Scalable depression monitoring with smartphone speech: a multimodal benchmark and topic analysis

Daniel Emden[1,2], Maike Richter[3], Astrid Chevance[4,5], Ramona Leenings[1], Julian Herpertz[1,3], Lara Gutfleisch[1], Anna Fleuchaus[1], Rogério Blitz[1], Vincent L. Holstein[2,6,7], Janik Goltermann[2], Nils R. Winter[2], Jennifer Spanagel[2], Susanne Meinert[2], Tiana Borgers[2], Kira Flinkenflügel[2], Frederike Stein[8], Nina Alexander[8], Hamidreza Jamalabadi[8], Jonathan Repple[2,9,10], Christian Dobel[11], Elisabeth J. Leehr[2], Ronny Redlich[2,12,13,14], Ulrich Ebner-Priemer[15], Igor Nenadić[8], Tilo Kircher[8,16], Udo Dannlowski[2], Tim Hahn[2], and Nils Opel[* 3,14]

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

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

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

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

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

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

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

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

[9]Department of Psychiatry, Psychosomatic Medicine and Psychotherapy, University Hospital Frankfurt, Frankfurt am Main, Germany

[10]Cooperative Brain Imaging Center - CoBIC, Goethe University Frankfurt, Frankfurt am Main, Germany

[11]Department of Otorhinolaryngology, Institute of Phoniatry and Pedaudiology, Jena University Hospital, Jena, Germany

[12]Department of Psychology, Martin Luther University Halle-Wittenberg, Halle, Germany

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

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

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

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

[*]Corresponding author. E-mail: nils.opel@charite.de

Abstract

Objective, scalable biomarkers are needed for continuous monitoring of major depressive disorder (MDD). Smartphone-collected speech is promising, yet extracting clinically useful signals remains difficult. We analysed 3 151 weekly voice diaries from 284 German-speaking adults (128 MDD, 156 controls) and regressed Beck Depression Inventory (BDI) scores. Sentence embeddings from the open-source 8-billion-parameter Qwen3-8B model predicted scores with MAE = 4.45 and $R^2 = 0.35$, explaining 16 more points of variance than the best traditional feature set (TF-IDF). Adding lexical–prosodic or TF-IDF features provided only marginal improvement (best MAE = 4.39). To interpret the embeddings we applied BERTopic and uncovered ten coherent themes; BDI scores peaked for "Persistent Low Mood" and "Pain Distress", confirming clinical relevance. Large-language-model embeddings therefore capture the dominant signal of depression severity in everyday speech and, paired with interpretable topic analysis, offer a privacy-preserving, scalable route to digital mental-health phenotyping.

Keywords: Major Depressive Disorder; Smartphone Speech; Digital Phenotyping; Natural Language Processing; Large Language Models; Sentence Embeddings; Topic Modelling

1 Introduction

Depression is a leading contributor to the global burden of disease, defined by its chronic and recurrent nature. Consequently, 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 bench-

mark 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.

2 Methods

2.1 Participants

The data for this study were drawn from a longitudinal project that used the ReMAP (Remote Monitoring Application in Psychiatry) smartphone application to acquire weekly 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. More than two-thirds of the participants in this dataset are part of the Marburg-Münster Affective Disorders Cohort Study (MACS) [19, 20], with data collected at two sites (Marburg and Münster, Germany) using identical study protocols. 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; [21]) 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 and Group Differences. 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.

2.2 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.

2.3 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.

2.3.1 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 [22]. 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.

2.3.2 Acoustic Feature Extraction

From the raw audio files, we extracted two well-established, high-dimensional acoustic feature sets using the Python port of the openSMILE toolkit: the comprehensive ComParE functionals set (6,373 features; [23]) and the psychologically relevant eGeMAPS v02 set (88 features; [24]).

2.3.3 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. The complete list of lexical and prosodic metrics is provided in Table 2.

- **TF-IDF Features:** Term Frequency–Inverse Document Frequency vectors calculated from 1- and 2-word n-grams in the normalized transcripts; they weight each term by its rarity across the corpus, thus highlighting words and phrases that uniquely characterise an individual diary.

- **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.

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 2. List of Lexical and Prosodic Metrics

2.4 Predictive Modelling

Predictive-modelling pipeline. We predicted the severity of depressive symptoms with a machine-learning pipeline that (1) standardised each feature matrix, (2) reduced dimensionality by principal-component analysis (PCA), and (3) fitted a support-vector regression (SVR) model. The ground-truth label was the Beck Depression Inventory (BDI) score recorded in the ReMAP app within ± 7 days of the corresponding voice diary. A separate pipeline was trained for every single-modal feature set and for each predefined multimodal combination.

Software. All analyses were run in Python 3.11 using scikit-learn 1.6; PCA and SVR were GPU-accelerated with cuML 25.02.

Cross-validation and hyper-parameter tuning. Model performance was estimated with nested, participant-stratified five-fold cross-validation (GroupKFold in both outer

and inner loops). The outer loop produced unbiased test estimates, whereas the inner loop tuned the SVR kernel, regularisation strength (C), ϵ-insensitive loss width, and the number of retained PCA components. Hyper-parameter search was orchestrated by Optuna's median-pruner algorithm. Performance is reported as mean absolute error (MAE) and explained variance (R^2).

Baselines. As points of reference we fitted (i) a `DummyRegressor` with strategy="mean" for all regression tasks and (ii) a `DummyClassifier` with strategy="stratified" for the HC vs. MDD supplementary classification.

Supplementary analyses. To contextualise the regression results we additionally (i) trained a binary classifier to distinguish healthy controls (HC) from participants with major depressive disorder (MDD), reporting balanced accuracy and AUROC, and (ii) repeated the full regression pipeline on the MDD subgroup alone to verify that performance reflected intra-group symptom variation rather than diagnostic differences.

Statistical significance. Significance was assessed with 1,000 label-shuffling permutations per feature set and 1,000 paired permutations comparing the leading models. Two-tailed p-values were computed as $(r + 1)/(N + 1)$, and all permutations were executed within the outer cross-validation folds to preserve the nested data-splitting structure.

2.5 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.

3 Results

3.1 Predictive Modelling

We evaluated a range of feature modalities for their ability to predict continuous BDI scores (Table 3). All feature-based models substantially outperformed the dummy regressor (MAE = 6.24). The model based solely on Qwen3-8B sentence embeddings achieved an MAE of 4.45 and an R^2 of 0.35, decisively outperforming every classical single-modality alternative (best classical: TF–IDF, MAE = 5.25, $R^2 = 0.19$).

Adding shallow features yielded small but consistent gains. Two combinations, namely embeddings + lexical–prosodic metrics and embeddings + TF–IDF, both yielded modest performance improvements (best: MAE = 4.39, $R^2 = 0.35$). Paired permutation tests confirmed that each of these multimodal models was significantly superior to embeddings alone ($p = 0.021$ and $p = 0.018$, respectively), although the absolute improvements were modest (ΔMAE < 0.07). No other fusion, acoustic or otherwise, matched this level of benefit.

Label-shuffling permutation tests (1,000 iterations per feature set) verified that every reported model performed above chance ($p < 0.001$). Effect sizes mirrored the descriptive results: embeddings delivered a large performance leap over the strongest classical baseline (ΔMAE = 0.80, $\Delta R^2 = 0.16$), whereas the multimodal additions provided only incremental refinement.

Feature Set	MAE (SD)	R^2 (SD)
Any Feature Set (Dummy Baseline)	6.24 (0.35)	-0.01 (0.05)
Single Modality		
Lexical & Prosodic Metrics	5.81 (1.04)	-0.15 (0.21)
TF–IDF	5.25 (0.44)	0.19 (0.07)
eGeMAPS acoustic	5.61 (0.77)	-0.10 (0.13)
ComParE acoustic	5.58 (0.74)	-0.04 (0.14)
Qwen3-8B embeddings	**4.45 (0.25)**	**0.35 (0.10)**
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	4.42 (0.24)	0.35 (0.11)
Qwen3-8B + TF–IDF	**4.39 (0.23)**	**0.35 (0.10)**
Qwen3-8B + eGeMAPS	4.46 (0.27)	0.35 (0.10)
Qwen3-8B + ComParE	4.69 (0.33)	0.29 (0.11)

Table 3. 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 single modality and multimodal combination are highlighted in bold.

To benchmark the robustness of our findings, we conducted two validity checks. First, we trained binary classifiers to distinguish participants with Major Depressive Disorder (MDD) from healthy controls (HC). Results (see Table 4) showed that the best-performing models combined Qwen3-8B embeddings with either lexical–prosodic or eGeMAPS acoustic features, each achieving a balanced accuracy of 0.70. The embeddings alone reached 0.68, markedly surpassing the strongest classical baseline (TF–IDF, 0.62) and all acoustic-only models, confirming that the semantic information captured by large-scale embeddings carries a strong, readily extractable signal of depression status. Label-shuffling permutation tests confirmed that every classification model performed significantly above chance ($p < 0.01$), while a paired permutation test found no significant difference between the two leading multimodal models (balanced-accuracy $\Delta \approx 0$, $p > 0.05$).

Second, to test whether our models were simply separating healthy controls from patients rather than tracking symptom variation within a clinical cohort, we repeated the

entire predictive-modelling pipeline using only the 1 271 diaries contributed by the 128 MDD participants (Table 5). In this more challenging setting the Qwen3-8B embeddings were the only feature set to demonstrate predictive power, achieving an MAE of 6.01 and an R^2 of 0.15. In contrast, all other feature modalities (lexical, acoustic, and TF-IDF) performed at, or slightly below, the dummy baseline (MAE = 7.37; R^2 = –0.06).

Feature Set	**Balanced Accuracy (SD)**	**AUC-ROC (SD)**
Any Feature Set *(Dummy Baseline)*	0.50 (0.00)	0.50 (0.00)
Lexical & Prosodic Metrics	0.59 (0.05)	0.68 (0.04)
eGeMAPS acoustic	0.55 (0.03)	0.55 (0.08)
ComParE acoustic	0.59 (0.05)	0.60 (0.14)
TF-IDF	0.62 (0.03)	0.70 (0.07)
Lexical & Prosodic + TF-IDF	0.62 (0.04)	0.70 (0.04)
Qwen3-8B Embeddings	0.68 (0.06)	0.76 (0.06)
Qwen3-8B + Lexical & Prosodic	**0.70 (0.03)**	**0.77 (0.04)**
Qwen3-8B + eGeMAPS acoustic	0.70 (0.04)	0.77 (0.05)

Table 4. Classification performance (mean (SD) across outer folds) for distinguishing participants with major depressive disorder from healthy controls. Balanced accuracy and area under the ROC curve (AUC-ROC) are averaged across outer folds of nested cross-validation; the dummy baseline predicts the majority class.

Feature Set	**MAE**	R^2
Any Feature Set (Dummy Baseline)	7.37 (1.60)	-0.06 (0.19)
Lexical & Prosodic Metrics	7.30 (1.48)	-0.21 (0.09)
eGeMAPS acoustic	7.26 (1.47)	-0.21 (0.09)
ComParE acoustic	7.21 (1.64)	-0.18 (0.21)
TF–IDF	6.93 (1.68)	-0.10 (0.18)
Qwen3-8B embeddings	**6.01 (1.45)**	**0.15 (0.17)**

Table 5. 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.

3.2 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 6). 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 reflecting emotional distress such as "Persistent Low Mood" (Mean BDI = 14.03) and "Pain Distress" (Mean BDI = 12.24) were associated with the highest depression scores. The most pronounced difference was found between "Persistent Low Mood" and the activity-focused topic "Planning & Leisure" (Mean BDI = 0.67), 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.

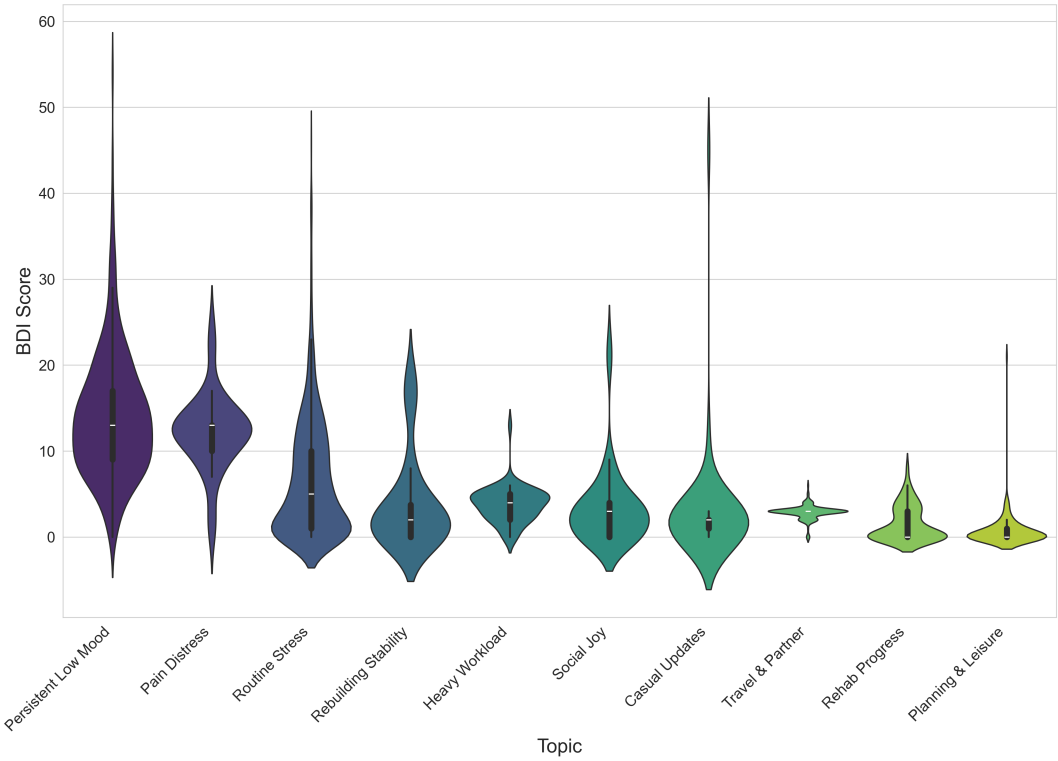

Figure 1. Distribution of Beck Depression Inventory (BDI) scores across 10 machine-learning-derived topics. The topics are arranged as violin plots along the horizontal axis in descending order of their mean BDI score. This ordering reveals a clear severity gradient. The leftmost topics, such as "Persistent 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 & representative documents)**	**Documents (n)**	**Subjects (N)**	**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 6. Topic Summary and BDI Statistics. Topics are ordered from highest to lowest Mean BDI score. Documents (n) = number of voice diary entries assigned to each topic; Subjects (N) = number of unique participants contributing to each topic. The outlier topic has been excluded from this analysis.

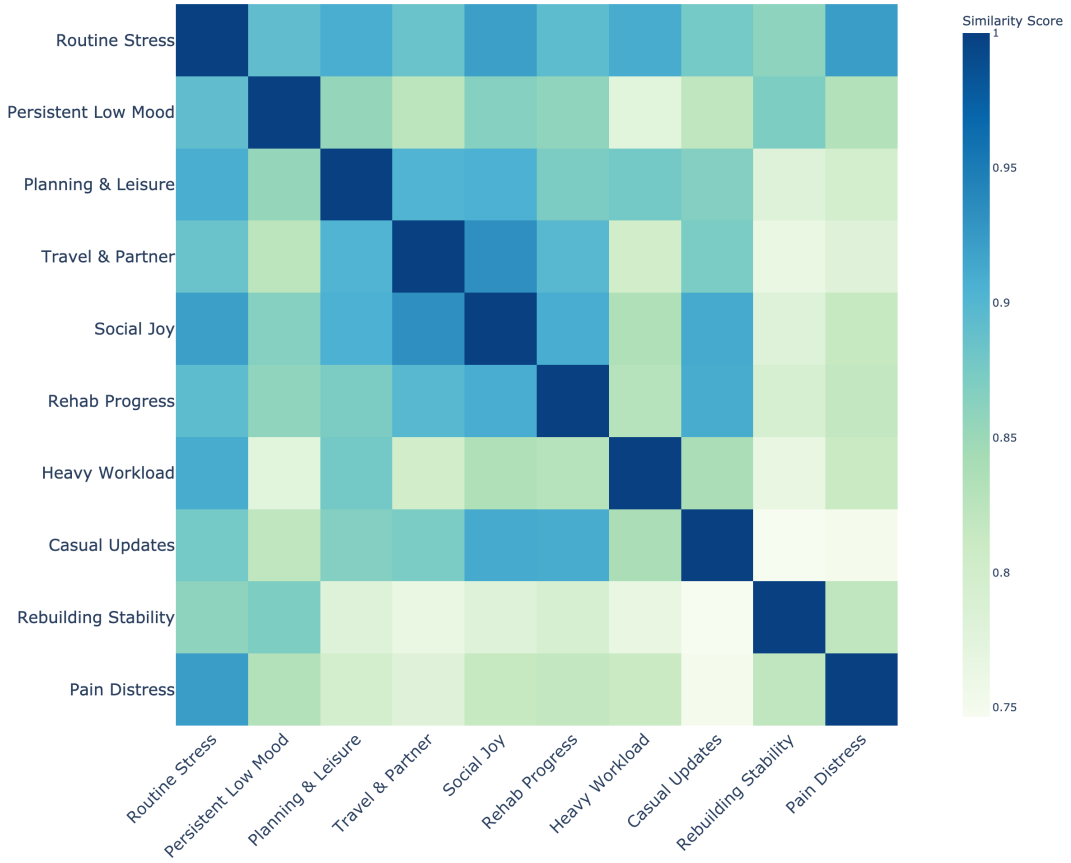

Figure 2. Pairwise cosine similarities among the ten BERTopic themes are displayed as a hierarchically ordered heat-map (Qwen3-8B embeddings; diagonal fixed at 1.0). A compact high-similarity block (deep blue) links the routine narratives Routine Stress, Social Joy and Casual Updates, while a second, 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.

4 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 adding either lexical–prosodic or TF–IDF features yielded slightly better scores (best: embeddings + TF–IDF, MAE = 4.39; $R^2 = 0.35$), this represents only a modest improvement, suggesting that embeddings capture the vast majority of the predictive signal. The small additional contribution from features like speech duration and pause ratios warrants further investigation, but it may indicate that basic speech timing and word rarity provide complementary, non-semantic information. 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 central contribution of this study is to show that high-performance models can be made interpretable rather than remaining inscrutable "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 [25–27].

Crucially, the ability of a separate model to predict BDI scores using only these topic probabilities confirms their clinical relevance. This dual approach of 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 [28].

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], which could potentially improve 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, our findings are based on a German-speaking cohort recruited primarily from two metropolitan areas (Marburg and Münster) using a bespoke research application (ReMAP). The specific demographics and clinical characteristics of this sample may therefore limit the broader generalizability of our findings to other languages, cultures, and healthcare systems. 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

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

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.

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.

Acknowledgements 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).

Author Contributions TH and NO had the initial idea of the ReMAP app and designed the project from which these data are drawn. DE conceived and designed the study, developed the ReMAP app, implemented all preprocessing and modelling code, performed statistical analyses, created visualisations and wrote the main manuscript text. MR, AC, RL, JH, VH and EL contributed to study design, data interpretation and substantive manuscript revisions. LG, AF, RB, JG, NW, JS, SM, TB, KF, FS, NA and HJ coordinated participant recruitment, collected speech diaries and associated clinical data, performed data quality assurance and edited the manuscript. JR, CD, RR, UEP, IN, TK, UD, TH and NO provided clinical oversight, project supervision and edited the manuscript. All authors provided critical feedback and approved the final version. DE and NO are the guarantors of the study.

References

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

2. König, H. *et al.* The excess costs of depression: a systematic review and meta-analysis. en. *Epidemiology and Psychiatric Sciences* **29.** Publisher: Cambridge University Press (CUP). ISSN: 2045-7960, 2045-7979. doi:10.1017/s2045796019000180. https://www.cambridge.org/core/product/identifier/S2045796019000180/type/journal_article (2025) (2020).

3. Melcher, J. *et al.* Digital phenotyping for mental health of college students: a clinical review. en. *Evidence Based Mental Health* **23.** Publisher: BMJ, 161–166. ISSN: 1362-0347, 1468-960X. doi:10.1136/ebmental-2020-300180. https://mentalhealth.bmj.com/lookup/doi/10.1136/ebmental-2020-300180 (2025) (Nov. 2020).

4. Bufano, P. *et al.* Digital Phenotyping for Monitoring Mental Disorders: Systematic Review. en. *Journal of Medical Internet Research* **25.** Publisher: JMIR Publications Inc., e46778. ISSN: 1438-8871. doi:10.2196/46778. https://www.jmir.org/2023/1/e46778 (2025) (Dec. 2023).

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

6. Wankhade, M. *et al.* A survey on sentiment analysis methods, applications, and challenges. en. *Artificial Intelligence Review* **55.** Publisher: Springer Science and Business Media LLC, 5731–5780. ISSN: 0269-2821, 1573-7462. doi:10.1007/s10462-022-10144-1. https://link.springer.com/10.1007/s10462-022-10144-1 (2025) (Oct. 2022).

7. Devlin, J. *et al.* *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* en. in *Proceedings of the 2019 Conference of the North* (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), 4171–4186. doi:10.18653/v1/N19-1423. http://aclweb.org/anthology/N19-1423 (2025).

8. Hossain, M. M. *et al.* Multi task opinion enhanced hybrid BERT model for mental health analysis. en. *Scientific Reports* **15.** Publisher: Springer Science and Business Media LLC. ISSN: 2045-2322. doi:10.1038/s41598-025-86124-6. https://www.nature.com/articles/s41598-025-86124-6 (2025) (Jan. 2025).

9. Pourkeyvan, A. *et al.* Harnessing the Power of Hugging Face Transformers for Predicting Mental Health Disorders in Social Networks. *IEEE Access* **12.** Publisher: Institute of Electrical and Electronics Engineers (IEEE), 28025–28035. ISSN: 2169-3536. doi:10.1109/access.2024.3366653. https://ieeexplore.ieee.org/document/10438433/ (2025) (2024).

10. Bey, R. *et al.* Natural language processing of multi-hospital electronic health records for public health surveillance of suicidality. en. *npj Mental Health Research* **3.** Publisher: Springer Science and Business Media LLC. ISSN: 2731-4251. doi:10.1038/s44184-023-00046-7. https://www.nature.com/articles/s44184-023-00046-7 (2025) (Feb. 2024).

11. Grootendorst, M. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* Version Number: 1. 2022. doi:10.48550/ARXIV.2203.05794. https://arxiv.org/abs/2203.05794 (2025).

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

13. Cummins, N. *et al.* A review of depression and suicide risk assessment using speech analysis. en. *Speech Communication* **71.** Publisher: Elsevier BV, 10–49. ISSN: 0167-6393. doi:10.1016/j.specom.2015.03.004. https://linkinghub.elsevier.com/retrieve/pii/S0167639315000369 (2025) (July 2015).

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

15. Low, D. M. *et al.* Automated assessment of psychiatric disorders using speech: A systematic review. en. *Laryngoscope Investigative Otolaryngology* **5.** Publisher: Wiley, 96–116. ISSN: 2378-8038, 2378-8038. doi:10.1002/lio2.354. https://onlinelibrary.wiley.com/doi/10.1002/lio2.354 (2025) (Feb. 2020).

16. Torous, J. *et al.* New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. en. *JMIR Mental Health* **3.** Publisher: JMIR Publications Inc., e16. ISSN: 2368-7959. doi:10.2196/mental.5165. http://mental.jmir.org/2016/2/e16/ (2025) (May 2016).

17. Goltermann, J. *et al.* Smartphone-Based Self-Reports of Depressive Symptoms Using the Remote Monitoring Application in Psychiatry (ReMAP): Interformat Validation Study. en. *JMIR Mental Health* **8.** Publisher: JMIR Publications Inc., e24333. ISSN: 2368-7959. doi:10.2196/24333. https://mental.jmir.org/2021/1/e24333 (2025) (Jan. 2021).

18. Emden, D. *et al.* Technical feasibility and adherence of the Remote Monitoring Application in Psychiatry (ReMAP) for the assessment of affective symptoms. en. *Journal of Affective Disorders* **294,** 652–660. ISSN: 01650327. doi:10.1016/j.jad.2021.07.030. https://linkinghub.elsevier.com/retrieve/pii/S0165032721007096 (2025) (Nov. 2021).

19. Vogelbacher, C. *et al.* The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data. en. *NeuroImage* **172,** 450–460. ISSN: 10538119. doi:10.1016/j.neuroimage.2018.01.079. https://linkinghub.elsevier.com/retrieve/pii/S105381191830079X (2025) (May 2018).

20. Kircher, T. *et al.* Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. en. *European Archives of Psychiatry and Clinical Neuroscience* **269,** 949–962. ISSN: 1433-8491. doi:10.1007/s00406-018-0943-x. https://doi.org/10.1007/s00406-018-0943-x (2025) (Dec. 2019).

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

22. Radford, A. *et al. Robust Speech Recognition via Large-Scale Weak Supervision* arXiv:2212.04356 [eess]. Dec. 2022. doi:10.48550/arXiv.2212.04356. http://arxiv.org/abs/2212.04356 (2025).

23. Eyben, F. *et al. Opensmile: the munich versatile and fast open-source audio feature extractor* en. in *Proceedings of the 18th ACM international conference on Multimedia* (ACM, Firenze Italy, Oct. 2010), 1459–1462. ISBN: 9781605589336. doi:10.1145/1873951.1874246. https://dl.acm.org/doi/10.1145/1873951.1874246 (2025).

24. Eyben, F. *et al.* The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* **7,** 190–202. ISSN: 1949-3045. doi:10.1109/TAFFC.2015.2457417. http://ieeexplore.ieee.org/document/7160715/ (2025) (Apr. 2016).

25. Rude, S. *et al.* Language use of depressed and depression-vulnerable college students. en. *Cognition & Emotion* **18.** Publisher: Informa UK Limited, 1121–1133. ISSN: 0269-9931, 1464-0600. doi:10.1080/02699930441000030. http://www.tandfonline.com/doi/abs/10.1080/02699930441000030 (2025) (Dec. 2004).

26. Ramirez-Esparza, N. *et al.* 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.** Publisher: Association for the Advancement of Artificial Intelligence (AAAI), 102–108. ISSN: 2334-0770, 2162-3449. doi:10.1609/icwsm.v2i1.18623. https://ojs.aaai.org/index.php/ICWSM/article/view/18623 (2025) (Sept. 2021).

27. De Choudhury, M. *et al.* Predicting Depression via Social Media. *Proceedings of the International AAAI Conference on Web and Social Media* **7.** Publisher: Association for the Advancement of Artificial Intelligence (AAAI), 128–137. ISSN: 2334-0770, 2162-3449. doi:10.1609/icwsm.v7i1.14432. https://ojs.aaai.org/index.php/ICWSM/article/view/14432 (2025) (Aug. 2021).

28. Jim, J. R. *et al.* Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. en. *Natural Language Processing Journal* **6.** Publisher: Elsevier BV, 100059. ISSN: 2949-7191. doi:10.1016/j.nlp.2024.100059. https://linkinghub.elsevier.com/retrieve/pii/S2949719124000074 (2025) (Mar. 2024).