Scalable, non-invasive depression monitoring with smartphone speech: a multimodal benchmark and topic analysis

Daniel Emden[1,3], Lara Gutfleisch[1], Julian Herpertz[1,2], Ramona Leenings[1], Rogério Blitz[1], Vincent L. Holstein[3,5,6], Janik Goltermann[3], Maike Richter[2], Astrid Chevance[11,12], Anna Fleuchaus[1], Nils R. Winter[3], Jennifer Spanagel[3], Susanne Meinert[3], Tiana Borgers[3], Kira Flinkenflügel[3], Frederike Stein[9], Nina Alexander[9], Hamidreza Jamalabadi[9], Elisabeth J. Leehr[3], Ronny Redlich[3,4,7,8], Ulrich Ebner-Priemer[13], Igor Nenadić[9], Tilo Kircher[9,10], Udo Dannlowski[3], Tim Hahn[3], and Nils Opel[*2,8]

[1]Department of Psychiatry and Psychotherapy, Jena University Hospital, Jena, Germany

[2]Department of Psychiatry and Neuroscience, Campus Benjamin Franklin, Charité–Universitätsmedizin Berlin, Berlin, Germany

[3]Institute for Translational Psychiatry, University of Münster, Münster, Germany

[4]Center for Intervention and Research on adaptive and maladaptive brain Circuits underlying mental health (C-I-R-C), Jena-Magdeburg-Halle, Germany

[5]Department of Psychiatry, Harvard Medical School, Boston, MA, USA

[6]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[7]Department of Psychology, University of Halle, Halle, Germany

[8]German Center for Mental Health (DZPG), Germany

[9]Department of Psychiatry and Psychotherapy, Philipps-University Marburg, Marburg, Germany

[10]Center for Mind, Brain and Behavior (CMBB), University of Marburg, Marburg, Germany

[11]Université Paris Cité and Université Sorbonne Paris Nord, INSERM INRAE, Centre for Research in Epidemiology and Statistics, Paris, France

[12]Centre d'Epidémiologie Clinique, AP-HP, Hôpital Hôtel Dieu, Paris, France

[13]Mental mHealth Lab, Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Karlsruhe, Germany

Abstract

Semantic patterns in weekly voice diaries provide a scalable window onto depression. We collected 3,151 weekly free-speech diaries from 284 German-speaking adults (156 controls, 128 with major depressive disorder) using the ReMAP smartphone app. Automatic speech recognition produced transcripts; lexical, prosodic and openSMILE acoustics, plus sentence embeddings from the 8-billion-parameter Qwen3-8B model, were extracted. Support-vector regression with nested cross-validation predicted Beck Depression Inventory (BDI) scores. Qwen3-8B embeddings substantially outperformed all classical features, achieving a mean absolute error of 4.56 points and $R^2 = 0.34$. The best-performing model combined embeddings with lexical and prosodic metrics (MAE = 4.44, $R^2 = 0.35$), representing only a modest improvement over embeddings alone. BERTopic applied to the embeddings revealed ten themes; depressive severity varied strongly between themes (Kruskal–Wallis H = 663.9, $p < 0.001$) with "Persistent Low Mood" and "Pain Distress" most associated with high BDI scores (Cliff's $\delta \leq 0.99$). These results show that modern language model embeddings capture the primary signal of depression severity in everyday speech, supporting privacy-preserving, non-invasive mental-health monitoring at population scale.

Depression | Digital Phenotyping | Speech Analysis | NLP | LLMs | Embeddings | Topic Modelling | Multimodal Machine Learning

Correspondence: nils.opel@charite.de

Introduction

Depression is a leading contributor to the global burden of disease; as a chronic condition often marked by recurrent episodes, its long-term observation is crucial for prevention, yet its clinical course is typically monitored through infrequent, clinic-based self-reports [1, 2]. This "snapshot" view often misses the day-to-day fluctuations that can signal relapse or treatment response. Digital phenotyping seeks to close that gap by continuously harvesting data from personal devices, turning everyday behaviour into continuous

mental-health read-outs [3, 4]. Among the many candidate signals, spoken language is unique: it is produced spontaneously, carries rich semantic content, and embeds prosodic cues that mirror affect and cognition [5, 6]. Voice diaries recorded in naturalistic (ecologically valid) contexts therefore offer a non-invasive lens on mood that can scale to the population level.

Early speech-based studies for depression focused on handcrafted acoustic markers or word-count approaches. While informative, these shallow features capture only limited facets of meaning and often fail to generalise across speakers and conditions. The paradigm shift towards transformer-based language models [7] now provides embeddings from large, pre-trained language models that compress nuanced semantics into dense vectors. These methods are increasingly being applied to mental health analysis [8, 9]. Yet few works have systematically benchmarked such embeddings against traditional lexical and acoustic pipelines on naturalistic data.

Equally important is the challenge of interpretability. Black-box predictions risk limited uptake in clinical settings unless researchers can articulate why a model outputs a high severity score [10]. Topic modelling methods such as BERTopic [11], an algorithm that automatically analyzes text to identify underlying discussion topics, bridge this gap by clustering embeddings into coherent themes whose prevalence can be related back to symptoms. This dual focus on performance plus insight is critical if digital phenotyping tools are to inform therapeutic decision-making rather than merely flag anomalies [12].

While promising, research in speech-based depression assessment has historically faced key limitations. Much of the foundational work has focused on classifying individuals with major depression versus healthy controls, often using speech elicited in controlled, laboratory settings (e.g., reading tasks or structured interviews) [13, 14]. This approach, while crucial for establishing a signal, does not capture the day-to-day symptom variability within a clinical population. More recent digital phenotyping studies have begun to leverage data from personal devices, but often remain cross-sectional or have yet to robustly model symptom severity from speech collected naturalistically over time [15, 16].

The present study directly addresses these gaps. We advance this agenda in three

ways. First, and most critically, we analyze a unique dataset of 3,151 weekly voice diaries from 284 participants with major depression, allowing us to move beyond simple case-control classification and instead model the nuanced, longitudinal fluctuations of symptom severity as they occur in a real-world clinical context. Second, we benchmark a multimodal feature set (classical lexical and basic prosodic metrics, TF-IDF, openSMILE acoustics, and Qwen3-8B sentence embeddings, a state-of-the-art 8-billion-parameter model) using a consistent support-vector regression protocol. Third, we pair the best-performing embeddings with BERTopic to expose the linguistic themes that drive prediction and to quantify their relationship with Beck Depression Inventory (BDI) scores.

Drawing on the ability of large language models to capture deep contextual information from text, we hypothesize that (i) Qwen3-8B embeddings will outperform all lexical and acoustic baselines in predicting continuous BDI scores. We further hypothesize that (ii) the resulting topic structure will reveal clinically coherent themes, contrasting expressions of internal distress with descriptions of external events, whose prevalence explains a substantial fraction of variance in concurrently measured symptom severity (BDI). By combining language representations from large-scale, billion-parameter models with unsupervised thematic analysis on a real-world corpus, our work aims to deliver a scalable and interpretable pipeline for continuous, voice-based depression monitoring.

Methods

Participants

The data for this study were drawn from the ReMAP study, a longitudinal project involving the weekly acquisition of speech samples from a cohort of both healthy controls (HC) and individuals with a lifetime or current diagnosis of Major Depressive Disorder (MDD). The full dataset consists of 3,151 weekly speech samples contributed by 284 unique participants (HC: 156; MDD: 128), collected between May 2019 and April 2025 over a mean period of 151 days per participant (median = 17 days, IQR = 0-137 days).

On average, each participant provided 11.1 samples (SD = 28.5); however, the number of contributions was highly variable, with a median of 2 samples per participant (IQR = 1-7) and a maximum of 282 samples from a single individual. The study design, recruitment procedures, and the feasibility of digital data acquisition via the ReMAP smartphone application have been detailed in previous publications [17, 18].

Briefly, participants were recruited from ongoing mental health cohort studies. The MDD group comprised both acutely depressed individuals and those with a lifetime diagnosis who were in remission at the time of participation thus including a wide range of depression symptom severity. The presence or absence of a lifetime mental disorder was confirmed using the Structured Clinical Interview for DSM-IV (SCID-IV; [19]) as part of the original cohort protocols. The primary outcome measure for this analysis, depressive symptom severity, was the temporally closest Beck Depression Inventory (BDI) score collected within a ±7 day window of each speech sample. Demographic and clinical characteristics of the sample are presented in Table 1.

Characteristic	**Healthy Controls**	**MDD**	**Total**	**Statistic**
No. of Participants, N (%)	156 (54.9%)	128 (45.1%)	284 (100.0%)	-
No. of Samples (n)	1,880	1,271	3,151	-
Age, years [Mean (SD)]	53.04 (11.73)	49.69 (11.45)	51.70 (11.73)	t = -7.31, $p < .001$
Sex, N (%)				$\chi^2 = 65.28$, $p < .001$
Female	107 (37.7%)	93 (32.7%)	200 (70.4%)	
Male	49 (17.3%)	35 (12.3%)	84 (29.6%)	
BDI Score at Time of Sample [Mean (SD)]	3.47 (4.29)	12.14 (9.36)	6.97 (8.03)	t = -30.90, $p < .001$

Table 1. Demographic and Clinical Characteristics of the Study Sample. MDD = Major Depressive Disorder; BDI = Beck Depression Inventory; SD = Standard Deviation. P-values from independent t-tests for continuous variables and Chi-squared tests for categorical variables.

Data Acquisition

Data were collected using the ReMAP application, a native smartphone app developed for both iOS and Android platforms. The data acquisition protocol has been detailed in

previous publications [18]. Following informed consent, participants installed the application on their personal smartphones, following a bring-your-own-device approach. The ReMAP app prompts users for active data collection, which includes weekly completion of the Beck Depression Inventory (BDI) and weekly voice samples. For the voice data collection central to this study, participants received randomized prompts during the day, asking them to speak freely for 1 to 3 minutes in response to the question: "How did you feel during the last week?" in German language.

Feature Extraction and Processing

To create a comprehensive feature set for modelling, we processed both the audio and the transcribed text from each speech sample.

Speech-to-Text Transcription and Preprocessing: All audio samples were first transcribed locally using the open-source OpenAI Whisper (large-v2) model to generate verbatim transcripts [20]. From this raw text, a second, normalized version was created for specific downstream tasks. This cleaning process, performed using spaCy, involved converting text to lowercase, removing stop words and punctuation, and lemmatizing tokens.

Acoustic Feature Extraction: From the raw audio files, we extracted two well-established, high-dimensional acoustic feature sets using the opensmile toolkit: the comprehensive ComParE_2016 set (6,373 features) and the psychologically-relevant eGeMAPS set (88 features).

Text-Based Feature Extraction: We extracted three distinct types of text-based features:

- **Lexical and Basic Prosodic Metrics:** A set of 18 features computed from the transcripts and audio timings, including metrics such as word count, lexical diversity, speech rate, and pause ratios. (see Supplementary Table S1 for a full list).

- **TF-IDF Features:** Term Frequency-Inverse Document Frequency features gener-

ated from n-grams (1- and 2-word phrases) in the normalized transcripts.

- **Sentence Embeddings:** State-of-the-art semantic representations were generated using the Qwen3-8B embeddings model. This open-source model was selected as its performance is comparable to leading proprietary models on public benchmarks (MTEB Leaderboard, checked 16 June 2025), while ensuring that sensitive participant data could be processed locally. To tailor the embeddings to our specific task, each verbatim transcript was prepended with the instruction: 'Generate a German embedding to analyse emotional and cognitive states from this weekly voice diary for a depression severity regression task.' This ensures the model generates representations optimized for our goal while preserving the full contextual and grammatical structure of the speech.

Predictive Modelling

The primary analysis predicted depressive-symptom severity using a pipeline of principal-component analysis (PCA) and support-vector regression (SVR) within a nested, participant-level cross-validation framework (see Supplementary Materials for details). Model performance was reported as mean absolute error (MAE) and R^2 against a mean-predicting baseline.

To contextualize the regression findings, we conducted two further validity checks. First, we examined how well each feature modality could classify participants as either HC or MDD. This procedure mirrored the main regression pipeline by substituting the SVR with a Support Vector Classifier (SVC), with full details provided in the Supplementary Materials. Performance was evaluated against a stratified DummyClassifier using Balanced Accuracy and AUC-ROC. Second, we repeated the full regression pipeline within the MDD group only to ensure that any predictive signal reflected variation in symptom severity rather than the broad HC–MDD contrast.

Topic Modelling

To gain insight into the linguistic themes associated with depression severity, we conducted an unsupervised topic analysis using BERTopic [11] on the best-performing text embeddings (Qwen3-8B). The BERTopic pipeline employed UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction and HDBSCAN for clustering. The topic modelling was performed on the lemmatized transcripts to ensure coherent keyword representations. The resulting topics were manually labelled based on an inspection of each topic's dominant keywords and a review of its most representative documents. After automatically reducing the number of topics to merge semantically similar clusters, we performed a series of statistical tests. A Kruskal-Wallis H-test was used to assess overall differences in BDI scores across topics, followed by pairwise Mann-Whitney U tests (with Benjamini-Hochberg FDR correction) to identify which specific themes differed significantly. Cliff's Delta (δ) was calculated as a non-parametric measure of effect size.

Results

Predictive Modelling

We evaluated a range of feature modalities for their ability to predict continuous BDI scores. The performance of key feature combinations is summarized in Table 2. All feature-based models substantially outperformed the dummy regressor baseline (MAE = 6.26). The model based solely on Qwen3-8B sentence embeddings significantly outperformed all other individual feature modalities, achieving a Mean Absolute Error (MAE) of 4.56 and an R^2 of 0.34. In contrast, models using classical features like TF-IDF, lexical metrics, or acoustic features performed substantially worse on their own, with the best classical approach (TF-IDF) achieving only MAE = 5.21 and $R^2 = 0.17$.

While the Qwen3-8B embeddings clearly dominated performance, we found that this powerful semantic signal could be modestly enhanced through multimodal fusion. The best-performing model combined embeddings with lexical and prosodic metrics, achieving

MAE = 4.44 and $R^2 = 0.35$, representing a relatively small improvement over embeddings alone. This suggests that while modern embeddings capture the primary semantic content driving depression prediction, basic speech timing and lexical diversity metrics (such as speech duration and pause ratios) may provide complementary information, though further research is needed to identify the specific contributing factors.

The results of the validity checks underscore the robustness and unique contribution of the sentence-embedding features (see Supplementary Materials, Tables S2 and S3 for full details). While embeddings demonstrated superior performance in the binary classification of Healthy Controls versus participants with Major Depressive Disorder (Balanced Accuracy $\approx 70\%$), their utility was most evident in the sensitivity analysis conducted exclusively within the MDD cohort. In this more challenging clinical context, the Qwen3-8B embeddings were the only feature set to demonstrate predictive power, achieving an MAE of 6.00 and an R^2 of 0.22. In contrast, all other feature modalities failed to perform better than a dummy baseline ($R^2 \leq 0$).

To address the possibility that the model's performance was primarily driven by separating the distinct distributions of HC and MDD participants (i.e., a spurious correlation), we conducted a sensitivity analysis. The entire predictive modelling pipeline was repeated using only the speech samples from the 128 participants in the MDD group. In this more challenging clinical context, the Qwen3-8B embeddings were the only feature set to demonstrate predictive power, achieving an MAE of 6.00 and an R^2 of 0.22. In contrast, all other feature modalities (lexical, acoustic, and TF-IDF) failed to perform better than a dummy baseline ($R^2 \leq 0$). This result strongly suggests that while many features can distinguish between healthy and depressed states, only the rich semantic representations from the Qwen3-8B embeddings are capable of capturing the nuanced variance in symptom severity within a clinical population.

Feature Set	MAE (SD)	R^2 (SD)
Any Feature Set (Dummy Baseline)	6.26 (0.28)	-0.08 (0.08)
Single Modality		
Lexical & Prosodic Metrics	5.81 (1.04)	-0.15 (0.21)
TF-IDF	5.21 (0.34)	0.17 (0.09)
Acoustic (eGeMAPS)	5.61 (0.77)	-0.10 (0.13)
Acoustic (ComParE)	5.58 (0.74)	-0.04 (0.14)
Qwen3-8B embeddings	4.56 (0.27)	0.34 (0.11)
Multimodal Combinations		
Lexical & Prosodic + TF-IDF	5.37 (0.75)	-0.09 (0.14)
Lexical & Prosodic + eGeMAPS	5.72 (1.02)	-0.14 (0.20)
Lexical & Prosodic + ComParE	5.64 (0.89)	-0.10 (0.18)
TF-IDF + eGeMAPS	5.74 (0.96)	-0.14 (0.19)
TF-IDF + ComParE	6.24 (0.63)	-0.09 (0.11)
Qwen3-8B + Lexical & Prosodic	**4.44 (0.20)**	**0.35 (0.10)**
Qwen3-8B + TF-IDF	4.61 (0.27)	0.33 (0.12)
Qwen3-8B + eGeMAPS	4.63 (0.27)	0.33 (0.12)
Qwen3-8B + ComParE	4.69 (0.33)	0.29 (0.11)

Table 2. Predictive Performance of Feature Sets for BDI Score Regression. All models are Support Vector Regressors with hyperparameters tuned via 5-fold nested cross-validation. MAE = Mean Absolute Error (lower is better); R^2 = Coefficient of Determination (higher is better); SD = Standard Deviation across outer folds. Best performance is highlighted in bold.

Topic Modelling

To understand the linguistic content driving the predictive performance, we conducted an exploratory topic analysis using BERTopic. This data-driven approach yielded 10 distinct, interpretable topics within the speech diaries that we labelled according to the dominant keywords and most representative documents (see Table 3). A Kruskal-Wallis H-test revealed a highly significant association between BDI scores and these discovered topics ($H = 663.92$, $p < .001$).

Post-hoc pairwise comparisons (Mann-Whitney U tests with FDR correction) confirmed that specific topics were strongly linked to symptom severity. In particular, themes of internal distress such as "Looking for Calm" and "Living with Pain" were associated with the highest BDI scores. The most pronounced difference was found between "Looking for Calm" and the daily-activity topic "School & Home Rhythm," which yielded the

largest effect size in the analysis (Cliff's $\delta = 0.99$, $p<.001$). The distribution of BDI scores for each topic is visualized in Figure 1. To visualize the relationships between these themes, we computed their pairwise semantic similarity, revealing the distinct topical structure of the diaries (Figure 2).

To quantify the predictive utility of these themes alone, we trained a separate regression model using only the topic probabilities as features. This topic-only model was able to predict BDI scores with an MAE of 5.68 ± 0.77 on 5-fold cross-validation, demonstrating that the discovered topics themselves capture a substantial and clinically meaningful portion of the variance in depression severity.

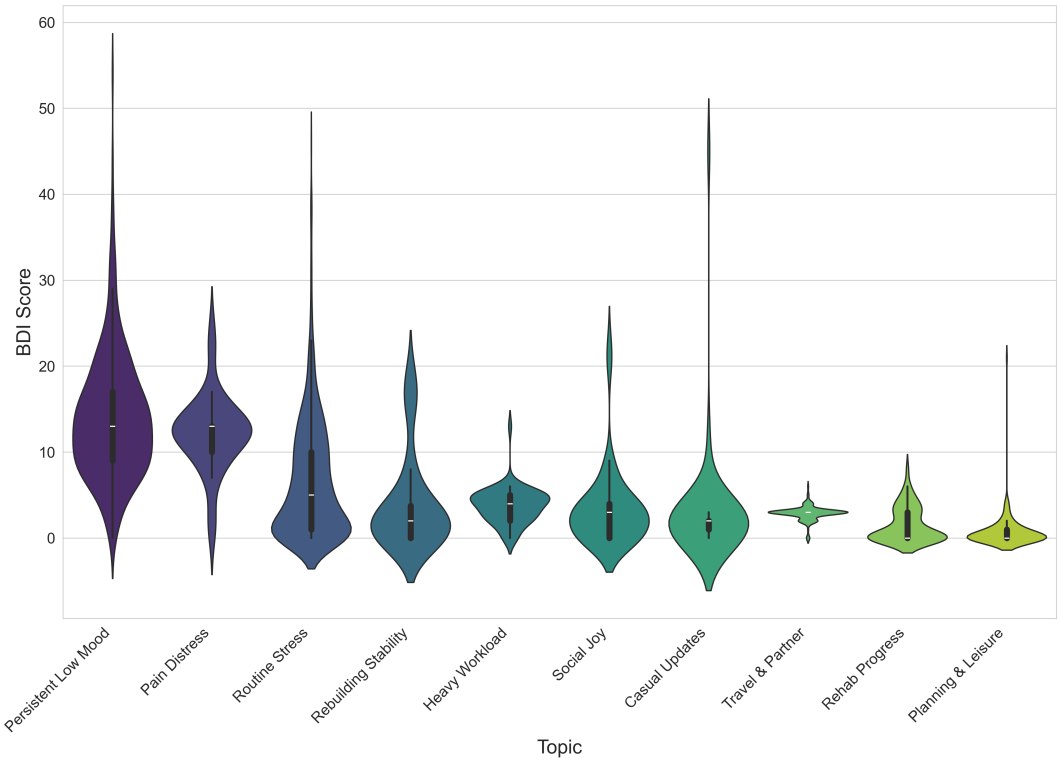

Figure 1. Distribution of Beck Depression Inventory (BDI) scores across 10 machine-learning-derived topics. The topics are arranged as violin plots along the horizontal axis in descending order of their mean BDI score. This ordering reveals a clear severity gradient. The leftmost topics, such as "Persistent Low Mood" and "Pain Distress," are visibly associated with higher BDI scores, whereas the rightmost topics, centered on daily activities like "Planning & Leisure," correspond to lower scores.

Topic Name	**Rationale (based on key words & texts)**	**Documents**	**Subjects**	**Mean BDI (SD)**
Persistent Low Mood	Mentions of chronic tiredness and ongoing therapy indicate a lingering dysphoric state conveyed with hesitant phrases like *halt* and *irgendwie*.	311	19	14.03 (7.40)
Pain Distress	*Schmerz* dominates, alongside talk of operations, health worries, and exhaustion, highlighting acute somatic discomfort driving emotional strain.	45	9	12.24 (4.55)
Routine Stress	Descriptions of ordinary work weeks, brief holidays, and household chores reflect manageable but recurring everyday pressures.	880	203	6.61 (6.95)
Rebuilding Stability	Narratives about moving, setting up a new home, and regaining strength (*Kraft aufbauen*, *stabil*) portray a constructive recovery phase.	46	3	3.98 (5.55)
Heavy Workload	Focus on budget planning, restructuring, and *Kurzarbeit* shows high task load and organizational change rather than emotional content.	57	4	3.75 (2.06)
Social Joy	Stories of birthdays, vacations, good weather, and family time emphasize pleasant social and seasonal experiences.	67	22	3.37 (4.59)
Casual Updates	Conversational check-ins with minor ailments (e.g., mild COVID colds) and everyday remarks resemble informal status updates.	47	7	2.91 (6.60)
Travel & Partner	Recurrent trips to Würzburg and time with a partner center on logistics and companionship in daily routine.	144	7	2.88 (0.80)
Rehab Progress	Detailed accounts of shoulder/knee therapy, orthoses, and physiotherapy track physical rehabilitation milestones.	67	9	1.51 (2.00)
Planning & Leisure	Entries about lesson prep, timetables, weekend walks, and family dinners highlight structured planning balanced with enjoyable downtime.	153	5	0.67 (1.92)

Table 3. Topic Summary and BDI Statistics. Topics are ordered from highest to lowest Mean BDI score. The outlier topic has been excluded from this analysis.

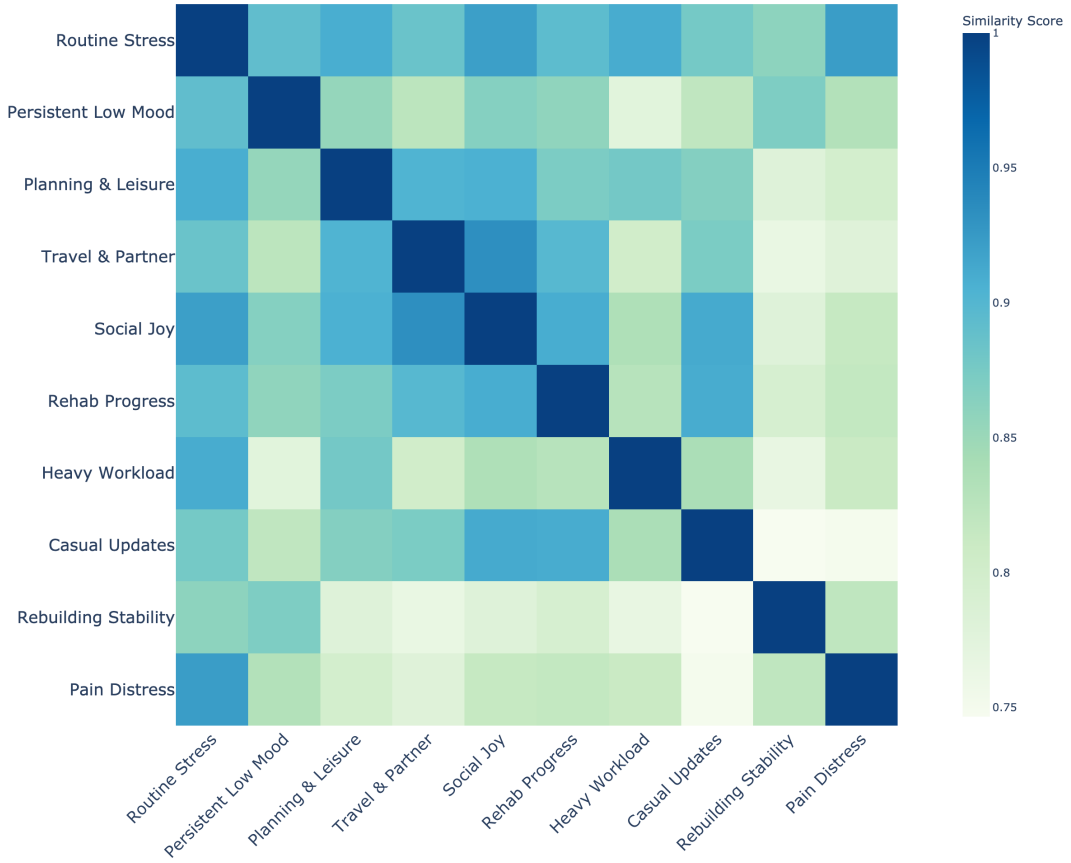

Figure 2. Pairwise cosine similarities among the ten BERTopic themes are displayed as a hierarchically ordered heat-map (Qwen3-8B embeddings; diagonal fixed at 1.0). A compact high-similarity block (deep blue) links the routine narratives Routine Stress, Social Joy and Casual Updates, while a second, looser block groups the domestic-routine themes Planning & Leisure and Travel & Partner. In contrast, the health-related topics diverge markedly: Rehab Progress shows only moderate affinity to Pain Distress, and Rebuilding Stability occupies the most isolated position in the matrix, underscoring its distinct lexical profile. The pale band separating these regions confirms the semantic distance between day-to-day event descriptions and more introspective health reflections, complementing the super-cluster pattern observed in the UMAP projection.

Discussion

This study developed and evaluated a pipeline for predicting depressive symptom severity from repeated, naturalistic voice diaries. Our findings yielded two principal insights: first, that large-scale sentence embeddings significantly outperform classical text and acoustic features for this task; and second, that an exploratory topic analysis of these embeddings can reveal clinically coherent themes that are strongly associated with depression severity.

Our primary predictive finding is the clear superiority of the Qwen3-8B sentence embeddings over all classical feature approaches. This result strongly supports the growing consensus that the semantic content of speech, as captured by large language models, is the dominant signal for detecting depressive states. While our best-performing model combined these embeddings with lexical and prosodic metrics (MAE = 4.44 vs. 4.56 for embeddings alone), this represents only a modest improvement, suggesting that embeddings capture the vast majority of the predictive signal. The specific contribution of features like speech duration and pause ratios warrants further investigation, but our findings clearly demonstrate that modern language model representations have fundamentally shifted the landscape of speech-based depression detection. Our sensitivity analysis, which evaluated the model exclusively on the MDD cohort, further strengthens this conclusion. This result strongly suggests that while many features can distinguish between healthy and depressed states, only the rich semantic representations from the Qwen3-8B embeddings are capable of capturing the nuanced variance in symptom severity within a clinical population.

A core contribution of our work is the demonstration that these high-performance models need not be "black boxes." By applying BERTopic to the embeddings, we moved beyond prediction and towards interpretation. Our data-driven approach independently uncovered topics that align closely with established clinical knowledge. The strong association between higher BDI scores and themes like "Pain Distress" and "Persistent Low Mood" suggests our model is sensitive to both somatic complaints and the language of internal coping or distress. This finding corroborates previous work showing that themes of self-focus and negative cognitive patterns are robust linguistic markers of depression [21, 22, 23].

Crucially, the ability of a separate model to predict BDI scores using only these topic probabilities confirms their clinical relevance. This dual approach—using embeddings for prediction and topic modelling for interpretation—offers a powerful model for future research. It allows for the development of accurate screening tools while also providing a pathway for clinicians to understand the potential drivers of a high predicted score,

thereby fostering trust and aiding therapeutic dialogue [24].

The present findings highlight the potential of using speech samples collected in naturalistic settings via participants' personal devices, combined with advanced natural language processing techniques, for monitoring depressive symptoms. This approach offers a scalable and ecologically valid method for digital phenotyping in mental health. Future research should investigate the generalizability of these findings across diverse populations, contexts, and technological platforms. Given the potential utility of such tools, further work is warranted that includes patient and public involvement and integrates the perspectives of caregivers and clinicians. This would help identify feasible and acceptable pathways for implementing speech-based monitoring in both research and clinical practice. Although not directly assessed in this study, it is worth considering the possibility that speech-based assessments may be perceived as less burdensome than traditional methods, such as text-based patient-reported outcomes or ecological momentary assessments. The act of recording speech can be conveniently integrated into daily activities—a concept already applied in emerging clinical decision support systems such as DAX Copilot[1]—potentially improving user engagement and adherence.

Several limitations should be acknowledged. First, our study relies on self-reported BDI scores as the primary outcome measure. While we previously demonstrated high correspondence between smartphone-derived self-reports via ReMAP and external clinical ratings of depression severity [17], the reliance on self-report remains a methodological consideration. Second, the specific demographics and clinical characteristics of the ReMAP cohort may limit the broader generalizability of our findings to other populations and healthcare contexts. Third, our analysis is fundamentally correlational in nature; while we can demonstrate strong associations between linguistic themes and BDI scores, we cannot establish causal relationships between specific language patterns and depressive symptoms. Finally, the naturalistic data collection approach, while ecologically valid, introduces variability in recording conditions and speech content that may affect model performance in more controlled clinical settings.

[1]Microsoft Dragon Copilot | Microsoft Cloud for Healthcare. Available at: `https://www.microsoft.com/en-us/health-solutions/clinical-workflow/dragon-copilot`

This study demonstrates that modern, open-source language models can predict depression severity from repeated, smartphone-collected voice diaries. We have shown that this predictive power is not opaque; it is driven by interpretable, clinically relevant linguistic themes that can be discovered directly from the data. By combining state-of-the-art language model embeddings with data-driven interpretation, this work represents a meaningful step towards developing scalable, non-invasive, and understandable tools for monitoring mental health in the real world.

Acknowledgments

This work was funded in part by the consortia grants from the German Research Foundation (DFG) FOR 2107 and SFB/TRR 393 (project grant no 521379614), as well as the DYNAMIC center, funded by the LOEWE program of the Hessian Ministry of Science and Arts (grant number: LOEWE1/16/519/03/09.001(0009)/98).

References

[1] N. Sartorius. "The economic and social burden of depression". eng. In: *The Journal of Clinical Psychiatry* 62 Suppl 15 (2001), pp. 8–11. ISSN: 0160-6689.

[2] H. König, H.-H. König, and A. Konnopka. "The excess costs of depression: a systematic review and meta-analysis". en. In: *Epidemiology and Psychiatric Sciences* 29 (2020). Publisher: Cambridge University Press (CUP). ISSN: 2045-7960, 2045-7979. DOI: 10.1017/s2045796019000180. URL: https://www.cambridge.org/core/product/identifier/S2045796019000180/type/journal_article (visited on 07/17/2025).

[3] Jennifer Melcher, Ryan Hays, and John Torous. "Digital phenotyping for mental health of college students: a clinical review". en. In: *Evidence Based Mental Health* 23.4 (Nov. 2020). Publisher: BMJ, pp. 161–166. ISSN: 1362-0347, 1468-960X. DOI: 10.1136/ebmental-2020-300180. URL: https://mentalhealth.bmj.com/lookup/doi/10.1136/ebmental-2020-300180 (visited on 07/17/2025).

[4] Pasquale Bufano et al. "Digital Phenotyping for Monitoring Mental Disorders: Systematic Review". en. In: *Journal of Medical Internet Research* 25 (Dec. 2023). Publisher: JMIR Publications Inc., e46778. ISSN: 1438-8871. DOI: 10.2196/46778. URL: https://www.jmir.org/2023/1/e46778 (visited on 07/17/2025).

[5] Rachid Riad et al. "Automated Speech Analysis for Risk Detection of Depression, Anxiety, Insomnia, and Fatigue: Algorithm Development and Validation Study". en. In: *Journal of Medical Internet Research* 26 (Oct. 2024). Publisher: JMIR Publications Inc., e58572. ISSN: 1438-8871. DOI: 10.2196/58572. URL: https://www.jmir.org/2024/1/e58572 (visited on 07/17/2025).

[6] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. "A survey on sentiment analysis methods, applications, and challenges". en. In: *Artificial Intelligence Review* 55.7 (Oct. 2022). Publisher: Springer Science and Business Media LLC, pp. 5731–5780. ISSN: 0269-2821, 1573-7462. DOI: 10.1007/s10462-022-10144-1. URL: https://link.springer.com/10.1007/s10462-022-10144-1 (visited on 07/17/2025).

[7] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". en. In: *Proceedings of the 2019 Conference of the North*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: http://aclweb.org/anthology/N19-1423 (visited on 07/17/2025).

[8] Md. Mithun Hossain et al. "Multi task opinion enhanced hybrid BERT model for mental health analysis". en. In: *Scientific Reports* 15.1 (Jan. 2025). Publisher: Springer Science and Business Media LLC. ISSN: 2045-2322. DOI: 10.1038/s41598-025-86124-6. URL: https://www.nature.com/articles/s41598-025-86124-6 (visited on 07/17/2025).

[9] Alireza Pourkeyvan, Ramin Safa, and Ali Sorourkhah. "Harnessing the Power of Hugging Face Transformers for Predicting Mental Health Disorders in Social Networks". In: *IEEE Access* 12 (2024). Publisher: Institute of Electrical and Electronics Engineers (IEEE), pp. 28025–28035. ISSN: 2169-3536. DOI: 10.1109/access.2024.3366653. URL: https://ieeexplore.ieee.org/document/10438433/ (visited on 07/17/2025).

[10] Romain Bey et al. "Natural language processing of multi-hospital electronic health records for public health surveillance of suicidality". en. In: *npj Mental Health Research* 3.1 (Feb. 2024). Publisher: Springer Science and Business Media LLC. ISSN: 2731-4251. DOI: 10.1038/s44184-023-00046-7. URL: https://www.nature.com/articles/s44184-023-00046-7 (visited on 07/17/2025).

[11] Maarten Grootendorst. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. Version Number: 1. 2022. DOI: 10.48550/ARXIV.2203.05794. URL: https://arxiv.org/abs/2203.05794 (visited on 07/17/2025).

[12] Jessica Irving et al. "Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk". en. In: *Schizophrenia Bulletin* 47.2 (Mar. 2021). Publisher: Oxford University Press (OUP), pp. 405–414. ISSN: 0586-7614, 1745-1701. DOI: 10.1093/schbul/sbaa126. URL: https://academic.oup.com/schizophreniabulletin/article/47/2/405/5918729 (visited on 07/17/2025).

[13] Nicholas Cummins et al. "A review of depression and suicide risk assessment using speech analysis". en. In: *Speech Communication* 71 (July 2015). Publisher: Elsevier BV, pp. 10–49. ISSN: 0167-6393. DOI: 10.1016/j.specom.2015.03.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167639315000369 (visited on 07/17/2025).

[14] Gratch, Jonathan et al. "The Distress Analysis Interview Corpus of human and computer interviews". In: *Proceedings of LREC 2014*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014, pp. 3123–3128.

[15] Daniel M. Low, Kate H. Bentley, and Satrajit S. Ghosh. "Automated assessment of psychiatric disorders using speech: A systematic review". en. In: *Laryngoscope Investigative Otolaryngology* 5.1 (Feb. 2020). Publisher: Wiley, pp. 96–116. ISSN: 2378-8038, 2378-8038. DOI: 10.1002/lio2.354. URL: https://onlinelibrary.wiley.com/doi/10.1002/lio2.354 (visited on 07/17/2025).

[16] John Torous et al. "New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research". en. In: *JMIR Mental Health* 3.2 (May 2016). Publisher: JMIR Publications Inc., e16. ISSN: 2368-7959. DOI: 10.2196/mental.5165. URL: http://mental.jmir.org/2016/2/e16/ (visited on 07/17/2025).

[17] Janik Goltermann et al. "Smartphone-Based Self-Reports of Depressive Symptoms Using the Remote Monitoring Application in Psychiatry (ReMAP): Interformat Validation Study". en. In: *JMIR Mental Health* 8.1 (Jan. 2021). Publisher: JMIR Publications Inc., e24333. ISSN: 2368-7959. DOI: 10.2196/24333. URL: https://mental.jmir.org/2021/1/e24333 (visited on 07/17/2025).

[18] Daniel Emden et al. "Technical feasibility and adherence of the Remote Monitoring Application in Psychiatry (ReMAP) for the assessment of affective symptoms". en. In: *Journal of Affective Disorders* 294 (Nov. 2021), pp. 652–660. ISSN: 01650327. DOI: 10.1016/j.jad.2021.07.030. URL: https://linkinghub.elsevier.com/retrieve/pii/S0165032721007096 (visited on 04/18/2025).

[19] First, Michael B. et al. *Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I)*. American Psychiatric Press, 1997.

[20] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv:2212.04356 [eess]. Dec. 2022. DOI: 10.48550/arXiv.2212.04356. URL: http://arxiv.org/abs/2212.04356 (visited on 07/17/2025).

[21] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. "Language use of depressed and depression-vulnerable college students". en. In: *Cognition & Emotion* 18.8 (Dec. 2004). Publisher: Informa UK Limited, pp. 1121–1133. ISSN: 0269-9931, 1464-0600. DOI: 10.1080/02699930441000030. URL: http://www.tandfonline.com/doi/abs/10.1080/02699930441000030 (visited on 07/17/2025).

[22] Nairan Ramirez-Esparza et al. "The Psychology of Word Use in Depression Forums in English and in Spanish: Testing Two Text Analytic Approaches". In: *Proceedings of the International AAAI Conference on Web and Social Media* 2.1 (Sept. 2021). Publisher: Association for the Advancement of Artificial Intelligence (AAAI), pp. 102–108. ISSN: 2334-0770, 2162-3449. DOI: 10.1609/icwsm.v2i1.18623. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/18623 (visited on 07/17/2025).

[23] Munmun De Choudhury et al. "Predicting Depression via Social Media". In: *Proceedings of the International AAAI Conference on Web and Social Media* 7.1 (Aug. 2021). Publisher: Association for the Advancement of Artificial Intelligence (AAAI), pp. 128–137. ISSN: 2334-0770, 2162-3449. DOI: 10.1609/icwsm.v7i1.14432. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14432 (visited on 07/17/2025).

[24] Jamin Rahman Jim et al. "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review". en. In: *Natural Language Processing Journal* 6 (Mar. 2024). Publisher: Elsevier BV, p. 100059. ISSN: 2949-7191. DOI: 10.1016/j.nlp.2024.100059. URL: https://linkinghub.elsevier.com/retrieve/pii/S2949719124000074 (visited on 07/17/2025).

[25] Sagar Jilka and Domenico Giacco. "Digital phenotyping: how it could change mental health care and why we should all keep up". en. In: *Journal of Mental Health* 33.4 (July 2024). Publisher: Informa UK Limited, pp. 439–442. ISSN: 0963-8237, 1360-0567. DOI: 10.1080/09638237.2024.2395537. URL: https://www.tandfonline.com/doi/full/10.1080/09638237.2024.2395537 (visited on 07/17/2025).

Supplementary Materials

Supplementary Methods

Metric	Description
duration_seconds	Total duration of the audio recording.
speak_seconds	Total duration of speech segments.
word_count	Total number of words spoken.
char_length	Total number of characters in the transcript.
unique_word_count	Number of unique words used.
lexical_diversity	Ratio of unique words to total words.
avg_word_length	Average length of words.
words_per_total_second	Speech rate calculated over the total recording duration.
words_per_speak_second	Speech rate calculated over speech segments only.
speech_ratio	Proportion of the recording that is speech.
pause_seconds	Total duration of pauses.
pause_ratio	Proportion of the recording that is pauses.
pause_to_speech_ratio	Ratio of pause duration to speech duration.
entity_count	Number of named entities (e.g., persons, locations) identified.
entity_density	Ratio of named entities to total words.
noun_ratio	Proportion of words that are nouns.
verb_ratio	Proportion of words that are verbs.
adj_ratio	Proportion of words that are adjectives.

Table S1. List of Lexical and Prosodic Metrics

Machine Learning Pipeline Details

All machine learning pipelines employed a nested cross-validation design for robust performance estimation and hyperparameter optimization. The outer loop used 5-fold GroupKFold cross-validation (grouped by participant ID to prevent data leakage), while the inner loop performed hyperparameter tuning using Optuna with MedianPruner for efficient optimization.

Implementation: Pipelines used scikit-learn (v1.6) with cuML (v25.02.01) for GPU-accelerated PCA and SVR. Hyperparameter optimization employed Optuna with MedianPruner. High-dimensional features (TF-IDF, acoustic, embeddings) underwent standardization → PCA → estimator; lexical features bypassed PCA.

Hyperparameters: PCA components: ['passthrough', 64-1024, 0.9-0.99 variance

ratios]; SVM kernel: ['rbf', 'poly', 'sigmoid']; C: [1e-3, 1e3]; gamma: ['scale', 'auto']; SVR epsilon: [1e-4, 1.0]; polynomial degree: [2-5]; coef0: [0.0-1.0]. Classification used balanced class weights.

Evaluation: Regression metrics: MAE, R^2 (vs. DummyRegressor mean baseline). Classification metrics: Balanced Accuracy, AUC-ROC (vs. DummyClassifier stratified baseline).

Supplementary Results

Group-Classification Analysis

To benchmark the discriminative value of each feature set, we trained binary classifiers to distinguish participants with Major Depressive Disorder (MDD) from healthy controls (HC). Results (see Table S2) showed that Qwen3-8B sentence embeddings yielded the highest performance, reaching a balanced accuracy of $\sim$ 70%. This exceeded the best classical text baseline (69%) and all acoustic-only models. The finding confirms that the semantic information captured by large-scale embeddings carries a strong, readily extractable signal of depression status and therefore provides a solid basis for the subsequent symptom-severity regression.

Feature Set	Model	Balanced Accuracy	AUC-ROC
Any Feature Set	Dummy	0.50	0.50
Lexical & Prosodic	SVC	0.62	0.66
TF-IDF	SVC	0.65	0.72
Lexical + TF-IDF	SVC	0.69	0.74
Embeddings	**SVC**	**0.70**	**0.76**

Table S2. Classification Performance for MDD vs. HC. Results for the best-performing embedding model are shown. SVC = Support Vector Classifier. The Dummy classifier uses a stratified strategy.

Sensitivity Analysis: Regression within MDD Cohort

To address the possibility that model performance was primarily driven by separating the distinct distributions of HC and MDD participants, we performed a sensitivity analysis by repeating the regression pipeline using only data from the MDD cohort. The results

are presented in Table S3. Only the Qwen3-8B sentence embeddings were able to predict BDI scores better than a chance baseline ($R^2 = 0.22$), confirming that these features capture genuine variance in symptom severity within the clinical population.

Feature Set	**MAE**	**R^2**
Any Feature Set (Dummy Baseline)	7.22	-0.08
Lexical & Prosodic Metrics	7.30	-0.12
Acoustic (eGeMAPS)	7.22	-0.08
Acoustic (ComParE)	7.27	-0.08
TF-IDF	6.71	-0.01
Qwen3-8B embeddings	**6.00**	**0.22**

Table S3. Predictive Performance within MDD Cohort Only (N=128). All models are Support Vector Regressors with hyperparameters tuned via 5-fold nested cross-validation. MAE = Mean Absolute Error; R^2 = Coefficient of Determination. Best performance is highlighted in bold.