group, Sant Joan de Déu Research Institut, Esplugues de Llobregat, Spain.

⁴. Mental Health Networking Biomedical Research Centre (CIBERSAM), Madrid, Spain

⁵. Department of Psychiatry, Amsterdam Public Health Research Institute and Amsterdam Neuroscience, Amsterdam University Medical Centre, Vrije Universiteit , Amsterdam, NL

⁶. School of Life Course & Population Sciences, King's College London, London, England

⁷. South London and Maudsley NHS Foundation Trust, London, UK

⁸. Aragón Institute of Engineering Research (I3A), University of Zaragoza, Zaragoza, Spain

⁹. Biomedical Research Networking Centre in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Spain

¹⁰. Janssen Research and Development LLC, Titusville, NJ, United States

¹¹. Institute of Health Informatics, University College London, London, UK

¹². Davos Alzheimer's Collaborative, Geneva, Switzerland

¹³. www.radar-cns.org

Abstract

Background: Captured between clinical appointments using mobile devices, spoken language has potential for objective, more regular assessment of symptom severity and earlier detection of relapse in major depressive disorder. However, research to date has largely been in non-clinical cross-sectional samples of written language using complex machine learning (ML) approaches with limited interpretability.

Methods: We describe an initial exploratory analysis of longitudinal speech data and PHQ-8 assessments from 5,836 recordings of 586 participants in the UK, Netherlands, and Spain, collected in the RADAR-MDD study. We sought to identify

interpretable lexical features associated with MDD symptom severity with linear mixed-effects modelling. Interpretable features and high-dimensional vector embeddings were also used to test the prediction performance of four regressor ML models.

Results: In English data, MDD symptom severity was associated with 7 features including lexical diversity measures and absolutist language. In Dutch, associations were observed with words per sentence and positive word frequency; no associations were observed in recordings collected in Spain. The predictive power of lexical features and vector embeddings was near chance level across all languages.

Limitations: Smaller samples in non-English speech and methodological choices, such as the elicitation prompt, may have also limited the effect sizes observable. A lack of NLP tools in languages other than English restricted our feature choice.

Conclusion: To understand the value of lexical markers in clinical research and practice, further research is needed in larger samples across several languages using improved protocols, and ML models that account for within- and between-individual variations in language.

1. Introduction

Major Depressive Disorder (MDD) affects over 10% of people worldwide (Bains & Abdijadid, 2023). It is associated with increased mortality risk (Pratt et al., 2016) and a substantial economic burden across Europe (Pappa et al, 2024; Vieta et al., 2021; Wijnen et al., 2023). MDD often follows a chronic, relapsing-remitting course, with individuals experiencing recurrent episodes of depression interspersed with periods of partial or full remission. This pattern underscores the need for ongoing management strategies; however, a lack of scalable, remote measurement tools hinders the ability to monitor symptoms for early indicators of relapse (Kennis et al., 2020).

In contrast, data collected using Remote Measurement Technology (RMT), like smartphones or wearables, offers the potential of frequent, objective and non-invasive approaches for monitoring chronic conditions like MDD outside of clinical visits (Matcham et al., 2022; Andrews et al., 2023). Among the different physiological and behavioural signals captured by RMT, speech is a promising

marker of MDD (Corbin et al., 2023). Depression severity has been associated with changes in both acoustic and linguistic patterns (Low et al., 2020; Koops et al., 2021). Digital speech-based tools could enhance personalised medicine, as unique speech and language patterns could be tracked to assess symptom progression (Bickman et al., 2016).

To realise the potential of digital speech tools, it is important to understand how language characteristics, such as lexical features, change with MDD symptom severity. Lexical features refer to characteristics of words used in spoken language or text, such as word frequency and vocabulary diversity. They are often used to provide measurable insights into content and style of language (Khurana et al., 2022). Lexical features offer advantages for monitoring of MDD in that they offer quantifiable and interpretable markers of underlying cognitive and emotional processes (Trifu et al., 2024). For example, lexical features such as an increase in the use of first-person pronouns, negative words, and absolutist words have all been associated with an increase in MDD severity (Savekar et al., 2022).

However, there are three main drawbacks in research undertaken into lexical features as digital markers of MDD to date. A first limitation is the use of cross-sectional, non-clinical datasets with limited sample sizes (Savekar et al., 2022). Results using such data seem promising but are likely to be underpowered and over-optimistic. A second limitation is that many works investigate written rather than spoken language. For example, many studies have relied on written sources like social media and online support forums, including Facebook and Reddit (Arachchige et al., 2021) and text messages (Tlachac et al., 2023). While benefitting from large-scale, readily available data, such resources typically lack clinical validation of participants' MDD status. Importantly, spoken language is a different form of communication from written language in its mode of production, stylistic properties and disfluency patterns, and neurocognitive correlates (Huang et al., 2001; Yule, 2022). At a high level, written language can be edited and is more deliberate (Horowitz & Samuels, 1987), and we cannot presume findings from one form of expression will hold in the other. A final limitation is that, with the increasing popularity of large language models (LLMs) and transformers, many recent studies have focused on machine learning (ML) approaches applied to high-dimensional lexical representations rather than investigating specific characteristics (Low et al.,

2020). While this shift toward high-dimensional representations can improve the predictive performances of models (Tahir et al., 2025), it comes at the expense of interpretability, making it difficult to understand which specific linguistic features are associated with depressive symptoms.

With these limitations in mind, our aim was to explore associations between lexical features extracted from spoken speech samples and depression symptom severity. We utilise the Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD) dataset (Matcham et al., 2022), which includes short speech samples taken from an observational trial of a large clinical population of individuals with a history of MDD. Previous research has been undertaken on the RADAR-MDD speech dataset. Cummins et al. (2023) focused on assessing changes in acoustic and prosodic features across the three languages. Several projects employed ML approaches to predict depression symptom severity, primarily using acoustic and prosodic information (White et al., 2025; Pérez-Toro et al., 2024; Campbell et al., 2023). Two papers have explored language changes; Pérez-Toro et al. (2024) presented a set of preliminary ML analysis utilising high-dimensional lexical representations. Zhang et al. (2024) used a deep learning model to identify depression-related topics in the English-language RADAR-MDD dataset. These previous studies used machine learning models with limited interpretability. In contrast, we explore interpretable lexical features; for example, response length, lexical diversity, and first-person pronoun use, to provide clearer insights into linguistic markers of MDD symptom severity.

Using the RADAR-MDD data, our specific objectives were to:

1. Examine if specific interpretable lexical features, extracted from remotely collected spoken language samples, are strongly associated with MDD symptom severity.
2. Assess the performance of these interpretable features when predicting MDD symptom severity.
3. Establish whether the inclusion of high-dimensional lexical feature representations improves prediction performance.

2. Methods

The key methodological steps involved in this work were collecting RADAR-MDD data, voice-to-text transcription, extracting both interpretable and high-dimensional lexical features, and analysing the relationships between interpretable features and depression and their ability to predict depression severity compared to the more complex features (Figure 1). Each step is described in detail in the subsections below.

2.1 Study Design

RADAR-MDD was an observational cohort study of individuals with a history of MDD from three recruitment sites: *King's College London* (KCL, London, United Kingdom); *Amsterdam UMC, Vrije Universiteit* (VUmc; Amsterdam, Netherlands); and *Centro de Investigación Biomédica en Red del Área Salud Mental* (CIBERSAM; Barcelona, Spain). The main eligibility criteria included a non-psychotic MDD episode according to the Diagnostic and Statistical Manual of Mental Health Disorders (DSM-5) within the two years before enrolment and having recurrent MDD (lifetime history of at least two episodes). A history of MDD was confirmed for all participants before the start of data collection; see Matcham (2019, 2022) for further details on the study protocol, inclusion, and exclusion criteria.

2.2 Ethics

The original ethics approval for collection of RADAR-MDD data in London was granted by the Camberwell St Giles Research Ethics Committee (17/LO/1154), *Fundacio Sant Joan de Deu* Clinical Research Ethics Committee (CI: PIC-128-17) in Barcelona, and from the *Medische Ethische Toetsingscommissie VUmc* (2018.012–NL63557.029.17) in Amsterdam.

2.3 Patient Involvement

The RADAR-MDD protocol was co-developed with a patient advisory board who shared their views on several user-facing aspects of the study, including the choice and frequency of survey measures, the usability of the study app, participant-facing documents, selection of optimal participation incentives, selection, and deployment of wearable devices, as well as analysis plans. The speech task and subsequent analysis were discussed specifically with the RADAR-CNS Patient Advisory Board.

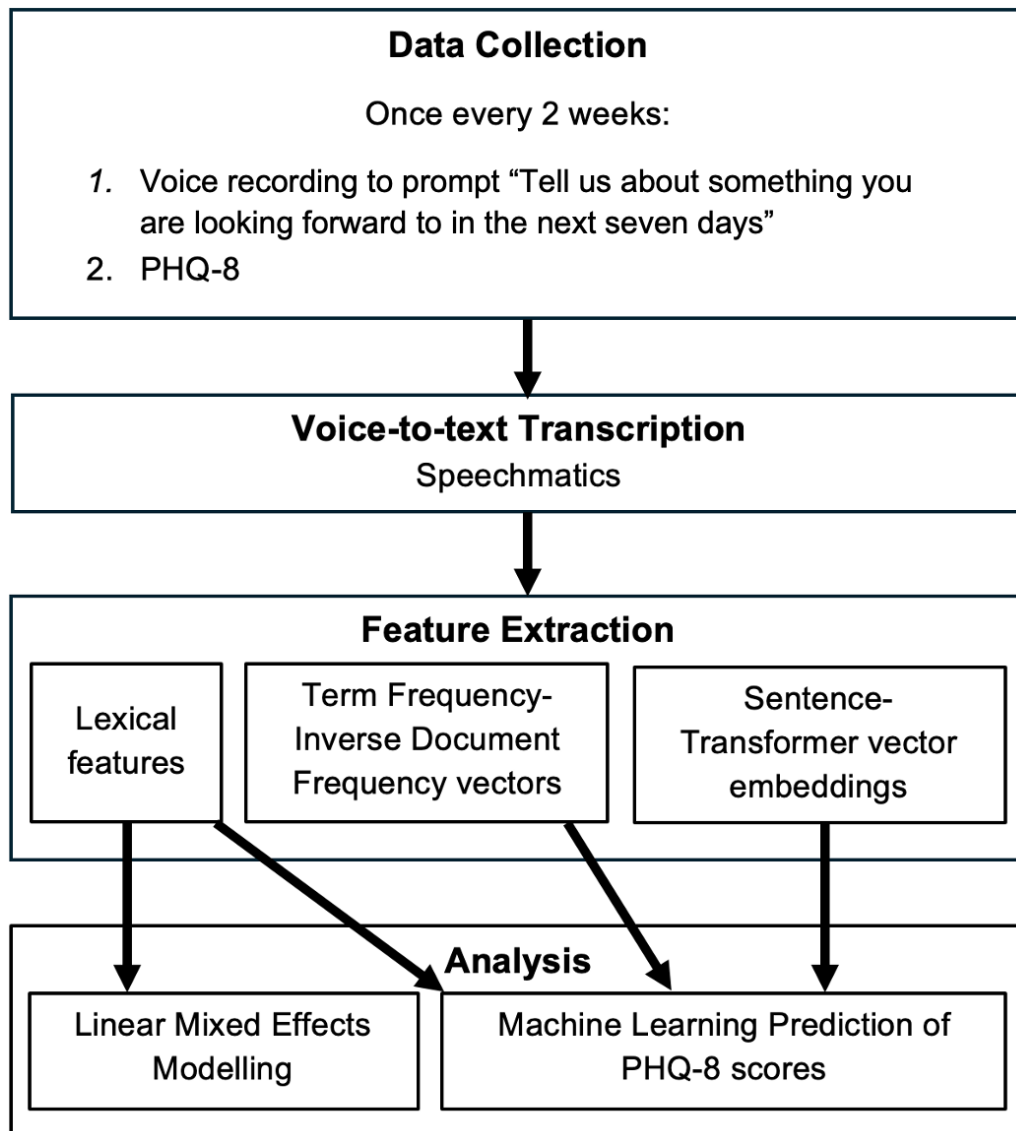


Figure 1. Overview of the key steps in the data processing pipeline.

2.4. Data Collection

Participants in the UK, Spain, and the Netherlands used an active Remote Measurement Technology (RMT) smartphone application developed for the study and installed on their Android devices (Ranjan et al., 2019). They were asked to complete speech tasks once every two weeks, beginning in August 2019 and ending in April 2021. Participants were able to re-record their responses or skip the task entirely. Speech tasks comprised a scripted task, where participants were asked to read aloud an excerpt of *The North Wind and The Sun*, and a free-response task with an open-ended cue “Tell us about something you are looking forward to in the next seven days” (Mundt et al., 2007). Only responses to the open-ended task were

analysed in this project, as they provided unscripted samples suitable for investigating lexical features.

To address safeguarding concerns and discourage participants from using the speech tasks to disclose suicidal thoughts or intentions, participants were informed that their free speech recordings would not be monitored during data collection. Upon completion, the recordings were encrypted, linked to the participant's study ID, and securely transferred to a protected server. To assess the severity of MDD symptoms, the 8-item Patient Health Questionnaire (PHQ-8) scores (0-24) (Kroenke et al., 2009) were scheduled for completion once every two weeks concurrently with the speech recordings. Participation in the speech recordings was optional, and many participants contributed multiple times, allowing us to treat this as a repeated measures design. Participation levels are described in Cummins et al. (2023). Some of the recordings were obtained during COVID lockdowns in the UK, Netherlands and Spain, defined as the period from 23 March 2020 to 11 May 2021 in work by Leightley et al. (2021). Detailed information about the speech tasks can be found in Cummins et al (2023).

2.5 Cohort Description and Bias Assessment

Sociodemographic variables including age, sex at birth, and years of education (as a proxy for reading ability and verbal IQ), collected as part of the wider study protocol have been shown to affect lexical aspects of speech (Le Dorze & Bedard, 1998) and were included in this analysis, consistent with acoustic and prosodic analysis of the RADAR-MDD data in Cummins et al (2023). Analyses of biases in the composition of analysable speech data in relation to sociodemographic and clinical variables were reported in the same work (Cummins et al., 2023).

2.6 Voice-to-Text Transcription

Recordings in English and Dutch were transcribed into text using the commercial ASR tool Speechmatics, as it has been shown to have higher accuracy than alternative open-source software such as Whisper (Russell et al., 2024). Recordings collected in Spain were spoken in either Spanish or Catalan and transcribed using Speechmatics' Spanish and Catalan ASR, respectively. Catalan transcripts were then translated into Spanish, also via Speechmatics software, enabling feature extraction via Spanish-language lexicons, to retain Catalan data in our sample for maximal statistical power.

Speech data collected in Spain is referred to as Spanish in our results and discussion for conciseness. Speechmatics transcribes disfluencies, including repetitions and filler words, that enabled more accurate measurement of response length-dependent features like word count and type-token ratio (Table 1). We chose to use automatic transcription due to resource constraints that prevented time-consuming manual transcriptions of the large RADAR-MDD dataset. Speechmatics is GDPR, SOC2 & HIPAA compliant; data security and privacy details are provided in their API documentation (Speechmatics, 2025).

2.7 Feature Extraction

We used Linguistic Inquiry and Word Count-22 (LIWC-22) software (Boyd et al., 2022) and custom Natural Language Toolkit (NLTK) 3.8.1 (Bird & Loper, 2004) code to extract interpretable lexical features from text, separately for each language (English, Dutch, Spanish). Features were selected for analysis based on a thematic review of the literature of characteristics reported to be affected by MDD. A list of lexical features, LIWC extraction categories, word examples, and expected associations between features and depression symptom severity based on previous studies is provided (Table 1). As a starting point in our exploratory work, we sought to avoid multiple hypothesis testing, as this practice is associated with an increased risk of false positive findings (Ranganathan et al., 2016).

We used two complementary measures of response length, total word count (WC) and the number of words per sentence (WPS), and two measures of lexical diversity (LD), response-length-sensitive type-token ratio (TTR) and length-insensitive Brunet's index (Table 1). Absolutist words were extracted for English; this was not possible for Dutch and Spanish due to the lack of suitable NLP tools and low feasibility of custom tool creation under the time constraints of the project. Information on a potential translation approach that we considered for absolutist word use analysis across the three languages can be found in the Supplementary Table 2 in the Appendix.

The more complex, high-dimensional features included term frequency-inverse document frequency (TF-IDF) vectors and dense vector embeddings. TF-IDF captures the importance of words in a text and is often used as a baseline for text-based ML prediction and comparison against more complex input like LLM

Table 1. Summary of lexical features and examples, LIWC categories, previous literature indicating an association between lexical features and MDD symptom severity, and expected direction of change in features with increasing MDD symptom severity. The ‘↑’ sign indicates an expected positive association between the feature and PHQ-8 scores, while ‘↓’ indicates an expected negative association between the feature and PHQ-8 scores.

Feature	Examples / Explanation	LIWC Category/ NLTK	Previous Literature	Expected Change
Negative word frequency	"Hate", "worthless", "miserable"	Affect – emotion – emo_neg	Rude et al., 2004; Weintraub et al., 2022; Hur et al., 2024	↑
Positive word frequency	"Happy", "pretty", "good"	Affect – emotion – emo_pos	Capecelatro et al., 2013; Himmelstein et al., 2018	↓
Past-focus word frequency	"Went", "did", "been"	Time orientation - focuspast	Trifu et al., 2017, Spruit et al., 2022	↑
Absolutist word frequency	"Never", "always", "completely"	Cognition - allnone	Al-Mosaiwi & Johnstone, 2018	↑
Singular First-person pronoun frequency	"I", "myself", "mine"	Function – pronoun – ppron-i	Rude et al., 2004; Edwards & Holtzman, 2017; Zimmermann et al., 2016	↑
Plural first-person pronoun frequency	"We", "us", "ours"	Function – pronoun – ppron-we	Ramirez-Esparza et al., 2008	↓
Words per sentence	5, 10, 15	WPS – words per sentence	Trifu et al., 2017	↓
Total word count	15, 20, 25	WC – word count	Trifu et al., 2017	↓
TTR	Unique words / total number of words	NLTK	Botelho et al., 2024	↓
Brunet's Index	Total text length ^ scaled power of unique words	NLTK	Botelho et al., 2024	↓

embeddings (Jalilifard et al., 2021; Lin et al., 2023). TF-IDF vectors were obtained using the scikit-learn TfidfVectorizer (Pedregosa et al., 2011) with the number of features limited to the 500 most frequent terms across the dataset to reduce the risk of overfitting. High-dimensional embeddings were extracted using a sentence-transformer model, paraphrase multilingual MPNet (masked and permuted pre-training for language understanding), available through Hugging Face (Reimers &

Gurevych, 2019). This model was selected because it is optimised for all three languages of the RADAR-MDD transcripts (English, Dutch, Spanish) (Song et al., 2020). Base model version 2 was used, producing embeddings with 768 dimensions.

Lexical features, continuous demographic variables including age and education years, and vector embeddings were pre-processed using standard scaling for normalisation ($M = 0$, $SD = 1$). In addition, a COVID variable was dummy encoded as 1 if the recording was made during the COVID lockdown and 0 if it was not. Categorical variables including sex and COVID lockdown status were retained as raw features without scaling. TF-IDF was not scaled as the scores it produced were already normalised (Robertson, 2004).

2.8 Relationships Between Lexical Features and MDD Symptom Severity

2.8.1. Mixed Effects Modelling

Separate linear mixed effects (LME) models for each language (English, Dutch, Spanish) were used to estimate the relationships between lexical features and MDD symptom severity (PHQ-8 scores) using the *lme4* package in R (Bates et al., 2015). This statistical approach is well-suited for handling missing and repeated measures data, as it can account for within-subject variability (Schober & Vetter, 2021). Each lexical feature was modelled separately as the dependent variable. Each model included fixed effects for PHQ-8 score, age, sex, years of education, and COVID lockdown status consistent Cummins et al. (2023). A random intercept for participant ID was included to account for inter-person variability in lexical features. Models were estimated using restricted maximum likelihood (REML) to obtain unbiased estimates of variance components (Harville, 1977). For each feature, estimated regression coefficients and 95% confidence intervals (CIs) were extracted and visualised using the *ggplot2* package (Wickham, 2016). Relationships between lexical features and PHQ-8 scores were considered strong (robust) where the 95% CI for the regression coefficient did not cross zero. Heatmaps with Pearson correlations between lexical features for each language were created using *ggplot2*.

2.8.2. Machine Learning Prediction

ML models were developed and evaluated in Python 3.10.9 using scikit-learn 1.4.2 (Pedregosa et al., 2011) to assess the potential of lexical features and the more complex TF-IDF vectors and vector embeddings to predict PHQ-8 scores. Four

regressor types were considered, with varying approaches to handling bias-variance trade-off (i.e., balance between underfitting and overfitting (Briscoe & Feldman, 2011)), sensitivity to noise, and model complexity (Supplementary Table 1). Each regressor included a corresponding set of hyperparameters (Supplementary Table 3), which were fine-tuned within the inner loop of nested five-fold validation to optimise model performance and improve generalisability to the outer loop.

Data were stratified by participant ID to ensure that the distribution of individuals was preserved across training and test sets in cross-validation and enable an accurate assessment of model generalisability. The models were evaluated using Root Mean Square Error (RMSE) and R^2 and were refitted using RMSE.

Each of the four ML regressors had six versions, producing a total of 24 regressor models with different types of input to systematically evaluate the effects of increasing input complexity on model performance. Feature selection using SelectKBest within each fold was set to preserve the top 100 features based on their F-statistic, with higher F-statistic indicating a stronger relationship between the feature and the target variable (PHQ-8 score). Truncated singular value decomposition (TSVD) was used for dimensionality reduction due to its suitability for sparse TF-IDF matrices (Udell et al., 2016) and was set to preserve top 100 components that capture most variance in the data. Unless stated otherwise, predictive performance values reported in the Results section refer to performance averaged across the five outer loops. Complete outer and inner loop results are provided in the Supplementary Tables 7-10.

3. Results

3.1 Dataset description (including bias assessment)

English language recordings comprised the majority of data, with 57% of participants ($N = 267$) and 68% of recordings ($N = 3,963$). Participant subgroups in all three cohorts were heavily skewed towards female participants, with women representing 78% of the English-speaking group, 70% of the Spanish sample, and 79% of the Dutch sample. Similarly, most recordings were made by female participants across the three countries. Mean PHQ-8 scores indicated that participants in Spain reported the highest average symptom severity ($M = 11.63$, $SD = 6.62$), consistent with moderate MDD. In contrast, the English and Dutch cohorts showed lower PHQ-8

Table 2. Summary of ML regressor types and corresponding input. *Notes.* ‘Significant lexical’ refers to lexical features that were shown to be robustly associated with PHQ-8 scores in mixed effects modelling, excluding the lexical features that were not strongly associated with PHQ-8 scores. ‘SelectKBest’ refers to an automatic feature selection approach, where top 100 features were preserved based on their F-statistic (higher F-statistic indicating a stronger relationship between the feature and PHQ-8 scores).

Model	Item						
	Demographics	All Lexical	Significant Lexical	Select KBest	TSVD	TF-IDF	MPNET
1.	✓	✓					
2.	✓	✓				✓	
3.	✓			✓	✓	✓	
4.	✓		✓			✓	
5.	✓						✓
6.	✓				✓		✓

scores ($M = 8.95$, $SD = 6.00$; $M = 8.46$, $SD = 5.03$, respectively), corresponding to mild depressive symptoms. The majority of recordings were made during COVID lockdown, with 68% of all English recordings (2,710/3,963), 78% of Spanish recordings (631/814), and 83% of Dutch recordings (877/1,059). Full details of the distribution of demographic and clinical variables are provided in Supplementary Table 4 and Figure 1. Additional information on the speech collection period and descriptive statistics for the whole RADAR-MDD dataset (including a scripted task not included in this project) have been provided previously (Cummins et al., 2023).

3.2 Mixed effects models

In the English-speaking dataset, mixed effects modelling indicated a robust positive association between PHQ-8 scores and TTR ($\beta = 0.012$, 95% CI [0.006, 0.018]), negative word frequency ($\beta = 0.009$, 95% CI [0.002, 0.015]), and a negative association with total word count ($\beta = -0.013$, 95% CI [-0.019, -0.007]), words per sentence ($\beta = -0.016$, 95% CI [-0.023, -0.008]), Brunet’s index ($\beta = -0.008$, 95% CI [-0.016, -0.001]), and frequency of first-person plural pronouns ($\beta = -0.009$, 95% CI [-0.016, -0.002]) and positive words ($\beta = -0.010$, 95% CI [-0.016, -0.003]) (Figure 2).

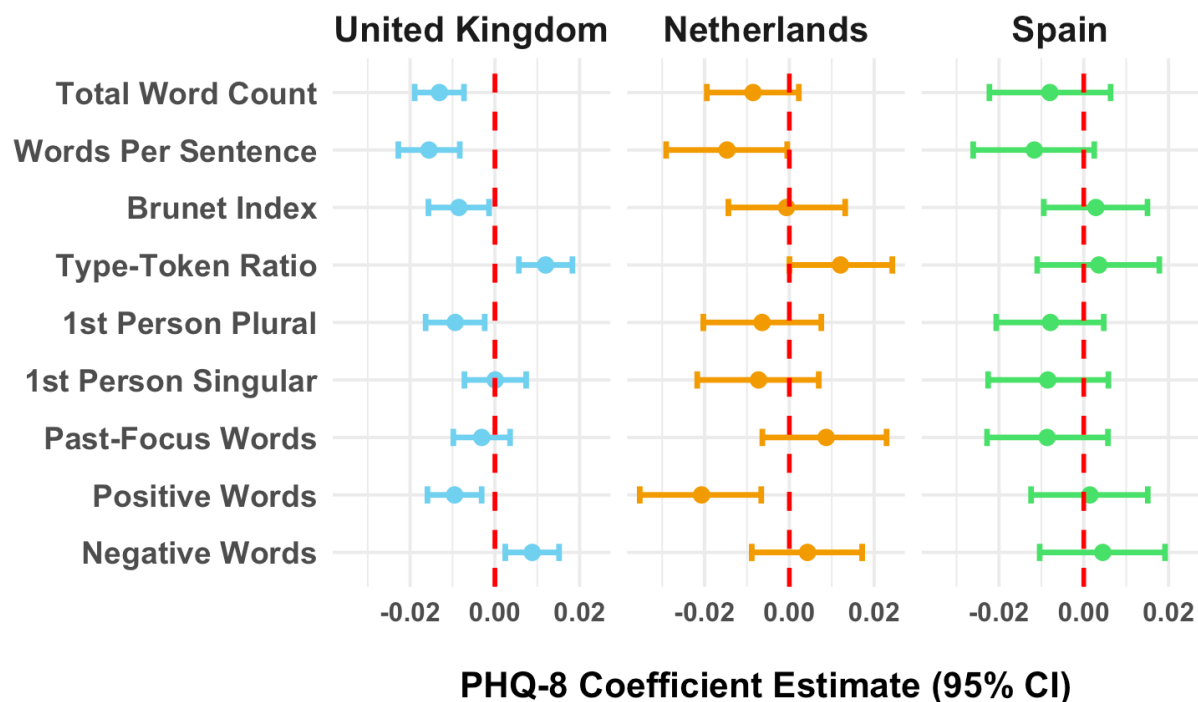


Figure 2. Association of speech features with PHQ-8 scores across the three languages (n = 467 individuals; 5863 observations). Notes. This figure shows fixed-effect estimates and 95% CIs from mixed effects modelling of the association between lexical features and PHQ-8 scores. Each point represents the standardised estimated effect of PHQ-8 scores on a lexical feature, with horizontal lines indicating 95% CIs. The vertical dashed red line represents the null effect (estimate = 0).

No strong associations with PHQ-8 scores were observed for first-person singular pronoun frequency and past-focus word frequency (Figure 2). In addition, a robust positive relationship was observed for absolutist word frequency (Supplementary Fig 2 and Table 6).

In the Dutch dataset, the number of words per sentence ($\beta = -0.015$, 95% CI [-0.029, -0.001]) and frequency of positive words ($\beta = -0.021$, 95% CI [-0.035, -0.007]) had a strong negative association with PHQ-8 scores, with the strongest relationship with positive words across the three languages (Figure 2). In the Spanish data, no robust associations between lexical features and MDD severity were observed (Figure 2).

Heatmaps for correlations between lexical features are presented (Supplementary Figs. 3-5). These demonstrated notable positive correlations (Pearson correlation coefficient >0.30) between total word count and words per sentence, and between

these two measures of response length and Brunet's index. We also observed negative correlations between TTR and Brunet's index, between these two measures of lexical diversity and TTR, and between TTR and Brunet's index across data from the three countries.

3.3. Machine learning prediction

Across all three languages and regressors, ML performance using exclusively lexical features was close to chance level, with R^2 values close to 0 and RMSE values over 4.90. The addition of TF-IDF vectors or the use of MPNet embeddings had very little influence on model performance. Overall, the highest-performing model across all languages was Elastic Net trained on vector embeddings (English: $R^2 = 0.07$, RMSE = 5.58; Dutch: $R^2 = 0.02$, RMSE = 4.94; Spanish: $R^2 = 0.01$, RMSE = 6.44). The low R^2 values, in particular, indicate that even our best-performing models are operating at near chance-level, performing no better than simple mean-based prediction.

4. Discussion

We report novel findings on the associations between depression symptom severity, as measured via remotely-collected PHQ-8 questionnaires and lexical features in spoken language of a longitudinal clinical cohort. We found changes in lexical diversity, response length, frequency of emotive words and first-person plural pronouns with increasing MDD symptom severity in English, reduced sentence length and positive word use in Dutch, but no robust associations in data collected in Spain. The RADAR-MDD dataset is unique in providing longitudinal recordings of natural speech collected using RMT from a clinical cohort recruited in the UK, Netherlands and Spain. In comparison, previous studies on lexical markers of MDD have used largely monolingual datasets drawn from written language resources, including social media and text messages (Arachchige et al., 2021; Tlachac et al., 2023).

We also explored different combinations of lexical feature representations in machine learning models to predict PHQ-8 scores. Our machine learning findings demonstrate that the models could not reliably predict depression symptom severity using lexical features; the inclusion of complex, multi-dimensional language representations did not contribute to a notable improvement in our models' prediction metrics, while reducing model interpretability. Our results also provide a

Table 3. Summary of ML performance across different regressor models and languages. Notes. ‘Significant lexical’ refers to models that only included lexical features that were robustly associated with PHQ-8 scores in mixed effects modelling. Results in bold represent performance values for the highest-performing model, Elastic Net trained on MPNet embeddings.

Regressor	Model	Language					
		English		Dutch		Spanish	
		RMSE	R ²	RMSE	R ²	RMSE	R ²
Elastic Net	All lexical	5.70	0.02	4.93	0.02	6.53	-0.02
	All lexical + TF-IDF	5.64	0.04	4.93	0.02	6.53	-0.02
	Significant lexical + TF-IDF	5.63	0.05	4.95	0.01	6.51	-0.01
	MPNet embeddings	5.58	0.07	4.94	0.02	6.44	0.01
XGBoost	All lexical	5.80	-0.01	5.01	-0.01	6.82	-0.10
	All lexical + TF-IDF	5.74	0.01	5.03	-0.02	6.87	-0.12
	Significant lexical + TF-IDF	5.74	0.01	5.03	-0.02	6.98	-0.16
	MPNet embeddings	5.63	0.00	5.03	-0.08	6.58	-0.09

unique insight into the gap between associations of lexical features with MDD severity and the weaker predictive performances of our ML models.

Our LME results characterising lexical characteristics of spoken language with increasing MDD symptom severity varied across data collected in three countries. Notably, we observed fewer strong associations in Dutch and Spanish, compared to English. These novel findings contrast with associations observed in Cummins et al (2023) between acoustic features and depression symptom severity that were largely similar across the three countries, suggesting that language and culture may have a stronger influence on lexical markers than on acoustic markers.

Cross-cultural differences have also been observed in written language of individuals with depression (Loveys et al., 2018; De Choudhury et al., 2017) and might have contributed to the differences between the countries in our project. Importantly, we observed that the Dutch and Spanish datasets produced wider CIs compared to English, reflecting greater statistical uncertainty of the estimates. This difference can be attributed to the smaller sample sizes of these groups, highlighting the importance of collecting larger spoken language datasets in non-English-speaking populations in future research (Név  l et al., 2018).

In the English language cohort, the strong negative relationships observed between PHQ-8 scores and both measures of response length, total word count and words per sentence, match previously reported results from a spoken narrative task (Trifu et al., 2017) and may reflect poverty of speech or lack of motivation observed in patients with depression (Cohen et al., 2014; Soini et al., 2024). We also observed that type-token ratio (TTR) was positively associated with MDD symptom severity, which, given previous results in the literature, was unexpected (Botelho et al., 2024). However, TTR is word count sensitive: TTR is higher when fewer words are used, creating an illusion of higher lexical diversity (Fergadiotis et al., 2015). The Brunet's index, which is length-insensitive, showed an expected negative association (Botelho et al., 2024). These differences highlight the benefit of using complementary measures when assessing the lexical diversity of speech (Lindsay et al., 2024).

Consistent with previous research, we found that individuals with greater depression symptom severity used more negative and fewer positive words, reflecting cognitive biases such as negative thinking (Rude et al., 2004). In addition, our analysis demonstrates that English-speaking individuals with more severe depressive symptoms used fewer first-person plural pronouns in their speech, which follows the social disengagement model of depression (Ramirez-Esparza et al., 2008). We could only analyse absolutist word frequency in English due to the lack of suitable lexicons for Dutch and Spanish. This feature was positively associated with MDD severity in the English-speaking population, mirroring black-and-white thinking observed in individuals with depression (Al-Mosaiwi & Johnstone, 2018).

In Dutch, the only robust associations observed included the use of shorter sentences and decreased use of positive words, similar to the English dataset. To the best of our knowledge, changes in these features with MDD symptom severity have not been previously reported in Dutch. No strong relationships were observed with Spanish data, which contrasted with the previous findings on changes in features like frequency of first-person singular and plural pronouns and negative words (Ramirez-Esparza et al., 2008).

Across all three languages, no positive relationships between MDD severity and frequency of first-person singular pronouns or past-focus words were observed, as has been previously demonstrated (Rude et al., 2004; Trifu et al., 2017). This could be attributed to our prompt phrasing ("Tell us about something you are looking

forward to in the next seven days”), as this encouraged participants to describe upcoming events rather than reflect on their current or past emotional experiences. Moreover, Spanish is a pronoun-dropping language where pronominal subjects can be omitted (Herbeck, 2021), which might have led to the lower overall use of first-person singular and plural pronouns in this cohort.

In general, machine learning regressors trained on lexical features achieved near chance-level predictive performance. These findings are novel for spoken language and contrast strongly with the previously reported accuracy of over 60% when screening text messages for moderate depression (Tlachak et al., 2023). We speculate that this difference is due to two novel qualities of this project. Firstly, we analysed spoken language from a clinical population, compared to the use of written non-clinical samples in previous research (Arachchige et al., 2021; Tlachac et al., 2023). Secondly, we used a robust nested cross-validation approach, which provides a more reliable estimate of out-of-sample performance compared to the broadly used alternative approaches (Ghasemzadeh et al., 2024). Our results suggest that the existing reports on predictive performance of lexical features in the MDD literature might be overly optimistic, reinforcing similar concerns in previous speech-based health research and highlighting the need for a wider implementation of more robust cross-validation techniques (Berisha et al., 2022).

In addition, the chance-level predictive power contrasted with the better performance during training and validation in the inner loop of all regressors (Supplementary Tables 7-10). This indicates that while models were able to learn from training data, they likely overfitted to a subgroup of participants and struggled to the patterns to the unseen test set (Aliferis & Simon, 2024). These findings reinforce the value of nested cross-validation for obtaining a realistic representation of performance and reflect high inter-person variability in lexical markers of MDD (Tackman et al., 2019), making generalisable prediction using lexical features more challenging.

Interestingly, the addition of more complex language representations, such as TF-IDF vectors and dense embeddings, did not have a notable influence on predictive performance. For example, the addition of TF-IDF vectors had very slight positive effects in some models and negative in others. In contrast, previous research on text-based mental health prediction showed that the combined use of lexical features and

TF-IDF led to a notable improvement in performance compared to lexical features alone (Yeskuatov et al., 2024). Similarly, models using dense vector embeddings showed minimal improvements in predictive power compared to regressors trained exclusively on lexical features. These observations were unexpected, as vector embeddings capture richer semantic and contextual information compared to distinct lexical features (Wolfrath et al., 2024). In addition, the inclusion of the more complex input reduced interpretability of the models, limiting transparency of the underlying decision processes involved in predicting MDD severity (Stiglic et al., 2020).

Our lower predictive performance contrasts strongly with robust associations between MDD symptom severity and several lexical features that were identified through LME modelling, especially in the English dataset. This discrepancy may arise because LMEs are well-suited for repeated measures, as they account for within-person variability (Schober & Vetter, 2021). In comparison, the ML models we applied did not model repeated measures or within-person variability; this might have limited the capture of person-specific patterns that were relevant to depression severity. In comparison, the inclusion of repeated measurements has been shown to improve predictive performance (Plate et al., 2019). Therefore, future work should investigate whether using models that account for intra-individual variability, such as mixed-effects or hierarchical approaches (Kilian et al., 2023; Hu et al., 2023, Hidalgo Julia et al., 2025), can improve the predictive performances of lexical features.

Several limitations of our project should be noted. Firstly, most participants in the RADAR-MDD dataset were middle-aged (Matcham et al., 2022), and a higher proportion were women, compared to typical clinical MDD samples (Salk et al., 2017). This might have introduced bias, as sex and gender differences have been reported in lexical aspects of spoken language (Newman et al., 2008). Secondly, a lack of NLP tools for Dutch and Spanish restricted our choice of analysable lexical features, e.g., an absolutist language lexicon was only available in English. This limitation highlights the ongoing imbalance in NLP resources and research, which remain disproportionately focused on English (Hovy & Prabhumoye, 2021). Related to this, our choice to translate Catalan into Spanish may have introduced noise into the lexical measures and embeddings. Thirdly, our Sentence-Transformer choice was

constrained to a multilingual model that supported all three languages. While this ensured compatibility, embedding quality might have been compromised by the “curse of multilinguality” (Conneau et al., 2020). Fourthly, dense vector embeddings that were used as ML input might have reflected gender and language biases present in the training data for Sentence-Transformers, such as a heavy focus on English datasets and text produced by male speakers (Hovy & Prabhumoye, 2021; Nemani et al., 2024). Finally, ASR accuracy can vary widely between audio recordings (Kuhn et al., 2023). We did not evaluate the accuracy of transcripts of this dataset, introducing uncertainty regarding transcription accuracy. In addition, we chose to use a commercial transcription tool, which typically offers higher accuracy compared to open-source alternatives. The use of commercial systems may pose cost or accessibility barriers for future replication, especially in resource-constrained settings (Russell et al., 2024).

Our analysis provides novel findings on the associations between lexical features and depression severity across languages. Effect sizes in our statistical analysis were small. The poor predictive performance of machine learning models using nested cross-fold validation also points to important generalisability challenges. However, the extracted features are specific to the protocol used, and our protocol design, including elicitation prompts, had no validated protocols to draw on (Cummins et al., 2024). Therefore, the value of remotely captured automatically extracted lexical markers in clinical research and practice cannot yet be fully understood from these initial exploratory findings.

Consequently, further research, informed by our experiences in RADAR-MDD, is needed; a broader, more open-ended prompt in combination with clinical interviews would provide a larger dataset of natural spoken responses as a primary outcome. Larger, multilingual spoken language datasets from diverse clinical populations are also required to improve the representativeness of findings. It is also crucial to address the persisting bias towards English in NLP tool development and research to promote greater generalisability and inclusivity in speech and mental health research. From an ML perspective, it would be beneficial to explore the predictive power of alternative approaches, such as hierarchical or mixed-effect models, which can account for both within- and between-person variability (Rezende et al., 2022; Kilian et al., 2023). With such advances in protocols and analytical tools, the promise

of the integration of spoken language measures into personalised approaches for predicting MDD symptom severity can be realised.

Credit authorship contribution statement

AT: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft; JD: Conceptualization, Methodology, Formal Analysis, Data Curation, Writing – Original Draft, Supervision; ZF: Methodology, Writing – Review; PC: Data Curation, Software; FM, SS and FL: Data Curation, Project administration, Writing – Review; EC: Methodology, Writing – Review; CO: Data Curation, Writing – Review; DL: Data Curation, Project Administration, Writing – Review, YZ: Writing – Review; AF: Data Curation, Software, Project Administration; JMH: Project Administration, Funding Acquisition, Writing – Review; BP: Project administration, Writing – Review; RB and SV: Project administration, Writing – Review; TW: Project Administration, Funding Acquisition, Writing – Review; VN: Project Administration, Supervision, Funding Acquisition; MH: Methodology, Project Administration, Supervision, Funding Acquisition; NC: Conceptualization, Methodology, Validation, Data Curation, Writing – Original Draft, Supervision.

Conflicts of Interest

RD is a director of CogStack Ltd and Onsientia Ltd. NC is a consultant to thymia Ltd and Noah Labs. All other authors have no other conflicts to declare.

Acknowledgements

The RADAR-CNS project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115902. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA (www.imi.europa.eu). We thank all the members of the RADAR-CNS patient advisory board for their contribution to the device selection procedures, and their invaluable advice throughout the study protocol design. This paper represents independent research part-funded by the NIHR Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Participant recruitment in Amsterdam was partially accomplished through Hersenonderzoek.nl, a Dutch online registry that facilitates participant recruitment for neuroscience studies (<https://hersenonderzoek.nl/>). Hersenonderzoek.nl is funded by ZonMw-Memorabel (project no 73305095003), a project in the context of the Dutch Deltaplan Dementie, Gieskes-Strijbis Foundation, the Alzheimer's Society in the Netherlands and Brain Foundation Netherlands. Participants in Spain were recruited through the following institutions: Parc Sanitari Sant Joan de Déu network of mental health services (Barcelona); Institut Català de la Salut primary care services (Barcelona); Institut Pere Mata-Mental Health Care (Terrassa); Hospital Clínico San Carlos (Madrid). This research was reviewed by a team with experience of mental health problems and their carers who have been specially trained to advise on research proposals and documentation through the Feasibility and Acceptability Support Team for Researchers (FAST-R): a free, confidential service in England provided by the National Institute for Health Research Maudsley Biomedical Research Centre via King's College London and South London and Maudsley NHS Foundation Trust. We thank all GLAD Study volunteers for their participation, and gratefully acknowledge the NIHR BioResource, NIHR BioResource centres, NHS Trusts and staff for their contribution. We also acknowledge NIHR BRC, King's College London, South London and Maudsley NHS Trust and King's Health Partners. We thank the National Institute for Health Research, NHS Blood and Transplant, and Health Data Research UK as part of the Digital Innovation Hub Programme.

Supplementary Material

Supplementary Table 1: Overview of the Machine Learning regressors used in the presented analysis.

Regressor	Definition
Random Forest (RF)	Ensemble model that averages the decisions made by its multiple decision trees to obtain a prediction and reduce overfitting (Breiman, 2001).
Elastic Net (EN)	Type of linear regression that combines Lasso (L1) and Ridge (L2) regularisation to reduce overfitting (Zou & Hastie, 2005).
Support Vector Regression (SVR)	Gradient boosting model that builds decision trees in a sequence (unlike the parallel trees in Random Forest Regression) and optimises each subsequent tree based on the previous prediction errors (Chen & Guestrin, 2016).
EXtreme Gradient Boosting (XGBoost)	Kernel-based type of regression that finds a line or a curve that fits the data the best within a margin while minimising prediction error (Smola & Schölkopf, 2004).

Supplementary Table 2: Summary of potential translations of 19 absolutist words defined by Al-Mosaiwi et al. (2018). Online translators used included Google Translate (<https://translate.google.co.uk/?sl=auto&tl=ru&op=translate>) and DeepL (<https://www.deepl.com/en/translator>) as per Adam-Troian et al. (2022). The decision not to proceed with manual translation and creation of a custom LIWC category for extraction was based on the lack of access to help of a professional linguist and native speakers for translation verification.

English	Spanish	Dutch
Absolutely	Absolutamente	Absoluut
All	Todo, toda, todos, todas	Alle, (allemaal)
Always	Siempre	Altijd
Complete	Completo, completa, completos, completas	Volledig
Completely	Completamente	Volledig, helemaal, geheel
Constant	Constante	Constant, constante
Constantly	Constantemente	Constant, voortdurend
Definitely	Definitivamente	Zeker
Entire	Entero, entera, enteros, enteras	Geheel
Ever	Alguna vez, nunca	Ooit, altijd
Every	Cada	Elk, elke, ieder, iedere
Everyone	Todo el mundo	Ledereen
Everything	Todo	Alles
Full	Completo, completa, ~	Volledig, volledige
Must	Include as a noun and verb conjugations?	Include as a noun and verb conjugations?
Never	Nunca, jamas, jamás	Nooit
Nothing	Nada	Niets, nul
Totally	Totalmente	Helemaal, totaal
Whole	Todo, entero, entera, enteros, enteras, ~completo, ~	Geheel, heel, volledig

Supplementary Table 3: Summary of hyper-parameters that were optimised in the inner loop of each Machine Learning regressor model. In models including embeddings (0.1, 0.5, 1) to avoid ConvergenceWarning error message.

Regressor	Model	Hyper-parameter				
		1	2	3	4	5
Random Forest (RF)	1-6	Tree total number (50, 100, 200, 500)	Max tree depth (50, 100, 200, 500)	Min samples for node split (2, 5, 10)		
Elastic Net (EN)	1-6	Alpha (0, 0.01, 0.1, 0.5, 1)	L1/L2 ratio (0.1, 0.5, 0.9).			
Support Vector (SVR)	1-6	C (0.1, 0.5, 1, 5, 10),	Kernel (linear, radial basis function)	Gamma (scale, auto)		
EXtreme Gradient Boosting (XGBoost)	1-6	Max decision tree depth (3, 5, 7, 9)	Learning rate (0.1, 0.01, 0.001)	Sub-sample (0.5, 0.7, 0.9)	L1 Alpha (0, 0.01, 0.1, 0.5, 1)	L2 Gamma (0, 0.01, 0.1, 0.5, 1)

Supplementary Table 4: A comparison of the per-country sociodemographic and baseline depression (Inventory of Depressive Symptomatology – Self Report; IDS-SR) characteristics the 467 participants were enrolled in RADAR-MDD during the speech collection period.

		United Kingdom (n = 267)	The Netherlands (n = 98)	Spain (n = 102)
Sex assigned at birth	Female	209	77	71
	Male	58	21	31
Age	Median	53	52	53
	IQR	34-63	27-59	43-60
Years in education	Median	17	17	13
	IQR	13-19	14-21	10-16
Baseline IDS-SR	Median	27	29	38
	IQR	20-38	20-37	29-52

Supplementary Table 5: The number of participants and files analysed for the scripted and free response task from each country, as well as descriptive statistics of the corresponding PHQ-8 scores.

	United Kingdom	Netherlands	Spain
# Participants	267	98	201
# Files	3,963	1,059	814
PHQ-8 Median	8	8	11
PHQ-8 IQR	4-13	5-11	6-17
PHQ-8 Range	0-24	0-24	0-24

Supplementary Table 6: Per speech task and country standardised β coefficients and 95% bootstrap CIs when using individual speech timing features to predict depression severity as determined by the 8-item Patient Health Questionnaire (PHQ-8). The model included an individual random intercept for participant ID and adjustment for age, gender, and years spent in education.

	United Kingdom		Netherlands		Spain	
	β	95% CI	β	95% CI	β	95% CI
Total Word Count	-0.013	[-0.019, -0.007]	-0.009	[-0.019, 0.002]	-0.008	[-0.022, 0.006]
Words Per Sentence	-0.016	[-0.023, -0.008]	-0.015	[-0.029, -0.001]	-0.012	[-0.026, 0.002]
Brunet's Index	-0.008	[-0.016, -0.001]	-0.001	[-0.014, 0.013]	0.003	[-0.009, 0.015]
Type-Token Ratio	0.012	[0.006, 0.018]	0.012	[0.000, 0.024]	0.004	[-0.011, 0.018]
1 st Person Plural Pronouns	-0.009	[-0.016, -0.002]	-0.006	[-0.020, 0.008]	-0.008	[-0.021, 0.005]
1 st Person Singular Pronouns	0.000	[-0.007, 0.007]	-0.007	[-0.022, 0.007]	-0.008	[-0.023, 0.006]
Absolutist Word Frequency	0.020	[0.013, 0.027]	Not extracted		Not extracted	
Past-Focus Word Frequency	-0.003	[-0.010, -0.004]	0.009	[-0.006, 0.023]	-0.009	[-0.023, 0.006]
Positive Word Frequency	-0.010	[-0.016, -0.003]	-0.021	[-0.035, -0.007]	0.001	[-0.012, 0.015]
Negative Word Frequency	0.009	[0.002, 0.015]	0.004	[-0.009, 0.017]	0.004	[-0.010, 0.019]

Supplementary Table 7: Complete table of ML performance across inner and outer loop for the English dataset. Highlighted are the results that were reported in the results section.

Regressor	Model type		Inner Loop Performance		Outer Loop Performance	
	Features	TF-IDF	RMSE	R ²	RMSE	R ²
Random Forest	All	×	3.09	0.71	5.70	0.02
	All	✓	3.16	0.70	5.97	-0.08
	SelectKBest, TSVD	✓	3.34	0.67	5.70	0.02
	Stats-based	✓	2.96	0.72	5.96	-0.08
	MPNET	×	1.95	0.89	5.67	-0.01
	MPNET, TSVD	×	2.28	0.88	6.56	-0.07
Elastic Net	All	×	5.58	0.08	5.70	0.02
	All	✓	5.34	0.16	5.64	0.04
	SelectKBest, TSVD	✓	5.39	0.14	5.65	0.04
	Stats-based	✓	5.34	0.16	5.63	0.05
	MPNet	×	5.28	0.18	5.58	0.07
	MPNet, TSVD	×	5.46	0.12	5.68	0.03
Support Vector	All	×	5.46	0.12	5.77	0.00
	All	✓	5.25	0.18	5.82	-0.02
	SelectKBest, TSVD	✓	5.47	0.11	5.73	0.01
	Stats-based	✓	5.25	0.18	5.79	-0.01
	MPNet	×	4.44	0.38	5.72	0.02
	MPNet, TSVD	×	5.02	0.25	5.71	0.02
XGBoost	All	×	5.40	0.14	5.80	-0.01
	All	✓	5.28	0.17	5.74	0.01
	SelectKBest, TSVD	✓	4.57	0.38	5.69	0.03
	Stats-based	✓	5.28	0.17	5.74	0.01
	MPNet	×	4.13	0.49	5.63	0.00
	MPNet, TSVD	×	4.36	0.44	5.74	-0.03

Supplementary Table 8: Complete table of ML performance across inner and outer loop for the Dutch dataset. Highlighted are the results that were reported in the results section.

Regressor	Model type		Inner Loop Performance		Outer Loop Performance	
	Features	TF-IDF	RMSE	R ²	RMSE	R ²
Random Forest	All	×	2.38	0.77	5.57	-0.25
	All	✓	2.96	0.65	5.31	-0.13
	SelectKBest, TSVD	✓	2.43	0.76	5.16	-0.07
	Stats-based	✓	3.05	0.63	5.26	-0.11
	MPNET	×	1.81	0.87	5.14	-0.12
	MPNET, TSVD	×	2.16	0.81	5.11	-0.11
Elastic Net	All	×	4.84	0.07	4.93	0.02
	All	✓	4.84	0.07	4.93	0.02
	SelectKBest, TSVD	✓	4.74	0.11	4.94	0.02
	Stats-based	✓	4.78	0.09	4.95	0.01
	MPNet	×	4.65	0.14	4.94	0.02
	MPNet, TSVD	×	5.81	0.22	6.68	-0.06
Support Vector	All	×	4.78	0.09	5.01	-0.01
	All	✓	4.77	0.10	5.03	-0.02
	SelectKBest, TSVD	✓	4.75	0.10	5.05	-0.03
	Stats-based	✓	4.56	0.16	5.21	-0.10
	MPNet	×	4.61	0.15	5.06	-0.03
	MPNet, TSVD	×	4.63	0.15	5.08	-0.04
XGBoost	All	×	4.91	0.04	5.01	-0.01
	All	✓	4.62	0.15	5.03	-0.02
	SelectKBest, TSVD	✓	4.24	0.28	4.99	-0.00
	Stats-based	✓	4.79	0.09	5.03	-0.02
	MPNet	×	3.65	0.46	5.03	-0.08
	MPNet, TSVD	×	3.71	0.45	5.07	-0.10

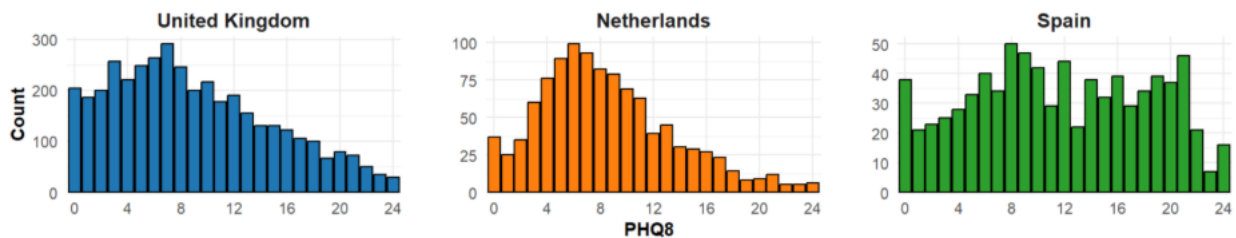
Supplementary Table 9: Complete table of ML performance across inner and outer loop for the Spanish dataset. Highlighted are the results that were reported in the results section.

Regressor	Model type		Inner Loop Performance		Outer Loop Performance	
	Features	TF-IDF	RMSE	R ²	RMSE	R ²
Random Forest	All	×	2.25	0.88	7.55	-0.37
	All	✓	2.05	0.90	7.32	-0.28
	SelectKBest, TSVD	✓	2.47	0.86	6.71	-0.07
	Stats-based	✓	2.03	0.90	7.27	-0.26
	MPNET	×	2.14	0.89	6.51	-0.07
	MPNET, TSVD	×	2.55	0.85	6.56	-0.07
Elastic Net	All	×	6.19	0.12	6.53	-0.02
	All	✓	6.19	0.12	6.53	-0.02
	SelectKBest, TSVD	✓	6.05	0.16	6.49	-0.01
	Stats-based	✓	5.85	0.21	6.51	-0.01
	MPNet	×	5.71	0.25	6.44	0.01
	MPNet, TSVD	×	5.81	0.22	6.68	-0.06
Support Vector	All	×	6.06	0.16	6.65	-0.06
	All	✓	6.02	0.17	6.76	-0.09
	SelectKBest, TSVD	✓	5.93	0.19	6.69	-0.07
	Stats-based	✓	6.08	0.15	6.72	-0.07
	MPNet	×	5.43	0.33	6.42	0.02
	MPNet, TSVD	×	5.42	0.33	6.41	0.02
XGBoost	All	×	5.17	0.36	6.82	-0.10
	All	✓	5.24	0.33	6.87	-0.12
	SelectKBest, TSVD	✓	5.07	0.39	6.65	-0.05
	Stats-based	✓	4.87	0.40	6.98	-0.16
	MPNet	×	3.45	0.71	6.58	-0.09
	MPNet, TSVD	×	3.25	0.71	6.61	-0.09

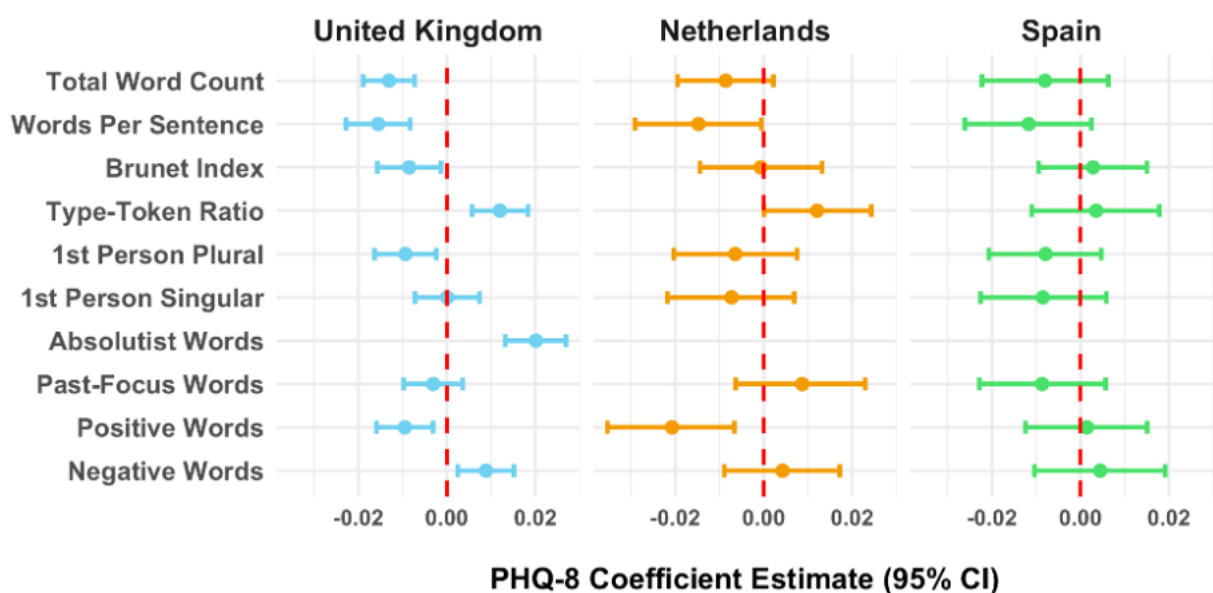
Supplementary Table 10: ML performance across inner and outer loop for the English dataset when absolutist word frequency was included as an input lexical feature.

Regressor	Model type		Inner Loop Performance		Outer Loop Performance	
	Features	TF-IDF	RMSE	R ²	RMSE	R ²
Random Forest	All	×	3.17	0.70	6.17	-0.16
	All	✓	2.67	0.77	6.01	-0.10
	SelectKBest, TSVD	✓	2.82	0.75	5.73	0.01
	Stats-based	✓	2.91	0.73	5.99	-0.09
Elastic Net	All	×	5.57	0.08	5.69	0.03
	All	✓	5.33	0.16	5.64	0.04
	SelectKBest, TSVD	✓	5.39	0.14	5.65	0.04
	Stats-based	✓	5.34	0.16	5.62	0.05
Support Vector	All	×	5.50	0.10	5.77	-0.00
	All	✓	5.20	0.19	5.82	-0.02
	SelectKBest, TSVD	✓	5.46	0.12	5.72	0.02
	Stats-based	✓	5.15	0.21	5.81	-0.02
XGBoost	All	×	5.40	0.14	5.80	-0.01
	All	✓	5.28	0.17	5.75	0.01
	SelectKBest, TSVD	✓	4.54	0.39	5.69	0.03
	Stats-based	✓	5.06	0.23	5.75	0.01

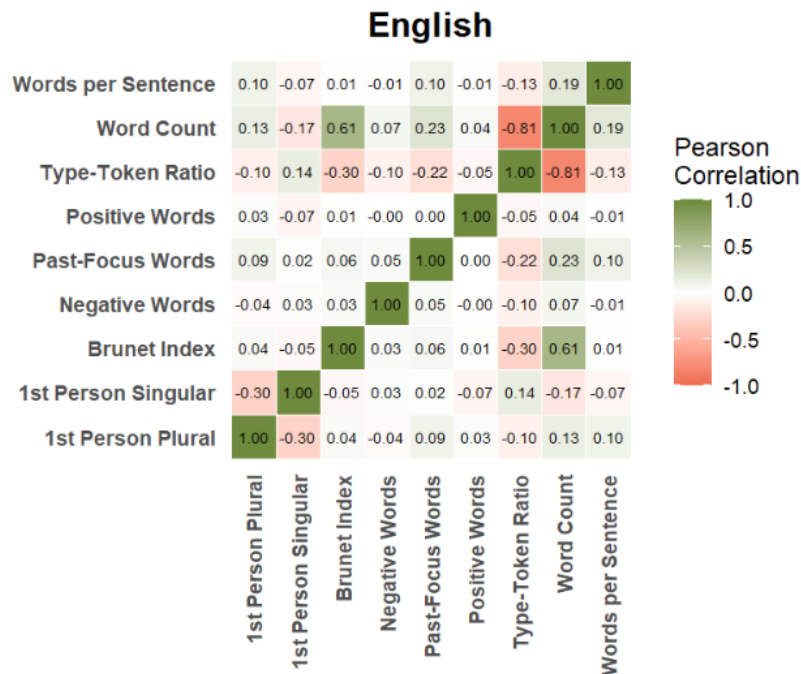
Supplementary Figure 1: A comparison of the distribution of 8-item *Patient Health Questionnaire* (PHQ-8) scores for the UK (left; $n = 3,963$), Dutch (middle; $n = 1,059$) and Spanish ($n = 814$) participants who provided analysable speech data.



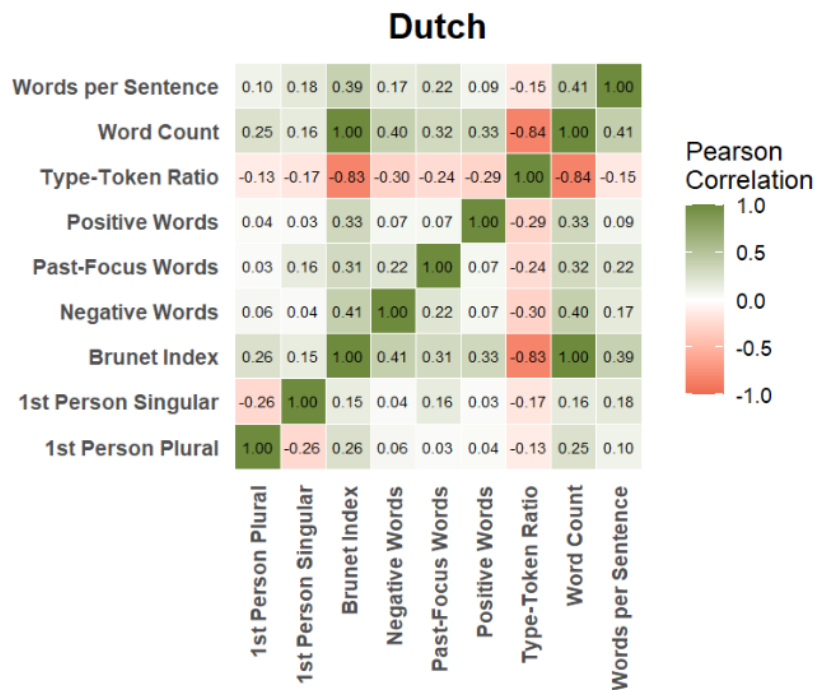
Supplementary Figure 2: Association of speech features with PHQ-8 score across the three languages including absolutist word frequency in the English-speaking cohort ($n = 467$ individuals; 5863 observations). *Notes.* This figure shows fixed-effect estimates and 95% CIs from mixed effects modelling of the association between lexical features and PHQ-8 scores. Each point represents the estimated effect of PHQ-8 scores on a lexical feature, with horizontal lines indicating 95% CIs. Estimates with CIs that do *not* cross zero indicate a robust association. The vertical dashed red line represents the null effect (estimate = 0).



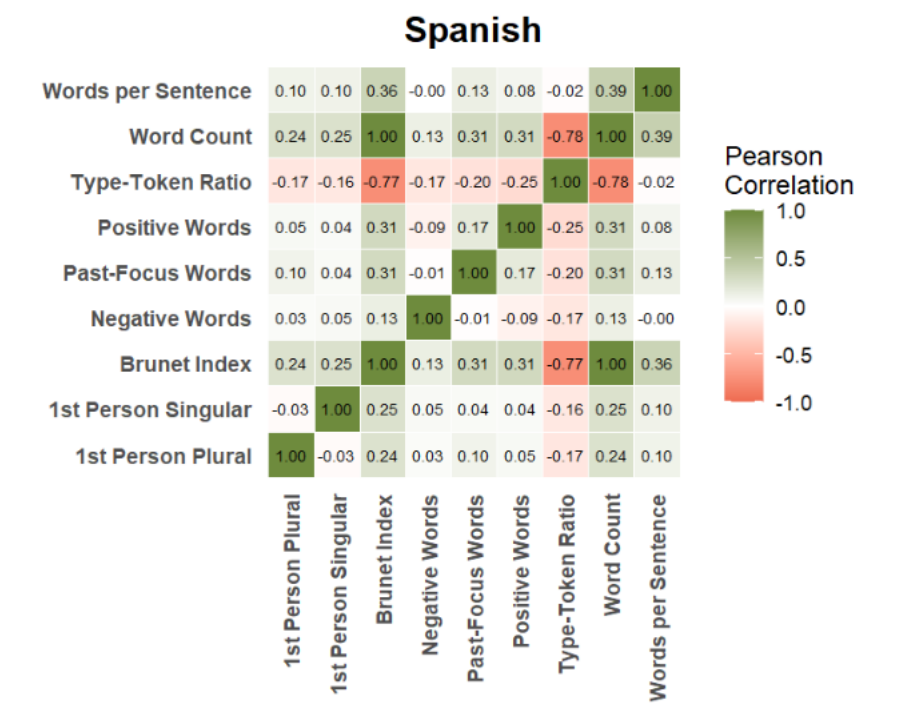
Supplementary Figure 3: Correlation Heatmap of the 9 lexical speech features extracted from recordings (n = 3,963) provided by the 267 UK participants in the RADAR-MDD cohort.



Supplementary Figure 4: Correlation Heatmap of the 9 lexical speech features extracted from recordings (n = 1,059) provided by the 98 Dutch participants in the RADAR-MDD cohort.



Supplementary Figure 5: Correlation Heatmap of the 9 lexical speech features extracted from recordings (n = 814) provided by the 201 Spanish participants in the RADAR-MDD cohort.



References

Al-Mosaiwi, M., & Johnstone, T. (2018). In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clinical Psychological Science*, 6(4), 529–542.
<https://doi.org/10.1177/2167702617747074>

Aliferis, C., & Simon, G. (2024). Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI. *Computers in Health Care (New York)*, 477–524.
https://doi.org/10.1007/978-3-031-39355-6_10

Andrews, J. L., Birrell, L., Chapman, C., Teesson, M., Newton, N., Allsop, S., McBride, N., Hides, L., Andrews, G., Olsen, N., Mewton, L., & Slade, T. (2023). Evaluating the effectiveness of a universal eHealth school-based prevention programme for depression and anxiety, and the moderating role of friendship network characteristics. *Psychological Medicine*, 53(11), 1–10.
<https://doi.org/10.1017/S0033291722002033>

Arachchige, I. A. N., Sandanapitchai, P., & Weerasinghe, R. (2021). Investigating Machine Learning & Natural Language Processing Techniques Applied for Predicting Depression Disorder from Online Support Forums: A Systematic Literature Review. *Information*, 12(11), 444.
<https://doi.org/10.3390/info12110444>

- Bains, N., & Abdijadid, S. (2023a). *Major Depressive Disorder*. PubMed; StatPearls Publishing. <https://www.ncbi.nlm.nih.gov/books/NBK559078/>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Berisha, V., Krantsevich, C., Stegmann, G., Hahn, S., & Liss, J. (2022). Are reported accuracies in the clinical speech machine learning literature overoptimistic? *Interspeech 2022*. <https://doi.org/10.21437/interspeech.2022-691>
- Bickman, L., Lyon, A. R., & Wolpert, M. (2016). Achieving Precision Mental Health through Effective Assessment, Monitoring, and Feedback Processes. *Administration and Policy in Mental Health and Mental Health Services Research*, 43(3), 271–276. <https://doi.org/10.1007/s10488-016-0718-5>
- Bird, S., & Loper, E. (2004, July 1). *NLTK: The Natural Language Toolkit*. ACLWeb; Association for Computational Linguistics. <https://aclanthology.org/PO4-3031/>
- Botelho, C., Abad, A., Schultz, T., & Trancoso, I. (2024). Speech as a Biomarker for Disease Detection. *IEEE Access*, 12, 184487–184508. <https://doi.org/10.1109/access.2024.3506433>
- Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. <https://www.liwc.app>; Austin, TX: University of Texas at Austin. <https://www.liwc.app>
- Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, 118(1), 2–16. <https://doi.org/10.1016/j.cognition.2010.10.004>
- Campbell, E. L., Dineley, J., Conde, P., Matcham, F., White, K. M., Oetzmann, C., Simblett, S., Bruce, S., Folarin, A. A., Wykes, T., Srinivasan Vairavan, Richard, Docio-Fernandez, L., Garcia-Mateo, C., Narayan, V. A., Hotopf, M., & Cummins, N. (2023). *Classifying depression symptom severity: Assessment of speech representations in personalised and generalised machine learning models*. <https://doi.org/10.21437/interspeech.2023-1721>
- Capecehatro, M. R., Sacchet, M. D., Hitchcock, P. F., Miller, S. M., & Britton, W. B. (2013). Major depression duration reduces appetitive word use: An elaborated verbal recall of emotional photographs. *Journal of Psychiatric Research*, 47(6), 809–815. <https://doi.org/10.1016/j.jpsychires.2013.01.022>
- Cohen, A. S., McGovern, J. E., Dinzeo, T. J., & Covington, M. A. (2014). Speech deficits in serious mental illness: A cognitive resource issue? *Schizophrenia Research*, 160(1-3), 173–179. <https://doi.org/10.1016/j.schres.2014.10.032>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. <https://doi.org/10.18653/v1/2020.acl-main.747>
- Corbin, L., Griner, E., Seyedi, S., Jiang, Z., Roberts, K., Boazak, M., Bahrami Rad, A., Clifford, G. D., & Cotes, R. O. (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.
<https://doi.org/10.1016/j.jadr.2023.100645>
- Cummins, N., Dineley, J., Conde, P., Matcham, F., Siddi, S., Lamers, F., Carr, E., Lavelle, G., Leightley, D., White, K. M., Oetzmman, C., Campbell, E. L., Simblett, S., Bruce, S., Haro, J. M., Penninx, B. W. J. H., Ranjan, Y., Rashid, Z., Stewart, C., & Folarin, A. A. (2023). Multilingual markers of depression in remotely collected speech samples: A preliminary analysis. *Journal of Affective Disorders*, 341, 128–136. <https://doi.org/10.1016/j.jad.2023.08.097>
- Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71(71), 10–49.
<https://doi.org/10.1016/j.specom.2015.03.004>
- Cummins, N., White, L. L., Rahman, Z., Lucas, C., Pan, T., Carr, E., Matcham, F., Downs, J., Dobson, R., Quatieri, T. F., & Dineley, J. (2024). A methodological framework and exemplar protocol for the collection and analysis of repeated speech samples (Preprint). *JMIR Research Protocols*, 14.
<https://doi.org/10.2196/69431>
- De Choudhury, M., Sharma, S. S., Logar, T., Eekhout, W., & Nielsen, R. C. (2017). Gender and Cross-Cultural Differences in Social Media Disclosures of Mental Illness. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*.
<https://doi.org/10.1145/2998181.2998220>
- Department of Health. (2011). *No health without mental health A cross-government mental health outcomes strategy for people of all ages*.
https://assets.publishing.service.gov.uk/media/5a7c348ae5274a25a914129d/dh_124058.pdf
- Edwards, T., & Holtzman, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, 68, 63–68. <https://doi.org/10.1016/j.jrp.2017.02.005>
- Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric Evaluation of Lexical Diversity Indices: Assessing Length Effects. *Journal of Speech, Language, and Hearing Research*, 58(3), 840–852.
https://doi.org/10.1044/2015_jslhr-l-14-0280
- Ghasemzadeh, H., Hillman, R. E., & Mehta, D. D. (2024). Toward Generalizable Machine Learning Models in Speech, Language, and Hearing Sciences: Estimating Sample Size and Reducing Overfitting. *Journal of Speech, Language, and Hearing Research*, 67(3), 753–781.
https://doi.org/10.1044/2023_jslhr-23-00273
- Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, 72(358), 320–338.
<https://doi.org/10.1080/01621459.1977.10480998>
- Herbeck, P. (2021). Perspectival factors and pro-drop: A corpus study of speaker/addressee pronouns with creer “think/ believe” and saber “know” in spoken Spanish. *Glossa: A Journal of General Linguistics*, 6(1).
<https://doi.org/10.16995/glossa.5873>

- Hidalgo Julia, N., Lewis, R., Ferguson, C., Goldberg, S., Lau, W., Swords, C., Valdivia, G., Wilson-Mendenhall, C., Tartar, R., Picard, R., & Davidson, R. (2025). Identifying Vocal and Facial Biomarkers of Depression in Large-Scale Remote Recordings: A Multimodal Study Using Mixed-Effects Modeling. *Proc. Interspeech 2025*, 5263-5267. <https://doi.org/10.21437/Interspeech.2025-2560>
- Himmelstein, P., Barb, S., Finlayson, M. A., & Young, K. D. (2018). Linguistic analysis of the autobiographical memories of individuals with major depressive disorder. *PLOS ONE*, 13(11), e0207814. <https://doi.org/10.1371/journal.pone.0207814>
- Horowitz, R., & Samuels, S. Jay. (1987). *Comprehending Oral and Written Language*. Elsevier Science Publishing Co Inc.
- Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8). <https://doi.org/10.1111/lnc3.12432>
- Hu, S., Wang, Y., Drovandi, C., & Cao, T. (2023). Predictions of machine learning with mixed-effects in analyzing longitudinal data under model misspecification. *Statistical Methods & Applications*, 32(2), 681–711. <https://doi.org/10.1007/s10260-022-00658-x>
- Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development* (pp. 855–858). Prentice Hall PTR.
- Hur, J. K., Heffner, J., Feng, G. W., Joormann, J., & Rutledge, R. B. (2024). Language sentiment predicts changes in depressive symptoms. *Proceedings of the National Academy of Sciences*, 121(39). <https://doi.org/10.1073/pnas.2321321121>
- Jalilifard, A., Caridá, V. F., Mansano, A. F., Cristo, R. S., & da Fonseca, F. P. C. (2021). Semantic Sensitive TF-IDF to Determine Word Relevance in Documents. *Lecture Notes in Electrical Engineering*, 736, 327–337. https://doi.org/10.1007/978-981-33-6987-0_27
- Kas, M. J. H., Hyman, S., Williams, L. M., Hidalgo-Mazzei, D., Huys, Q. J. M., Hotopf, M., Cuthbert, B., Lewis, C. M., De, L. J., Lalouis, P. A., Etkin, A., Modinos, G., & Marston, H. M. (2024). Towards a consensus roadmap for a new diagnostic framework for mental disorders. *European Neuropsychopharmacology*, 90, 16–27. <https://doi.org/10.1016/j.euroneuro.2024.08.515>
- Kennis, M., Gerritsen, L., van Dalen, M., Williams, A., Cuijpers, P., & Bockting, C. (2020). Prospective biomarkers of major depressive disorder: a systematic review and meta-analysis. *Molecular Psychiatry*, 25(2), 321–338. <https://doi.org/10.1038/s41380-019-0585-z>
- Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural Language processing: State of the art, Current Trends and Challenges. *Multimedia Tools and Applications*, 82(3), 3713–3744. <https://doi.org/10.1007/s11042-022-13428-4>
- Kilian, P., Ye, S., & Kelava, A. (2023). *Mixed effects in machine learning – A flexible mixedML framework to add random effects to supervised machine learning*

- regression*. Transactions on Machine Learning Research.
<https://openreview.net/forum?id=MKZyHtmfwH&layout=2&sort=date-desc>
- Koops, S., Brederoo, S. G., de Boer, J. N., Nadema, F. G., Voppel, A. E., & Sommer, I. E. (2021). Speech as a Biomarker for Depression. *CNS & Neurological Disorders - Drug Targets*, 20(2).
<https://doi.org/10.2174/1871527320666211213125847>
- Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3), 163–173.
<https://doi.org/10.1016/j.jad.2008.06.026>
- Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2023). Measuring the Accuracy of Automatic Speech Recognition Solutions. *ACM Transactions on Accessible Computing*, 16(4). <https://doi.org/10.1145/3636513>
- Latif, S., Qadir, J., Qayyum, A., Usama, M., & Younis, S. (2021). Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art. *IEEE Reviews in Biomedical Engineering*, 14, 342–356.
<https://doi.org/10.1109/rbme.2020.3006860>
- Le Dorze, G., & Bedard, C. (1998). Effects of age and education on the lexico-semantic content of connected speech in adults. *Journal of Communication Disorders*, 31(1), 53–71. [https://doi.org/10.1016/S0021-9924\(97\)00051-8](https://doi.org/10.1016/S0021-9924(97)00051-8)
- Leightley, D., Lavelle, G., White, K. M., Sun, S., Matcham, F., Ivan, A., Oetzmman, C., Penninx, B. W. J. H., Lamers, F., Siddi, S., Haro, J. M., Myin-Germeys, I., Bruce, S., Nica, R., Wickersham, A., Annas, P., Mohr, D. C., Simblett, S., Wykes, T., & Cummins, N. (2021). Investigating the impact of COVID-19 lockdown on adults with a recent history of recurrent major depressive disorder: a multi-Centre study using remote measurement technology. *BMC Psychiatry*, 21(1). <https://doi.org/10.1186/s12888-021-03434-5>
- Leykin, Y., Dunn, L. B., & Muñoz, R. F. (2017). The effect of depression on the decision to join a clinical trial. *Journal of Consulting and Clinical Psychology*, 85(7), 751–756. <https://doi.org/10.1037/ccp0000212>
- Lin, Y.-C., Chen, S.-A., Liu, J.-J., & Lin, C.-J. (2023). *Linear Classifier: An Often-Forgotten Baseline for Text Classification*. 2, 1876–1888.
<https://aclanthology.org/2023.acl-short.160.pdf>
- Lindsay, H., Albertin, G., Schwed, L., Linz, N., & Tröger, J. (2024). Cross-Lingual Examination of Language Features and Cognitive Scores From Free Speech. *ACL Anthology*, 16–25. <https://aclanthology.org/2024.rapid-1.3/>
- Loveys, K., Torrez, J., Fine, A., Moriarty, G., & Coppersmith, G. (2018, June 1). *Cross-cultural differences in language markers of depression online*. ACLWeb; Association for Computational Linguistics.
<https://doi.org/10.18653/v1/W18-0608>
- Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1), 96–116.
<https://doi.org/10.1002/lio2.354>
- Matcham, F., Barattieri di San Pietro, C., Bulgari, V., de Girolamo, G., Dobson, R., Eriksson, H., Folarin, A. A., Haro, J. M., Kerz, M., Lamers, F., Li, Q., Manyakov, N. V., Mohr, D. C., Myin-Germeys, I., Narayan, V., BWJH, P.,

- Ranjan, Y., Rashid, Z., Rintala, A., & Siddi, S. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. *BMC Psychiatry*, 19(1). <https://doi.org/10.1186/s12888-019-2049-z>
- Matcham, F., Leightley, D., Siddi, S., Lamers, F., White, K. M., Annas, P., de Girolamo, G., Difrancesco, S., Haro, J. M., Horsfall, M., Ivan, A., Lavelle, G., Li, Q., Lombardini, F., Mohr, D. C., Narayan, V. A., Oetzmman, C., Penninx, B. W. J. H., Bruce, S., & Nica, R. (2022). Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. *BMC Psychiatry*, 22(1). <https://doi.org/10.1186/s12888-022-03753-1>
- Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., & Geralt, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *Journal of Neurolinguistics*, 20(1), 50–64.
- Nemani, P., Joel, Y. D., Vijay, P., & Liza, F. F. (2024). Gender bias in transformers: A comprehensive review of detection and mitigation strategies. *Natural Language Processing Journal*, 6, 100047. <https://doi.org/10.1016/j.nlp.2023.100047>
- Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical Natural Language Processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantics*, 9(1). <https://doi.org/10.1186/s13326-018-0179-8>
- Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, 45(3), 211–236.
- Pappa, S., Shah, M., Young, S., Anwar, T., & Ming, T. (2024). Health-care utilisation and associated economic burden in Major Depressive Disorder and Treatment-Resistant Depression in a longitudinal population-based cohort study of 110.000 patients. *Journal of Psychiatric Research*. <https://doi.org/10.1016/j.jpsychires.2024.11.02>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85), 2825–2830.
- Pérez-Toro, P. A., Dineley, J., Kaczowska, A., Conde, P., Zhang, Y., Matcham, F., Siddi, S., Haro, J. M., Bruce, S., Wykes, T., Bailón, R., Vairavan, S., Dobson, R. J. B., Maier, A., Nöth, E., Orozco-Arroyave, J. R., Narayan, V. A., & Cummins, N. (2024). Longitudinal Modeling of Depression Shifts Using Speech and Language. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12021–12025. <https://doi.org/10.1109/icassp48485.2024.10447195>
- Plate, J. D. J., van de Leur, R. R., Peelen, L. P. H., Hietbrink, F., Peelen, L. M., & Eijkemans, M. J. C. (2019). Incorporating repeated measurements into prediction models in the critical care setting: a framework, systematic review and meta-analysis. *BMC Medical Research Methodology*, 19(1). <https://doi.org/10.1186/s12874-019-0847-0>

- Pratt, L. A., Druss, B. G., Manderscheid, R. W., & Walker, E. R. (2016). Excess mortality due to depression and anxiety in the United States: results from a nationally representative survey. *General Hospital Psychiatry*, 39(1), 39–45. <https://doi.org/10.1016/j.genhosppsych.2015.12.003>
- Ramirez-Esparza, N., Chung, C., Kacewicz, E., & Pennebaker, J. (2008). The Psychology of Word Use in Depression Forums in English and in Spanish: Testing Two Text Analytic Approaches. *Proceedings of the International AAAI Conference on Web and Social Media*, 2(1), 102–108. <https://doi.org/10.1609/icwsm.v2i1.18623>
- Ranganathan, P., Pramesh, C., & Buyse, M. (2016). Common pitfalls in statistical analysis: The perils of multiple testing. *Perspectives in Clinical Research*, 7(2), 106. <https://doi.org/10.4103/2229-3485.179436>
- Ranjan, Y., Rashid, Z., Stewart, C., Conde, P., Begale, M., Verbeeck, D., Boettcher, S., Dobson, R., & Folarin, A. (2019). RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices. *JMIR MHealth and UHealth*, 7(8), e11734. <https://doi.org/10.2196/11734>
- Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (pp. 3982–3992). <https://aclanthology.org/D19-1410.pdf>
- Rezende, P. M., Xavier, J. S., Ascher, D. B., Fernandes, G. R., & Pires, D. E. V. (2022). Evaluating hierarchical machine learning approaches to classify biological databases. *Briefings in Bioinformatics*, 23(4). <https://doi.org/10.1093/bib/bbac216>
- Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5), 503–520. <https://doi.org/10.1108/00220410410560582>
- Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8), 1121–1133. <https://doi.org/10.1080/02699930441000030>
- Russell, S. O., Gessinger, I., Krason, A., Vigliocco, G., & Harte, N. (2024). What automatic speech recognition can and cannot do for conversational speech transcription. *Research Methods in Applied Linguistics*, 3(3), 100163. <https://doi.org/10.1016/j.rmal.2024.100163>
- Salk, R. H., Hyde, J. S., & Abramson, L. Y. (2017). Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. *Psychological Bulletin*, 143(8), 783–822. <https://doi.org/10.1037/bul0000102>
- Savekar, A., Tarai, S., & Singh, M. (2022). Structural and functional markers of language signify the symptomatic effect of depression: A systematic literature review. *European Journal of Applied Linguistics*, 0(0). <https://doi.org/10.1515/eujal-2022-0022>
- Schober, P., & Vetter, T. R. (2021). Linear Regression in Medical Research. *Anesthesia & Analgesia*, 132(1), 108–109. <https://doi.org/10.1213/ane.0000000000005206>
- Soini, E., Rosenström, T., Määtänen, I., & Jokela, M. (2024). Physical activity and specific symptoms of depression: A pooled analysis of six cohort studies.

- Journal of Affective Disorders*, 348, 44–53.
<https://doi.org/10.1016/j.jad.2023.12.039>
- Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). *MPNet: Masked and Permuted Pre-training for Language Understanding*. Neural Information Processing Systems; Curran Associates, Inc.
<https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0cab5b7b67e-Abstract.html>
- Speechmatics. (2025). *API reference*. Speechmatics.
<https://docs.speechmatics.com/api-ref>
- Spruit, M., Verkleij, S., de Schepper, K., & Scheepers, F. (2022). Exploring Language Markers of Mental Health in Psychiatric Stories. *Applied Sciences*, 12(4), 2179.
<https://doi.org/10.3390/app12042179>
- Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5).
<https://doi.org/10.1002/widm.1379>
- Strawbridge, R., McCrone, P., Ulrichsen, A., Zahn, R., Eberhard, J., Wasserman, D., Brambilla, P., Schiena, G., Hegerl, U., Balazs, J., Caldas de Almeida, J., Antunes, A., Baltzis, S., Carli, V., Quoidbach, V., Boyer, P., & Young, A. H. (2022). Care pathways for people with major depressive disorder: A European Brain Council Value of Treatment study. *European Psychiatry*, 65(1), e36.
<https://doi.org/10.1192/j.eurpsy.2022.28>
- Tackman, A. M., Sbarra, D. A., Carey, A. L., Donnellan, M. B., Horn, A. B., Holtzman, N. S., Edwards, T. S., Pennebaker, J. W., & Mehl, M. R. (2019). Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of Personality and Social Psychology*, 116(5), 817–834.
<https://doi.org/10.1037/pspp0000187>
- Tahir, W. B., Khalid, S., Almutairi, S., Abohashrh, M., Memon, S. A., & Khan, J. (2025). Depression Detection in Social Media: A Comprehensive Review of Machine Learning and Deep Learning Techniques. *IEEE Access*, 13, 1–1.
<https://doi.org/10.1109/access.2025.3530862>
- Tlachac, M., Shrestha, A., Shah, M., Litterer, B., & Rundensteiner, E. A. (2023). Automated Construction of Lexicons to Improve Depression Screening With Text Messages. *IEEE Journal of Biomedical and Health Informatics*, 27(6), 2751–2759. <https://doi.org/10.1109/jbhi.2022.3203345>
- Trifu, R. N., Nemeş, B., Bodea-Haţegan, C., & Cozman, D. (2017). Linguistic indicators of language in major depressive disorder (MDD). An evidence based research. *Journal of Evidence-Based Psychotherapies*, 17(1), 105–128.
<https://doi.org/10.24193/jebp.2017.1.7>
- Trifu, R. N., Nemeş, B., Herta, D. C., Bodea-Hategan, C., Talaş, D. A., & Coman, H. (2024). Linguistic markers for major depressive disorder: a cross-sectional study using an automated procedure. *Frontiers in Psychology*, 15.
<https://doi.org/10.3389/fpsyg.2024.1355734>
- Udell, M., Horn, C., Zadeh, R., & Boyd, S. (2016). Generalized Low Rank Models. *Foundations and Trends® in Machine Learning*, 9(1), 1–118.
<https://doi.org/10.1561/22000000055>

- Vieta, E., Alonso, J., Pérez-Sola, V., Roca, M., Hernando, T., Sicras-Mainar, A., Sicras-Navarro, A., Herrera, B., & Gabilondo, A. (2021). Epidemiology and costs of depressive disorder in Spain: the EPICO study. *European Neuropsychopharmacology*, 50, 93–103. <https://doi.org/10.1016/j.euroneuro.2021.04.022>
- Weintraub, M. J., Posta, F., Ichinose, M. C., Arevian, A. C., & Miklowitz, D. J. (2023). Word usage in spontaneous speech as a predictor of depressive symptoms among youth at high risk for mood disorders. *Journal of Affective Disorders*, 323, 675–678. <https://doi.org/10.1016/j.jad.2022.12.047>
- White, L. L., Carr, E., Dineley, J., Botelho, C., Conde, P., Matcham, F., Oetzmam, C., Folarin, A. A., Fairs, G., Norbury, A., Gorla, S., Wykes, T., Vairavan, S., Dobson, R. J., Narayan, V. A., Hotopf, M., Abad, A., Trancoso, I., & Cummins, N. (2025). Speech reference intervals: an assessment of feasibility in depression symptom severity prediction. *Accepted for Publication in Proc. Interspeech 2025*.
- Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. In *Use R!* Springer International Publishing. <https://doi.org/10.1007/978-3-319-24277-4>
- Wijnen, B. F. M., Ten Have, M., de Graaf, R., van der Hoek, H. J., Lokkerbol, J., & Smit, F. (2023). The economic burden of mental disorders: results from the Netherlands mental health survey and incidence study-2. *The European Journal of Health Economics*, 25, 925–934. <https://doi.org/10.1007/s10198-023-01634-2>
- Yang, Y., Fairbairn, C., & Cohn, J. F. (2013). Detecting Depression Severity from Vocal Prosody. *IEEE Transactions on Affective Computing*, 4(2), 142–150. <https://doi.org/10.1109/t-affc.2012.38>
- Yeskuatov, E., Chua, S.-L., & Foo, L. K. (2024). Detecting Suicidal Ideations in Online Forums with Textual and Psycholinguistic Features. *Applied Sciences*, 14(21), 9911–9911. <https://doi.org/10.3390/app14219911>
- Yule, G. (2022). Language and the Brain. In *The Study of Language* (pp. 191–207). Chapter 12, Cambridge: Cambridge University Press.
- Zhang, Y., Folarin, A. A., Dineley, J., Conde, P., de Angel, V., Sun, S., Ranjan, Y., Rashid, Z., Stewart, C., Laiou, P., Sankesara, H., Qian, L., Matcham, F., White, K., Oetzmam, C., Lamers, F., Siddi, S., Simblett, S., Schuller, B. W., & Vairavan, S. (2024). Identifying depression-related topics in smartphone-collected free-response speech recordings using an automatic speech recognition system and a deep learning topic model. *Journal of Affective Disorders*, 355, 40–49. <https://doi.org/10.1016/j.jad.2024.03.106>
- Zimmermann, J., Brockmeyer, T., Hunn, M., Schauenburg, H., & Wolf, M. (2016). First-person Pronoun Use in Spoken Language as a Predictor of Future Depressive Symptoms: Preliminary Evidence from a Clinical Sample of Depressed Patients. *Clinical Psychology & Psychotherapy*, 24(2), 384–391. <https://doi.org/10.1002/cpp.2006>