scientific reports

Check for updates

OPEN

Understanding mental health discourse on Reddit with transformers and explainability

Irene Sánchez Rodríguez[1,3,4]✉, John Bianchi[2,3,4], Fabio Pinelli[2], Folco Panizza[1], Emiliano Ricciardi[1], Pietro Pietrini[1] & Marinella Petrocchi[2,3]

Social media is a powerful tool also for discussing mental health. The conversations that take place in these spaces provide a unique insight into how users talk about the issue. This study uses fine-tuned pretrained transformer models (BERT and MentalBERT), to classify Reddit posts about anxiety, depression, bipolar disorder and borderline personality disorder (BPD) in specialised subreddits. By assessing how well subreddit conversations align with their intended mental health focus, the analysis ensures that these communities are effectively serving their purpose as support spaces. Our classification models achieve an average accuracy of 82%, with MentalBERT slightly outperforming BERT. To ensure transparency, we use Local Interpretable Model-agnostic Explanations (LIME) to identify key linguistic patterns that influence the model predictions. The outcome reveals distinct language use across conditions: as examples, discussions in bipolar disorder subreddits often refer to mood instability, while BPD communities emphasise challenges in emotional regulation. By integrating classification with explainability, this study offers insights into thematic patterns in online discourse that can support mental health professionals in identifying trends. While our models are not diagnostic tools, they function as subreddit-alignment classifiers, helping to uncover how different topics are discussed across communities. These insights may inform human-in-the-loop community management strategies and contribute to raising awareness and reducing stigma around mental health issues, ultimately fostering more supportive digital environments.

Mental health remains a critical global public health issue, with disorders such as anxiety, depression, bipolar disorder and borderline personality disorder (BPD) placing a significant burden on individuals and health systems. These conditions contribute to disability, reduced quality of life and increased mortality, highlighting the urgent need for improved mental health interventions. Depression and anxiety are the most common and severe mental disorders worldwide, and their cost is projected to increase until at least 2044, disproportionately affecting women and middle-aged adults[1]. Bipolar disorder and Borderline Personality Disorder (BPD) remain a persistent global challenge and are associated with severe emotional distress and high suicide risk[2,3]. Despite advancements in diagnostic frameworks and therapeutic interventions, challenges remain in ensuring timely and accurate diagnosis and treatment of mental health conditions[4,5]. These challenges are compounded by the stigma associated with mental illness, which often discourages individuals from seeking professional help[6]. When seeking help feels stigmatizing, many people look for safer, more accessible alternatives. Anonymous online communities have become one of the main outlets for this. The proliferation of social media platforms has fundamentally changed the way individuals interact, share information, and seek support. They serve as important arenas for social interaction, providing spaces where individuals can freely express themselves and exchange ideas on various topics[7]. Because users can post under pseudonyms, social platforms reduce the fear of being publicly labeled, allowing more candid discussion of mental health conditions. At the same time, these spaces enable large-scale data analysis of mental health discourse[8–10].

In this paper, we consider Reddit (https://www.reddit.com/), a social media platform organized into topic-focu s communities called subreddits. Reddit has emerged as an important space for users to share experiences related to sensitive and stigmatized issues[11]. Its unique features, such as user anonymity, accessibility, large user base, and engagement-driven content structure, make it an ideal source for studying online mental health discussions and understanding how individuals communicate their psychological distress[12]. Because of their public accessibility, mental health subreddits are not limited to individuals who have received a clinical diagnosis, but

[1]Molecular Mind Lab, IMT School for Advanced Studies Lucca, Lucca 55100, Italy. [2]SysMA, IMT School for Advanced Studies Lucca, Lucca 55100, Italy. [3]Istituto di Informatica e Telematica - CNR, Pisa 56124, Italy. [4]These authors contributed equally this work: Irene Sánchez Rodríguez and John Bianchi. ✉email: irene.sanchez@imtlucca.it

attract people with mental health concerns. Thus, subreddits are treated as "spaces of concern" where users with common challenges share advice, vent, and seek support. By studying these spaces, researchers can gain insight into how mental health concerns manifest themselves in unstructured, natural conversations[13]. Reddit data are particularly valuable because they reflect real-world experiences, often including the perspectives of people who may be reluctant to seek help from health professionals. However, the sheer volume and heterogeneity of online content pose significant challenges for manual analysis, requiring the use of advanced computational techniques.

Recent studies demonstrate that machine learning models can effectively classify social media posts that discuss mental health topics by analyzing linguistic patterns and semantic features[14–18]. Most of these models classify content into specific categories by learning from labeled data and exploiting the latest advances in natural language processing. However, a key challenge remains: improving explainability, i.e., understanding how predictions are generated. The opacity of modern high-capacity neural models, especially deep learning architectures, makes it difficult to directly interpret their decision-making processes[19–22]. This "black box" problem complicates the adoption of AI in mental health, as clinicians need to understand how decisions are made. Identifying key features, such as emotionally charged words or disorder-related patterns, is crucial to increase confidence in the reliability of the prediction model[23].

To address this issue, explainability methods such as Local Interpretable Model-agnostic Explanations (LIME)[24] and SHAP[25] have been proposed. Below, we briefly summarize relevant prior work in classification and explainability for mental-health social media data.

Related work

Early work on using social media to detect mental health discourse relied on traditional machine learning pipelines with engineered n-gram or psycholinguistic features[18]. Soon after, large pre-trained language models such as BERT and its derivatives became the dominant approach: Kim et al.[10] achieved an F1 score of 0.86 for depression using a CNN-based multiclass Reddit model, while Le et al.[15] demonstrated that DistilBERT outperforms CNN/LSTM baselines for social anxiety detection. Zhang et al.[14] review natural language processing methods for detecting mental illness across heterogeneous sources (clinical notes, EHRs, social media), spanning a decade and demonstrating how traditional machine learning classifiers have gradually been replaced by deep learning-based classifiers, including transformers. Mostly focusing on depression and suicidal intent detection tasks, this review emphasizes the need for explainability. Domain-adapted checkpoints such as MentalBERT[26] offer only modest improvements in colloquial Reddit language[16]. It is worth noting that the cited studies have mostly focused on binary or single-diagnosis setups. Multi-class setups spanning several related conditions are less common, often collapsing comorbid disorders into a single label and precluding cross-condition error analysis. Recent work evaluates LLaMA-3 strategies for mental-health text classification and finds that RAG (Retrieval-Augmented Generation) currently lags behind fine-tuning on multi-class settings (e.g[27].,: $\sim$56% vs. $\sim$80-82%), suggesting that RAG is at the time of writing suboptimal for this classification task.

Since our work addresses a balanced, four-class classification task involving anxiety, depression, bipolar disorder and BPD, we perform classification based on BERT-like transforme models. As a second step for contextualizing, we compare results to those obtained from a zero-shot large language model

Explainability of classification results. Transformer attention maps offer a first glimpse of model reasoning[21]. However, in practical clinical workflows, after the model is trained, most people still rely on add-on tools like LIME, which explain each prediction by showing which input words mattered most. Among post hoc explainers, LIME works by slightly perturbing the input text, observing how the model's prediction shifts, and then training a simple local surrogate to mimic that behavior. On the other hand, SHAP, by contrast, draws on cooperative game theory to quantify how much each feature contributes to the prediction, producing both instance-level attributions and a consistent global view of model behaviour[24,25].

Some works target causal relations and causal inference in mental health text[28,29]. Garg et al[28]. provide CAMS, a corpus of mental health posts (including Reddit) annotated by the authors for cause/effect categories. Saxena et al[29]. then train explainable classifiers to predict these causal categories, explicitly aiming to "identify the cause behind the user's intention", and explain them with LIME/Integrated Gradients. Thus, these works aim to classify posts for causal analysis, i.e. to recognise the cause of discomfort and the resulting effect in each post. While these works are clearly relevant, as they help clinicians to better understand the causes and effects of mental illness, they differ from our own work, which aims to understand how relevant a Reddit post is to its community.

A growing body of research underscores the clinical relevance of such tools. Band et al.[30] systematically review explainability methods and confirm LIME/SHAP as the prevailing techniques in mental-health applications. Jo et al.[31] combine the two within a depression-prediction framework, achieving 99.3 % accuracy with a Random-Forest model while clarifying how co-morbid indicators (schizophrenia, bipolar disorder, anxiety) influence the decision. For social media text, Alghazzawi et al.[32] employ SHAP with an ensemble to detect suicidal ideation, providing global explanations.

Despite this progress, two gaps persist when considering social media texts: (i) Often, posts, although belonging to different categories, have many words in common, and in the literature it is rare that an explanation is given as to why the model classifies the post as belonging to one category or another, and (ii) few studies compare which words are shared across disorders and which are unique when the model must choose among several conditions at once.

We address both the gaps by aggregating LIME explanations over 5-fold cross-validation and visualizing term importance distributions, while also highlighting cross-condition overlaps. Computational constraints also influenced our decision to use LIME rather than SHAP for large-scale Reddit data, as explained in more detail later in the paper. By integrating explainability techniques into our pipeline, we aim to increase the transparency and effectiveness of deep-learning-based mental health interventions.

What we do in this study

We analyze the posts of four subreddits dedicated to four different mental disorders, anxiety, depression, bipolar disorder, and BPD; classifying each post according to its subreddit-aligned discourse category. For the classification, we use two pretrained transformer models, BERT[33] and MentalBERT[26]. Our classification results indicate an accuracy of approximately 82%, with MentalBERT slightly outperforming BERT in all metrics. The explainability analysis further reveals key linguistic markers unique to each category. Words like "panic" and "attack" dominate anxiety-related discussions, whereas "hopeless" and "worthless" are prominent in depression-related content. BPD-related posts highlight terms like "abandonment" and "attachment," while bipolar-related discussions feature words such as "mood," "mania," and "cycle." These insights underscore the importance of explainability in deep-learning-based mental health applications, enhancing transparency and trust in computational approaches for analyzing mental health discourse.

What we do not do in this study

It is important to emphasize that this study does not aim to diagnose users or infer their mental health status based on their posts. A user's presence in a particular mental health subreddit does not mean that they have the disorder or condition discussed in that community. Online forums serve a diverse range of purposes, including seeking information, providing support, sharing experiences, or engaging in general discussions about mental health. Moreover, while our classification model effectively predicts subreddit alignment based on linguistic patterns, it does not assess the accuracy, validity, or clinical significance of user-generated content. Our goal is strictly to assess the thematic alignment between subreddit discussions and their intended focus, providing insights into online mental health discourse. As such, any potential application of this research should be framed within the context of online discourse analysis rather than individual psychological or psychiatric assessment. Our classifier is not a clinical screening tool and should not be used as a proxy for diagnosis.

Our contribution

We present an applied, balanced four-class Reddit study (r/Anxiety, r/Depression, r/Bipolar, r/BPD) using fine-tuned BERT and MentalBERT. We aggregate LIME attributions across 5-fold cross-validation to derive per-class lexical profiles, and we include a zero-shot LLM baseline for context. Prior studies often focus on binary or single-condition setups or report explanations at the instance level; here we integrate supervised classification with fold-aggregated post-hoc explanations to examine subreddit-aligned discourse across four communities. We position this as an incremental, reproducible contribution for online discourse analysis and not as a clinical diagnostic tool.

Applications

The findings of this research have practical implications, particularly in three key areas:

Online support analysis: Mental health professionals can use our findings to gain insights into how users articulate their mental health experiences, express distress, and seek peer support in online spaces. Recognizing common discussion topics and unexpected themes in subreddits can help identify areas where people are struggling to find support, allowing for the creation of more tailored resources and interventions.

Tools for Clinicians and Mental Health Research: While not intended for clinical purposes, our approach may help clinicians and researchers analyze linguistic patterns in online mental health discussions. By understanding how individuals describe their experiences compared to clinical terminology, this research can inform studies on early indicators of mental distress and refine strategies for patient communication. Any clinical deployment would require a different study design and regulatory validation; this work is strictly retrospective and research oriented.

Enhancing Discussion Quality: explainability findings can assist platform moderators in ensuring thematic consistency within mental health communities. By assessing whether posts align with the intended focus of a subreddit, this approach can reduce misinformation, foster relevant engagement, and support users in finding the most appropriate spaces for discussion.

We would like to remark that any operational use of these models must retain a human-in-the-loop and cannot substitute clinical judgment.

Research questions

To frame the study, we pose three research questions:

RQ1. Predictive performance: How accurately can transformer-based text classifiers distinguish among the four mental-health discourse categories represented by the subreddits r/Depression, r/Anxiety, r/BPD and r/Bipolar?

RQ2. Discourse patterns: Which lexical cues drive the classification, and do explanation methods reveal features that are shared across, or specific to, particular conditions?

RQ3. Comparative baseline: How does the proposed multi-class transformer pipeline compare, in predictive quality, with a zero-shot large-language-model baseline?

Methods

Data preparation

The initial dataset[17] includes posts from seven subreddits related to mental health: r/depression (258,496 posts), r/anxiety (86,243 posts), r/bipolar (41,493 posts), r/mentalhealth (39,373 posts), r/BPD (38,216 posts), r/schizophrenia (17,506 posts), and r/autism (7,142 posts). We selected four subreddits from the full set: r/anxiety, r/depression, r/bipolar, and r/BPD. These four conditions were chosen as the focus of this study because of their frequency in the clinical setting, diversity of symptomatology, and the role of online communities in mental

health discourse. r/mentalhealth was excluded because it is a broad, non-specific subreddit that covers many different mental health topics. Each post is labeled by its subreddit of origin (r/Anxiety, r/Depression, r/Bipolar, r/BPD), and each post belongs to exactly one of the four classes. Accordingly, we treat this as a single-label, multi-class classification task, following standard definitions[34].

Anxiety disorders and depression are among the most prevalent mental health conditions worldwide, affecting approximately 301 million people with anxiety and 280 million people with depression in 2019, contributing significantly to the global burden of disease, with incidence rates steadily increasing over the past few decades[1]. Bipolar disorder, which affects around 40–50 million people, is associated with significant mood instability and a high risk of suicide[35], while borderline personality disorder (BPD) is characterized by emotional dysregulation and impulsivity, contributing to severe psychosocial impairment[36]. The last two disorders remain among the most stigmatized mental health conditions, often leading individuals to seek peer support in online spaces to mitigate discrimination and access emotional validation[37]. Given these factors, the selected subreddits provide a diverse yet interconnected dataset that allows for a nuanced analysis of how people with different mental health conditions engage in online discourse and seek community support.

To mitigate class imbalance and ensure a fair evaluation across categories, we constructed a balanced dataset by downsampling the larger subreddits (r/anxiety, r/depression, r/bipolar) to match the size of the smallest class (r/BPD, with 38,216 posts). While this procedure removes a substantial number of posts from the larger subreddits, it avoids biasing the classifier toward majority classes and ensures that standard performance metrics (e.g., accuracy, F1) reflect true discriminative ability across all categories. This choice alters the original class distribution, which was highly skewed in favor of r/anxiety and r/depression. By enforcing uniform class priors, we make the classification task more balanced and statistically controlled, at the cost of discarding potentially informative data from the overrepresented classes. However, using unbalanced data would have resulted in inflated performance for dominant classes and reduced comparability between categories. We chose class balancing over alternative methods such as class-weighted loss or data augmentation to maintain a clean and interpretable experimental setup, particularly for explainability analysis, where uneven sample sizes could distort feature attribution and interpretation.

We randomly selected posts from the larger subreddits to match the smallest class. The final dataset comprised 152,864 posts (38,216 posts for each class: r/anxiety, r/BPD, r/bipolar, and r/depression), as shown in Fig. 1. We use this balanced dataset for all subsequent experiments, including model training and evaluation, and explainability analysis. With four balanced classes, a random-guess classifier would achieve an accuracy of 25%, which serves as a baseline for evaluating our models[38,39].

Classification task

In this section, we describe the models (and their configurations) used to classify a post as belonging to one of the four subreddits selected from the original dataset. The task is to perform a multi-class classification at the post level, categorizing each post into one of four different classes by using the following transformer-based models (It is important to emphasize that this study is a non-clinical, offline analysis of social media text and is not intended to be used for diagnostic purposes.):

- BERT (Bidirectional Encoder Representations from Transformers): A general-purpose, pre-trained language model designed to understand context through bi-directional representations[40].

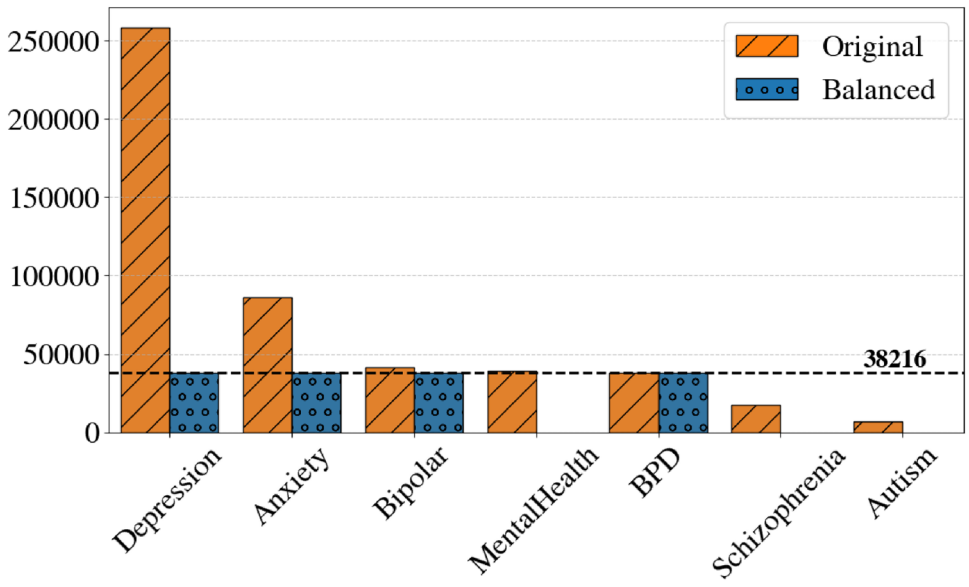


Fig. 1. Comparison of class distributions between the original and balanced dataset. The x-axis displays the different classes (subreddits) and the y-axis the number of samples per class.

- MentalBERT: A BERT variant pre-trained on mental health data that improves the detection and analysis of mental disorders from social content[41].

The two models were chosen because of reproducibility, cost-accessibility, and compatibility with word-level post-hoc explanations[42]. To contextualize our approach, we include a zero-shot LLM baseline and compare against it in the Results. This design aligns with calls to pair modern models with transparent interpretability practices in mental health NLP[14]. Sections "Explainability task" and "Explainability results" will explore this in detail.

We fine-tune both models with a maximum sequence length of 256 tokens to balance computational efficiency with performance. This short sequence length accelerates training and inference. 33,500 out of 152,864 posts (22%) exceeded this limit, meaning that we are using most of the available data (see Fig. 1 in Supplementary Material (S1) for a detailed view of the token distribution). To get an idea of which subreddits have posts longer than 256 tokens on average, we did some analysis and found that posts from the BPD subreddit exceed the 256-token limit by an average of 227 words, and the Anxiety subreddit exceed the same limit by an average of 201 words. Without truncation, post length could itself become a distinguishing feature, potentially biasing the classifier. By enforcing a fixed token limit, we ensure that the analysis remains focused on the linguistic content within the initial 256 tokens, thereby reducing the influence of overall post length on model predictions.

We train the models for four epochs with a learning rate of 2×10^{-5} and a batch size of 64. In order to process our large dataset efficiently, we use mixed precision training (FP16)[43], which significantly reduces the training time.

We use 5-fold cross-validation by splitting the dataset into five subsets to allow the models to be trained and evaluated on different pieces of data, ensuring consistency of results (see Fig. 2, first, second, third layer and fourth layer, from the top).

To check the robustness and consistency of the models under slight variations in the composition of the dataset, we perform an additional experiment. Since r/depression had the largest number of posts in the original dataset, we select a second random sample of 38,216 posts from this subreddit to replace the original r/depression subset used in the balanced dataset. The posts from r/anxiety, r/BPD, and r/bipolar remain unchanged. This approach allows us to test whether different samples of the same class may affect the models' performance while maintaining the overall dataset size and class balance. The results show no significant change in performance compared to the original sample.

We do not apply any textual pre-processing to the dataset. This decision is made to preserve the integrity of the original posts and to allow the models' explainability results to reflect a 100% realistic scenario, preserving all textual nuances.

To maximize the information available to the models, we manage each subreddit post by concatenating the title and text of the post, to provide as much contextual detail as possible for the models to learn from. The

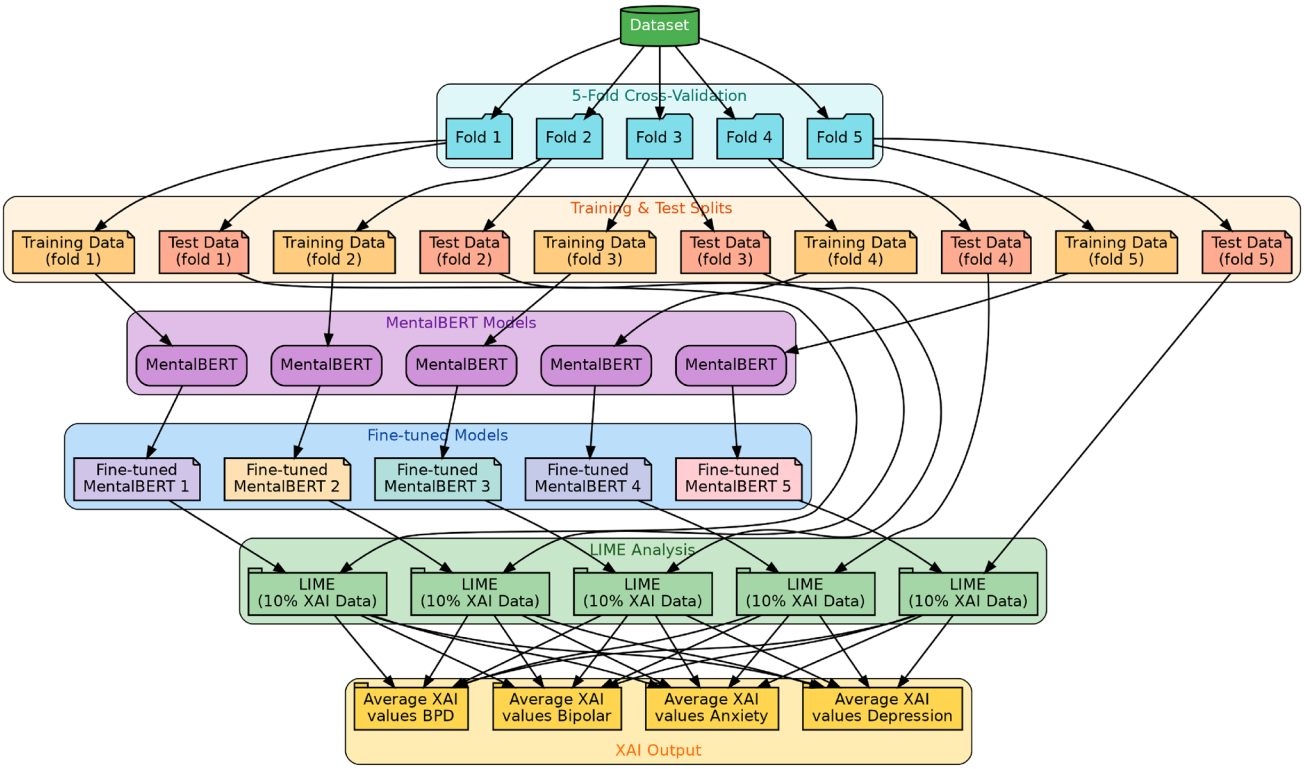


Fig. 2. Pipeline for classification and explainability tasks, with MentalBERT (same pipeline with BERT).

procedure does not favor certain classes. Analysis of the dataset revealed that titles are consistently short across all classes, contributing an average of just 4% to a post's total word count. This length varies minimally, with a standard deviation of 0.4 words. The models processed the concatenated text as a single, continuous sequence with no delimiter to distinguish between the title and body. This ensures that the structural element of a title does not influence the model's classification process disproportionately.

We ensure reproducibility by using a fixed random seed in all experiments. This means that operations like data splitting and other randomized processes consistently use the same seed, resulting in repeatable outcomes.

Comparison with a Zero-Shot LLM Baseline to validate the performance of our BERT and MentalBERT models on the multi-class subreddit classification task, we compared them against GPT-4.1 mini in a zero-shot setting. We selected GPT-4.1 mini based on its balance between performance and cost-effectiveness for experimental use, as noted in the OpenAI release announcement (OpenAI. (2025, April 14). Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/).

For the zero-shot evaluation, we used the GPT-4.1 mini model via the Azure OpenAI client, configuring it with a fixed random seed (42) for reproducibility. The system prompt instructed the model to act as an "expert text classification assistant" and to respond with only the exact category label. Each user message included the input text and the list of valid categories. We set the temperature to 0 to ensure deterministic output and limited the response length by setting `max_tokens` to 50.

Binary classification: mental health vs other topics

In addition to distinguishing between different mental health subreddits, we also evaluated the model's ability to separate mental health–related content from generic online discourse. This task reflects a realistic deployment scenario, where systems must first detect whether a post pertains to mental health at all.

To assess this, we utilized two distinct datasets to determine whether mental health-related posts can be reliably distinguished from general content. For mental health content, we used the balanced dataset described in Sect. "Data preparation". For the general content, we downloaded the wenknow/reddit_dataset_44 dataset from Hugging Face (https://huggingface.co/datasets/wenknow/reddit_dataset_44), which was the most downloaded Reddit dataset at the time of writing. This dataset is extremely large, with an estimated 131,958,531 rows. We focused on the most common subreddit topics within the wenknow/reddit_dataset_44 dataset, selecting the top 10 by total count, as shown in Table 1. We then filtered this dataset to include only rows marked as posts, excluding comments.

After filtering, the generic content dataset was smaller than our original, reduced mental health dataset (120,184 versus 152,864 rows). Consequently, we reduced the size of the mental health dataset to match that of the general content dataset. We performed a stratified selection across the four classes to ensure an equal number of posts per class in the mental health dataset, using a fixed random seed.

This resulted in two datasets of equal size. For classification, we transformed all four labels in the mental health dataset into a single 'mentalHealth' label, and all labels in the general content dataset into a single 'generic' label. We then launched the training process using both BERT and MentalBERT models in a 5-fold classification setup.

Explainability task

To better interpret the predictions of the classification models, we apply LIME[24], an Explainable AI (xAI) framework[44], to analyze the importance of features across different mental health classes. This task aims to identify a list of key terms, one list for each subreddit/class, that contribute to the models' decision-making process, taking into account the measured importance of each term calculated across different folds.

We use LIME instead of SHAP for two practical reasons. Firstly, recent benchmarks on transformer models for text classification demonstrate that, while SHAP provides attributions at the level of individual tokens that are comparable in quality to those of LIME, it requires 10–40 times more runtime and GPU memory because it must sample permutations for each input token[42]. Secondly, our analysis requires us to generate explanations for 10% of the test set in each of the five cross-validation splits (approximately 7,600 posts of up to 256 tokens each), followed by submodular pick aggregation. Under these conditions, SHAP would require several days of computation on a high-end GPU, whereas LIME takes under four hours with no loss of local fidelity for the mental

Rank	Topic	Total Count	Percentage
1	r/AskReddit	647595	0.43%
2	r/marvelrivals	548000	0.36%
3	r/AITAH	540780	0.36%
4	r/NoStupidQuestions	530502	0.35%
5	r/AmIOverreacting	501090	0.33%
6	r/teenagers	487369	0.32%
7	r/politics	482609	0.32%
8	r/mildlyinfuriating	447357	0.30%
9	r/Market76	423161	0.28%
10	r/worldnews	412381	0.27%

Table 1. Top 10 Most Common Subreddit Topics from wenknow/reddit_dataset_44.

health domain. We therefore adopt LIME as a widely used, tractable explainer that balances interpretability with computational feasibility[24].

We want the explanations to be generated from data that the models have never seen, so in LIME we input each fine-tuned model and the corresponding test dataset (see the second layer from the bottom in Fig. 2). The choice to have more than one test dataset in LIME's input was to increase its generalization ability, as it can analyze different linguistic patterns. Considering that LIME with submodular pick has a high computational cost, we decided to analyze 10% of the test set per model.

To generate an explanation, LIME receives as input the fine-tuned model and the test data, then it proceeds as follows: First, it creates a modified version of the test data. In the case of text data, LIME changes or removes some words and observes how the model changes its prediction. As an example, if removing a word drastically changes the classification, it means that the word is strongly influencing the model's decision. LIME builds a locally interpretable model that approximates the behavior of the model only on that specific instance. After perturbing the input several times, it is possible to identify the words with the greatest influence on the model's classification.

In our setting, for each text in the test data, we consider 500 perturbed samples. Then, we apply the Submodular Pick algorithm to select 50 representative explanations, to collectively capture a diverse range of model behaviors across the data. In the original LIME paper[24], the number of perturbations varies depending on the application context: the authors report using values such as 5,000 and 15,000. Similarly, the number of global explanations selected via Submodular Pick is determined by practical constraints, with typical values ranging from 5 to 10. Our configuration of 500 perturbations and 50 explanations reflects a trade-off between computational efficiency and explanation coverage, especially given the size of our dataset and the need for qualitative validation by domain experts.

More precisely, regarding explanations, the `num_exps_desired` parameter specifies the number of representative explanations selected by the Submodular Pick algorithm to summarize the model's decision-making patterns. Rather than generating explanations for all instances, the algorithm identifies a subset that collectively captures the diversity of the model's behavior. For instance, in classifying mental health subreddits, a model might encounter terms that overlap across the classes like "stress" (common in anxiety and depression posts) and class-specific phrases like "maniac episodes" (indicative of bipolar disorder) or "fear of abandonment" (characteristic of BPD). Setting num_exps_desired=50 ensures that LIME selects posts that highlight both the shared and unique linguistic markers.

This approach balances interpretability and computational efficiency, as formalized in the SP-LIME framework[24].

At the end of the process, we obtain four lists of terms – one for each class – for each of the five folds. Each list contains words paired with their corresponding importance scores, indicating their contribution to the model's predictions for that class. To produce a single consolidated list per class, we merge the five fold-specific lists by averaging the importance scores of each word across folds. For example, for the *Anxiety* class, if the word "worried" has importance scores of 0.8, 0.7, 0.6, 0.9, and 0.75 across the five models, its final score is the average: 0.75. If a word is missing from a list, its importance score is set to zero in the averaging process, ensuring that its absence is properly reflected in the final score. Throughout, we use "stability" to mean that token-level LIME attributions are averaged across the 5 cross-validation folds, yielding per-class lexical profiles that are robust to train/test splits; these attributions are non-causal.

At this point, the words in the four lists, one for each class, are processed as follows. We lemmatize the words using the WordNetLemmatizer library (https://www.nltk.org/api/nltk.stem.WordNetLemmatizer.html?highlight=wordnet) to reduce words such as "running" and "run" to the base form while preserving their linguistic meaning. Unlike stemming, which reduces words to their root form, lemmatizing converts them to the dictionary form. This results in a more accurate and interpretable representation of words. For example, the words "dogs" and "dog" would both be lemmatized to "dog", and "running" would be lemmatized to "run". It is important to note, however, that lemmatization treats verbs, nouns and adjectives separately. For example, the adjective "depressed" is lemmatized to "depressed", but it is not aggregated with the noun "depression", as the lemma is different and retains its meaning as a separate concept. We then sum the importance scores of words that share the same lemma across instances, removing numeric tokens and stop words.

To facilitate rapid and intuitive inspection of the explanatory features, we supplement the numeric rankings with word cloud visualizations. Word clouds offer a readily interpretable 'lexical field' overview that clinicians and other non-technical audiences can easily understand, while still reflecting the underlying mean LIME importance scores (font size is proportional to the log-scaled importance of the top 70 lemmas per class). For readers who prefer exact effect sizes, we provide the corresponding numerical values in Table 4, as well as a fully quantitative bar plot representation in Supplementary Figures S2.

Finally, we extend the analysis from the word level to the sentence level. We employed *spaCy* (https://spacy.io/.) to segment each post into grammatically complete units. This segmentation process decomposes complex constructions into main and subordinate clauses, resulting in shorter and more coherent phrases. Subsequently, the importance score of each segment was computed as the average of the importance scores of its constituent words. In this manner, we obtain, for each post, a set of chunk-level importance scores directly related to the classification task under consideration.

Results

Classification performance

The classification task involves classifying Reddit posts from four mental health-related subreddits (depression, anxiety, bipolar disorder, BPD) using BERT and MentalBERT. The metrics are aggregated across all categories. For completeness, we also evaluate GPT-4.1 mini in a zero-shot setting on the same classification task.

Model	Accuracy	Precision (Micro)	Precision (Macro)	Recall (Micro)	Recall (Macro)	F1 (Micro)	Balanced Accuracy
BERT	0.819	0.819	0.821	0.819	0.819	0.819	0.819
MentalBERT	0.822	0.822	0.826	0.822	0.822	0.822	0.822
GPT-4.1 mini	0.755	0.755	0.772	0.755	0.755	0.755	0.755

Table 2. Average Classification Metrics Across 5 Folds for BERT, MentalBERT, and ChatGPT.

Model	Accuracy	Precision (Micro)	Precision (Macro)	Recall (Micro)	Recall (Macro)	F1 (Micro)	F1 (Macro)	Balanced Accuracy
BERT	0.952	0.952	0.952	0.952	0.952	0.952	0.952	0.952
MentalBERT	0.954	0.954	0.955	0.954	0.954	0.954	0.954	0.954

Table 3. Average Classification Metrics for BERT and MentalBERT in Binary Classification of Mental Health vs. Generic Subreddit Posts.

The results are shown in Table 2. MentalBERT slightly outperforms Base BERT across all the reported metrics, though the gains are not drastic.

The best-performing model achieved an accuracy of 0.822, representing a $+0.3$ percentage point improvement (approximately 0.37% relative gain) over the second-performing model (accuracy of 0.819). The observed gain is consistent with prior studies comparing domain-adapted checkpoints to Base-BERT on social-media mental-health tasks, where absolute improvements typically remain below one percentage point once extensive fine-tuning is applied[14,26].

This small uplift is expected for two reasons: First, Reddit mental health posts still rely heavily on everyday English vocabulary that Base BERT already models well; domain-specific terms (e.g., "lamictal," "hypomania") form only a minor portion of the token distribution, so the additional pretraining in MentalBERT yields diminishing returns. Second, our fine-tuning set is large (152k posts) and balanced, allowing Base BERT to adapt effectively to the task; prior work shows that domain checkpoints confer larger benefits mainly when downstream data are scarce or highly technical[26,45].

Table 2 also reports the performance of the LLM. Both BERT variants outperformed GPT-4.1 mini across all evaluated metrics, with MentalBERT achieving the best overall results, confirming its suitability for this task. The interested reader can refer to Fig. 8 in Supplementary Materials (S4) for the confusion matrices of the LLM. Because the evaluation set is perfectly balanced across four classes (25% each), a uniform-random classifier has expected accuracy $1/4 = 25\%$. Under perfect balance, an always-predict-the-most-frequent-class baseline is also 25%[38,46].

Classification performance for the binary classification task

Table 3 presents the performances achieved by BERT and MentalBERT for distinguishing between mental health-related posts and generic posts. For a detailed breakdown, the interested reader is referred to the confusion matrices in Figures 9 and 10 in Supplementary Materials (S4).

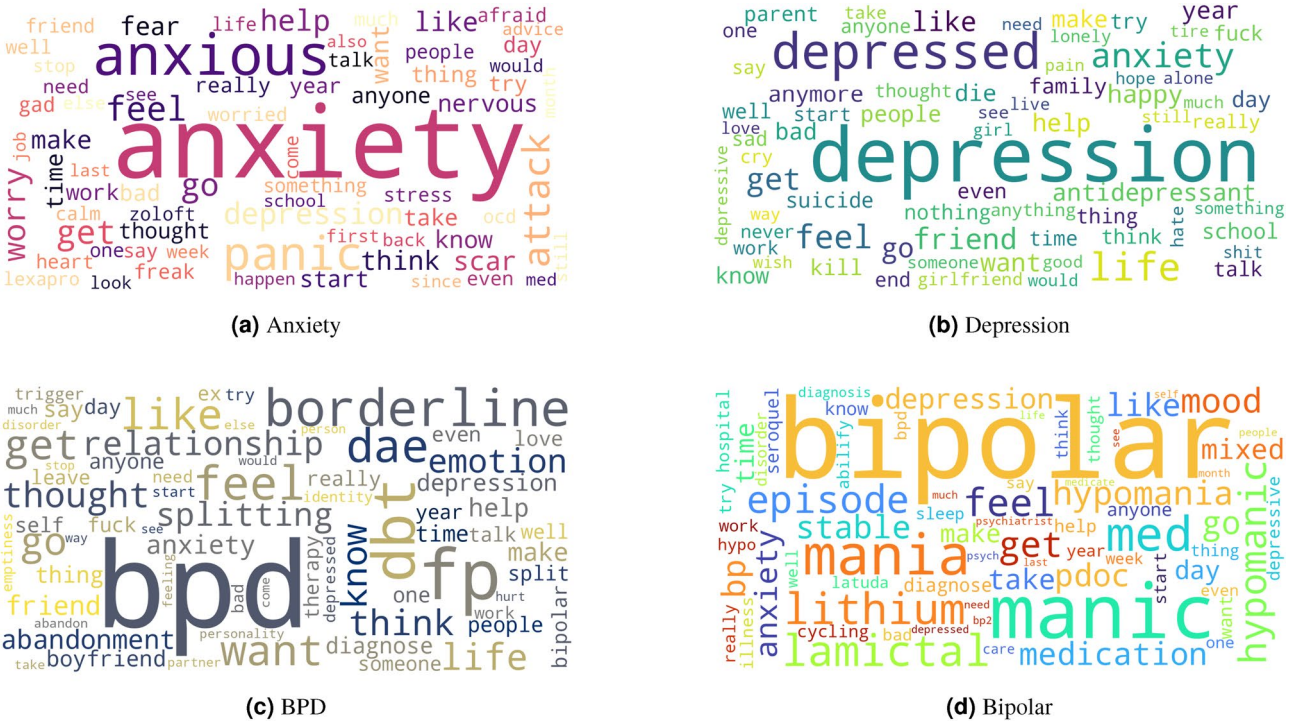
Explainability results

Explainability analysis identifies key linguistic features that contribute to the classifier's decisions. In the following, we show the results of the analysis on MentalBERT because, although its classification performance is similar to that of BERT, it has been specifically pre-trained on mental health-related data, making it the most relevant model for interpreting linguistic patterns in this context.

By aggregating word-level explanations across all folds (as described in Sect. "Binary classification: mental health vs other topics"), we identify the most influential terms that drive predictions for each class. Specifically, for each word, the average importance score across the five folds is calculated, with missing words in specific folds assigned a score of zero. The top words with the highest average importance scores highlight both class-specific patterns and overlapping terms across folds. For example, words such as "thought" and "feel" are common to all classes, whereas words such as 'mania' and 'panic' are specific to bipolar and anxiety, respectively.

In general, the results show that the most influential terms are semantically relevant to their respective classes, in line with the known symptomatology of each disorder. For example, in the anxiety class (Fig. 3a), terms such as "panic", "attack", and "anxiety" dominate. These reflect the core features of anxiety disorders,including intense worry, panic attacks, and physical manifestations such as hypervigilance and fear responses[47]. In the depression class (Fig. 3b), key words include "depression", "life", "help", and "feel", which correspond to hallmark symptoms of major depressive disorder such as persistent sadness, anhedonia, and emotional distress[48]. Many terms are present in both the anxiety and depression subreddits, suggesting common expressions of distress for people posting in these communities.

The BPD class (Fig. 3c) is characterized by words such as "abandonment", "relationship", and "fp (favorite person)", which are central to borderline personality disorder. These reflect the emotional dysregulation and fear of rejection that often characterize the condition[49,50].



(a) Anxiety **(b)** Depression

(c) BPD **(d)** Bipolar

Fig. 3. Each cloud shows the 70 highest-ranked lemmas for that class, ordered by their mean LIME importance across five cross-validation folds (absent terms receive importance = 0). Font size is proportional to the log-scaled importance; colors have no semantic meaning. The precise ranking (top-20) is provided in Table 4, and bar-plots with the exact values of the importance score appear in Supplementary Fig. S2.

Finally, the bipolar class (Fig. 3d) includes words such as "mania", "manic", "stable", and "lamictal", capturing the mood instability and treatment context associated with bipolar disorder[51,52]. Bipolar-related posts indicate the alternation between manic episodes (characterized by high energy, impulsivity, and reduced need for sleep) and depressive episodes (characterized by low energy, fatigue and feelings of hopelessness).

These linguistic patterns match the clinical descriptions of each condition, reinforcing the validity of the model's classifications. Table 4 shows, for each mental condition, the top 20 words, ordered by their average importance score, identified by the explainability analysis.

We perform a further analysis to verify that the most important words (Fig. 3, Table 4) also have the greatest influence on classification at the post level. In the case of correctly classified posts, this correspondence is precisely observed. Figure 4ashows an example of a post that was correctly classified as anxiety-related.

By contrast, the mistakes reveal two systematic errors: (i) anxiety-depression confusion, and (ii) borderline personality disorder-bipolar disorder confusion. In the former, anxiety posts are categorized as depression when the text reflects hopelessness or suicidality, but never mentions somatic cues of arousal, such as panic or palpitations. Conversely, depression posts are mistaken for anxiety posts when they include references to panic but little sustained description of mood. Both patterns reflect the shared "negative-affect" core of the two conditions and the lack of longitudinal context in a single post[53,54]. For the second error, BPD posts are mislabelled as bipolar when they highlight rapid mood changes or mention mood-stabilizing drugs, such as lamotrigine or lithium. The opposite error occurs when bipolar posts focus on intense relationship conflict or self-harm and do not specify episode duration or rhythmicity. This ambiguity echoes clinical findings on the overlap between the two disorders[55,56]. In Supplementary Material, S5 and S6, we show some examples of correctly and erroneously classified posts.

In addition to the token-level plot in Fig. 4a, the Fig. 4bpresents a sentence-level explanation of the same post. We segment the text into clauses with spaCy and assign each clause the mean of its tokens' LIME attributions. Clauses that support the predicted label are shown in green, those that oppose it in red; darker shades indicate larger magnitude. This complements the word-level display by summarizing how evidence is distributed across phrases, not only individual words (Methods §2.4).

Discussion

This study examined linguistic patterns in mental health discourse on Reddit. The Transformer-based models BERT and MentalBERT were used to classify posts from four subreddits, each covering anxiety, depression, borderline personality disorder, and bipolar disorder, as belonging or not belonging to these categories.

We formulated three research questions at the beginning of the paper, and we now provide answers to these questions.

Anxiety	
Word	**Importance**
anxiety	67.56
anxious	11.87
panic	5.71
attack	3.77
feel	3.72
get	3.17
go	2.98
worry	2.62
depression	2.58
like	2.33
fear	2.13
help	2.10
think	1.92
want	1.87
make	1.83
gad	1.78
know	1.77
take	1.64
start	1.59
time	1.58
Depression	
Word	**Importance**
depression	31.46
depressed	9.68
life	5.30
anxiety	4.36
feel	4.32
get	4.19
friend	3.55
like	3.23
go	3.15
want	2.74
bad	2.45
help	2.38
make	2.31
happy	2.23
year	2.10
talk	2.05
know	2.03
family	1.95
thing	1.91
anymore	1.90
Borderline Personality Disorder (BPD)	
Word	**Importance**
bpd	34.74
fp	6.37
dbt	4.43
borderline	2.91
dae	2.83
feel	2.56
like	2.26
get	2.17
want	2.06
friend	1.81

Continued

Borderline Personality Disorder (BPD)	
Word	**Importance**
go	1.79
anxiety	1.79
know	1.55
depression	1.52
think	1.42
relationship	1.39
life	1.35
emotion	1.34
make	1.33
thing	1.26
Bipolar Disorder	
Word	**Importance**
bipolar	24.69
manic	8.36
mania	4.39
lamictal	3.93
med	3.70
lithium	2.91
hypomanic	2.82
episode	2.50
feel	2.44
get	2.30
go	1.97
hypomania	1.96
stable	1.81
depression	1.72
mood	1.67
take	1.56
anxiety	1.54
like	1.50
pdoc	1.48
diagnose	1.41

Table 4. The top 20 words identified in the explainability analysis for the different mental health conditions, ranked by mean importance score.

In response to RQ1: How accurately can transformer-based text classifiers distinguish among the four mental-health discourse categories represented by the subreddits r/Depression, r/Anxiety, r/BPD and r/Bipolar? Our results show that transformer models capture the linguistic nuances of mental health discussions, with MentalBERT achieving slightly better classification performance than BERT (for example, we obtained an F1 of 0.822 with MentalBERT and 0.819 with BERT). In addition, the integration of explainability techniques provides insight into the most influential features driving model decisions, thereby improving the interpretability of results and confidence in the adoption of classification tools by clinicians. A contribution of this study is the integration of explainability techniques to improve model transparency. While deep learning models achieve strong classification performance, their lack of explainability remains a barrier to adoption in clinical applications[19,22].

In response to RQ2: Which lexical cues drive the classification, and do explanation methods reveal features that are shared across, or specific to, particular conditions? Our analysis highlights the importance of transparency in machine-learning-based mental health classification. Using LIME, we identified the key linguistic features that contribute to the model's predictions, which is critical not only for verifying the model's validity, but also for ensuring ethical use of deep-learning tools for mental health applications. Studies have shown that clinicians are more likely to use machine-learning tools when they align with established diagnostic criteria and when the meaning of features can be explicitly understood[6]. Our approach follows this principle, making model decisions more transparent and accessible to researchers and healthcare professionals. Our findings confirm that specific linguistic markers are tied to specific disorders. For example, words such as "panic" and "attack" are strongly associated with anxiety[14]. Similarly, our model identifies distinctive language markers for BPD, such as discussions of abandonment and attachment issues, which are consistent with the clinical literature[2]. In the case of bipolar disorder, key terms related to mood swings and mania are highly influential in classification, supporting previous research linking social media discourse to mood instability[35].

Successfully got married and went on myhoneymoonwith no major panic attacks! Hi reddit, I have found comfort in this community of fellow anxious people so I thought I'd share an update on my anxiety. I've had anxiety my entire life and it has become debilitating at times. It makes me scared to do things I want to do, such as travel. This year is a big year for me - I just got married last Saturday and I just got back from my honeymoon in California. Those two things alone were enough to keep me up at night anxiously worrying months before. BUT - I did it. It was a perfect trip and I'm so glad I didn't let my anxiety ruin it. I'm gonna list a few things that have helped meimmensely and allowed me to enjoy my time this week. \-First, I went to my doctor and got a prescription for 10 0.5mg Klonopin pills. My plan was to use them for times of extreme anxiety (flying) only. These were very helpful to me. I was really anxious on my wedding morning and almost couldn't get out of my car to go to the hair salon, so I popped half a pill. About an hour into getting my hair and makeup done, I realized I was fine. The rest of the day was absolutely beautiful and I had no social anxiety and didn't feel any panic. I also took a whole pill the morning of my flight to California. I was perfectly relaxed on the plane and was pleasantly sleepy. I have no intentions of taking these pills on a regular day - just when I'm doing something I know will send me into a panic.\-Secondly is just simply not letting my anxiety stop me from doing things. For example, I was very anxious to get on our first flight and almost found myself saying "I can't do this." Had I actually not done it and gone home, I would have been even more scared of flying in the future. But, I did it anyway and I'm so glad I did. Once I was in the air, I turned on The Office and played games on my phone and before I knew it, we were landing. Now I'm excited to fly again, and am already planning my next trip. I truly believe that exposure is helping my anxiety. \-Third, I have been taking CBD oil (both in oil form and through a vape pen) consistently. This has decreased my daily anxiety a noticeable amount. \-Fourth, I cut out coffee for the most part and try to stay away from refined sugar. I noticed a correlation between my anxiety levels and what I was eating/drinking, so I just started cutting things out of my diet. I still eat sugar and drink an occasional latte, but it doesn't give me the same anxiety as when I was eating/drinking these things daily. I just really wanted to post and let people know that while it may seem like a never-ending battle, anxiety can definitely be helped. I am slowly getting better and getting to know myself in the process. Much love to my fellow anxiety-ridden folks.

(a) An example post classified correctly as belonging to the Anxiety category. Green indicates the model's confidence for the Anxiety class, whereas red indicates tokens whose contribution decreases the probability of being assigned to the correct class. The character ** in the text represents a "zero-width space," a non-printing character used in HTML environments like Reddit for formatting purposes. It should not be considered a meaningful part of the text and has no impact on the model's behavior.

(b) The same post as above, with sentence-level local explanation. The post is segmented into clauses with spaCy; each clause is colored by the mean of its tokens' LIME attributions. Green indicates clauses that increase the predicted class probability; red indicates clauses that decrease it; darker shades denote larger magnitude. Texts without a background color indicate that they had no impact on the probability of the predicted class.

Fig. 4. Local explanations for a single Reddit post. (a) Token-level attributions per word. (b) Sentence-level attributions by clause (mean of token attributions). Colors: green = positive, red = negative; intensity encodes magnitude and is normalized within the post.

We build upon several previous studies on the interpretability of mental health text classification. For example, Kim et al.[10] used CNN saliency maps and LIWC categories to highlight generic distress words such as "depression", "feel", "help", "life", and "suicide". All five of these words appear in the top 20 LIME scores for the depression class (Table 4). Similarly, Tadesse et al.[18] reported that first-person pronouns and negative emotion terms dominate depression posts. Our model surfaces these patterns across both the Depression and Anxiety classes. More recent explainability work employing SHAP or LIME, such as that by Jo et al.[31] for depression

prediction and Alghazzawi et al.[32] for suicidal ideation detection, confirms the importance of the terms "hopeless", "worthless" and "panic", which again rank highly in our aggregated scores.

What distinguishes our study is its balanced four-class design, which reveals markers that become uniquely significant only when the model must discriminate between closely related disorders. "Panic", "attack", and "hyperventilation" for r/Anxiety; "lithium", "hypomania", and "mania" for r/Bipolar; and "abandonment", "DBT" (Dialectical Behaviour Therapy), and "FP" (Favourite Person) for r/BPD. The analysis reveals both shared and condition-specific lexical patterns in mental health discourse.

In response to RQ3: How does the proposed multi-class transformer pipeline compare, in predictive quality, with a zero-shot large-language-model baseline? Our results demonstrate that transformer-based models outperform zero-shot GPT-4.1 in terms of classification performance. Although GPT-4.1 is a powerful general-purpose language model, our fine-tuned MentalBERT model is better aligned to the mental health domain and is more computationally efficient. These findings emphasize the importance of domain adaptation and task-specific fine-tuning in the development of models for mental health applications.

It is important to clarify that our study does not aim to diagnose mental health conditions but rather to analyze how users express mental health concerns within online communities. Unlike clinical settings where an individual is assessed by a health professional, our dataset consists only of textual posts tagged by the subreddit in which they are posted. Therefore, our findings reflect patterns in community discourse rather than confirmed diagnoses. Contrary to clinical diagnostic tools, our model does not attempt to classify users by psychiatric criteria; it determines whether a post's language aligns with the thematic focus of a given subreddit, providing discourse-level insights rather than medical inference. This distinction is crucial because online mental health communities are not intended for diagnosing individuals. Rather, they serve as indicators of how mental health issues are publicly expressed[12].

The results obtained have practical implications. For example, they could inform the creation of more specialized online resources, prevent unexpected topics from being discussed in inappropriate communities and define new communication strategies to improve dialogue between clinicians and people experiencing mental distress. To facilitate real-world use, we have made the five fine-tuned MentalBERT checkpoints publicly available on OSF (https://osf.io/mgveh/). There is one checkpoint for each cross-validation fold. As they adhere to the native HuggingFace format, models can be loaded with just a few lines of Python code and run on an everyday laptop CPU; no specialized hardware or cloud service is required. These resources enable at least two immediate, human-in-the-loop applications: (i) dashboard tools that alert subreddit moderators when discussion drifts off topic; (ii) rapid labeling of large Reddit datasets to support future mental-health NLP research. The models are operational, but they still require human oversight to validate outputs and mitigate potential harm.

Our contribution is complementary to prior studies. Reviews such as[14] emphasize both the promise of deep learning and the need for interpretability; our fold-aggregated post-hoc lexical profiles directly support this aim. Retrieval-Augmented Generation systems constitute a promising avenue for tackling the task addressed in this study, which led us to explore their applicability. Nevertheless, recent analyses indicate that their classification performance remains below that achieved by transformer-based models[27] (Mental-health subreddits classification with LLaMA-3). Future work may further investigate the evolution of this technology and evaluate its potential to reach or surpass state-of-the-art performance on the task considered herein.

Certain limitations should be acknowledged. Firstly, although Reddit is a valuable source of discourse on mental health, it may not be representative of all individuals with these conditions, particularly those who do not participate in online forums. While the dataset is balanced across categories, it does not account for variations in demographic factors, such as age, gender, and geographical location, that may influence linguistic expressions[6]. Also, our labels index community membership, not diagnoses. While comorbidity at the person level is common, a multi-label formulation of the classification task would require reliable multi-label ground truth (e.g., expert annotation or explicit multi-community tags), which we do not have in this corpus. We also wanted to point out that sentence-level displays (Fig. 4b) are exploratory and intended for qualitative inspection only; clinical validation of phrase-level explanations by domain experts is left for future work.

Future research can extend this work by incorporating data from multiple platforms to compare linguistic trends across different social media environments. For example, while Reddit is widely used in the US, it is important to explore whether the relationship between linguistic patterns and discussion categories is also found on more widely used platforms in Europe and other non-Anglophone countries. Additionally, analyzing data from different platforms could provide insights into how people express mental health issues in different cultural contexts, highlighting potential variations in discourse. Multimodal data, such as videos and images, as well as user interactions, could further enhance our understanding of mental health discussions and the diverse ways people engage with these topics online.

In conclusion, this study contributes to the growing body of research on deep-learning-based detection of mental health conditions by demonstrating the effectiveness of transformer-based classifiers and highlighting the importance of explainability. Our findings highlight the potential of social media data for mental health monitoring, while emphasizing the need for transparent, interpretable and ethically sound AI applications in this area. By providing a robust framework for analyzing mental health discussions on social media using fine-tuned transformer models, we offer a balanced approach that ensures both accuracy and interpretability. Unlike previous studies that focus solely on classification performance, our research prioritizes transparency, making deep-learning-based mental health insights more accessible to researchers and healthcare professionals. Ultimately, by analyzing online discussions, we contribute valuable insights into mental health discourse, supporting efforts to identify trends in mental well-being and inform future research in computational psychiatry. This, in turn, may contribute to spreading awareness about mental health conditions and overcoming the stigma that still persists towards psychiatric disorders[57]. Nevertheless, we caution that the lexical cues identified by LIME are context-dependent and should never be used in isolation for automated flagging or triage. Any

practical deployment must involve human oversight to review model outputs and prevent erroneous labeling or unintended stigmatization.

Data availability

The dataset analysed during the current study is not publicly available because it was obtained directly from the authors of a previously published study (Kim, J., Lee, J., Park, E. et al. A deep learning model for detecting mental illness from user content on social media. Sci. Reports 10, 11846, doi:10.1038/s41598-020-68764-y (2020)). The data were shared upon request for research purposes and are therefore subject to the original authors' terms of use. They are available from the corresponding author of that study on reasonable request. We provide five MentalBERT models, each fine-tuned on a different fold of our cross-validation setup. Models available at https://osf.io/mgveh/?view_only=eec4c5188a0645efafa09f863627e427.

Received: 7 April 2025; Accepted: 8 January 2026
Published online: 31 January 2026

References

1. Liu, J., Ning, W., Zhang, N., Zhu, B. & Mao, Y. Estimation of the global disease burden of depression and anxiety between 1990 and 2044: An analysis of the global burden of disease study 2019. *Healthcare* **12**, https://doi.org/10.3390/healthcare12171721 (2024).
2. Söderholm, J. J., Socada, J. L., Rosenström, T. H., Ekelund, J. & Isometsä, E. Borderline personality disorder and depression severity predict suicidal outcomes: A six-month prospective cohort study of depression, bipolar depression, and borderline personality disorder. *Acta Psychiatr. Scand.* **148**, 222–232. https://doi.org/10.1111/acps.13586 (2023).
3. Zhu, X. & Lv, Q. Sex difference in incidence of bipolar and anxiety disorders: findings from the global burden of disease study 2021. *medRxiv* https://doi.org/10.1101/2024.10.27.24316200 (2024). https://www.medrxiv.org/content/early/2024/10/29/2024.10.27.24316200.full.pdf.
4. Cheng, Y. et al. The burden of depression, anxiety and schizophrenia among the older population in ageing and aged countries: an analysis of the global burden of disease study 2019. *Gen. Psychiatry* **37**, https://doi.org/10.1136/gpsych-2023-101078 (2024).
5. Patel, V. et al. Addressing the burden of mental, neurological, and substance use disorders: Key messages from disease control priorities, 3rd edition. *The Lancet* **391**, 1672–1685, https://doi.org/10.1016/S0140-6736(18)30491-6 (2018).
6. Corrigan, P. W., Druss, B. G. & Perlick, D. A. The impact of mental illness stigma on seeking and participating in mental health care. *Psychol. Sci. Public Interest* **15**, 37–70. https://doi.org/10.1177/1529100614531398 (2014).
7. Kuss, D. J. & Griffiths, M. D. Social networking sites and addiction: Ten lessons learned. *Int. J. Environ. Res. Public Heal.* **14**, 311. https://doi.org/10.3390/ijerph14030311 (2017).
8. Mansoor, M. A. & Ansari, K. H. Early detection of mental health crises through artificial-intelligence-powered social media analysis: A prospective observational study. *J. Pers. Medicine* **14**, 958. https://doi.org/10.3390/jpm14090958 (2024).
9. Orehek, E. & Human, L. J. Self-expression on social media: Do tweets present accurate and positive portraits of impulsivity, self-esteem, and attachment style?. *Pers. Soc. Psychol. Bull.* **43**, 60–70. https://doi.org/10.1177/0146167216675332 (2017).
10. Kim, J. et al. A deep learning model for detecting mental illness from user content on social media. *Sci. Reports* **10**, 11846. https://doi.org/10.1038/s41598-020-68764-y (2020).
11. Chan, G., Fung, M., Warrington, J. & Nowak, S. Understanding health-related discussions on reddit: Development of a topic assignment method and exploratory analysis. *JMIR Form. Res.* **9**, e55309. https://doi.org/10.2196/55309 (2025).
12. Chandrasekaran, R., Mehta, V., Valkunde, T. & Moustakas, E. A mental health informatics perspective on social media data during the covid-19 pandemic: Systematic review. *J. Med. Internet Res.* **23**, e27848. https://doi.org/10.2196/27848 (2021).
13. Morini, V., Sansoni, M., Rossetti, G., Pedreschi, D. & Castillo, C. Participant behavior and community response in online mental health communities: Insights from reddit. *Comput. Hum. Behav.* **165**, 108544. https://doi.org/10.1016/j.chb.2024.108544 (2025).
14. Zhang, T., Schoene, A. M., Ji, S. & Ananiadou, S. Natural language processing applied to mental illness detection: A narrative review. *npj Digital Medicine* **5**, 46, https://doi.org/10.1038/s41746-022-00589-7 (2022).
15. Le Glaz, A. et al. Machine learning and natural language processing in mental health: systematic review. *J. medical Internet research* **23**, e15708 (2021).
16. Saxena, R. R. Applications of natural language processing in the domain of mental health. *Authorea Prepr.* (2024).
17. Kim, J. et al. Machine learning for mental health in social media: bibliometric study. *J. Med. Internet Res.* **23**, e24870 (2021).
18. Tadesse, M. M., Lin, H., Xu, B. & Yang, L. Detection of depression-related posts in reddit social media platform using deep learning. *BMC Psychiatry* **19**, 413. https://doi.org/10.1109/ACCESS.2019.2909180 (2019).
19. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215. https://doi.org/10.1038/s42256-019-0048-x (2019).
20. Wiegreffe, S. & Pinter, Y. Attention is not not explanation. In Inui, K., Jiang, J., Ng, V. & Wan, X. (eds.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 11–20, https://doi.org/10.18653/V1/D19-1002 (Association for Computational Linguistics, 2019).
21. Vig, J. A multiscale visualization of attention in the transformer model. In Costa-jussà, M. R. & Alfonseca, E. (eds.) *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, 37–42, https://doi.org/10.18653/V1/P19-3007 (Association for Computational Linguistics, 2019).
22. Arrieta, A. B. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities, and challenges toward responsible ai. *Inf. Fusion* **58**, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012 (2020).
23. Caruana, R. et al. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730, https://doi.org/10.1145/2783258.2788613 (2015).
24. Ribeiro, M. T., Singh, S. & Guestrin, C. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144, https://doi.org/10.1145/2939672.2939778 (2016).
25. Lundberg, S. M. & Lee, S. A unified approach to interpreting model predictions. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 4765–4774 (2017).
26. Ji, S., Pan, S., Zhang, H. & Cambria, E. Mentalbert: A pretrained transformer for mental healthcare. *arXiv preprint* arXiv:2205.07643 (2022).
27. Kermani, A., Perez-Rosas, V. & Metsis, V. A systematic evaluation of llm strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. rag. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)* (2025).

28. Garg, M. et al. Cams: An annotated corpus for causal analysis of mental health issues in social media posts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, 6387–6396 (European Language Resources Association, Marseille, France, 2022).
29. Saxena, C., Garg, M. & Ansari, G. Explainable causal analysis of mental health on social media data. In *International Conference on Neural Information Processing (ICONIP 2022)*, 172–183 (Springer, Cham, 2022).
30. S Band, S. et al. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked* **40**, 101286, https://doi.org/10.1016/j.imu.2023.101286 (2023).
31. Jo, A. A., Raj, E. D., Vino, A. S. & Menon, P. V. Exploring explainable ai for enhanced depression prediction in mental health. In *2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC)*, 1–7 (IEEE, 2024).
32. Alghazzawi, D., Ullah, H., Tabassum, N., Badri, S. K. & Asghar, M. Z. Explainable ai-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique. *Sci. Reports* **15**, 1111 (2025).
33. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805 (2019).
34. Tsoumakas, G. & Katakis, I. Multi-label classification: An overview. *Int. J. Data Warehous. Min.* **3**, 1–13. https://doi.org/10.4018/jdwm.2007070101 (2007).
35. Oliva, V. et al. Bipolar disorders: an update on critical aspects. *Lancet Reg. Heal. - Eur.* **48**, 101135. https://doi.org/10.1016/j.lanepe.2024.101135 (2024).
36. Gunderson, J. G., Herpertz, S. C., Skodol, A. E., Torgersen, S. & Zanarini, M. C. Borderline personality disorder. *Nat. Rev. Dis. Primers* **4**, 18029. https://doi.org/10.1038/nrdp.2018.29 (2018).
37. Hawke, L. D., Parikh, S. V. & Michalak, E. E. Stigma and bipolar disorder: a review of the literature. *J. Affect. Disord.* **150**, 181–191. https://doi.org/10.1016/j.jad.2013.05.030 (2013).
38. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
39. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York, NY, USA, 2009), 2nd edn.
40. Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C. & Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186, https://doi.org/10.18653/V1/N19-1423 (Association for Computational Linguistics, 2019).
41. Ji, S. et al. Mentalbert: Publicly available pretrained language models for mental healthcare. In Calzolari, N. *et al.* (eds.) *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, 7184–7190 (European Language Resources Association, 2022).
42. Fantozzi, P. & Naldi, M. The explainability of transformers: Current status and directions. *Comput.* **13**, 92. https://doi.org/10.3390/COMPUTERS13040092 (2024).
43. Micikevicius, P. et al. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* (OpenReview.net, 2018).
44. Longo, L. et al. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Inf. Fusion* **106**, 102301, https://doi.org/10.1016/j.inffus.2024.102301 (2024).
45. Lyu, D., Wang, X., Chen, Y. & Wang, F. Language model and its interpretability in biomedicine: A scoping review. *iScience* **27**, 109334, https://doi.org/10.1016/j.isci.2024.109334 (2024).
46. Murphy, K. P. *Machine Learning: A Probabilistic Perspective* (MIT Press, 2012).
47. Horowitz, B. E. Psychological perspectives: Anxiety disorders: Identification and intervention. In *Communication apprehension: Origins and management*, 1st edn. chap. Psychological perspectives: Anxiety disorders: Identification and intervention (Singular/Thomson Learning, 2002).
48. Bhowmik, D., Kumar, S., Srivastava, S., Paswan, S. & Dutta, A. Depression-symptoms, causes, medications and therapies. *Pharma Innov.* **1**, 32–45 (2012).
49. Aaronson, C. J. et al. Comparison of attachment styles in borderline personality disorder and obsessive-compulsive personality disorder. *Psychiatr. Q.* **77**, 69–80. https://doi.org/10.1007/s11126-006-7962-x (2006).
50. Brüne, M. Borderline personality disorder: Why 'fast and furious'?. *Evol. Medicine, Public Heal.* **52–66**, 2016. https://doi.org/10.1093/emph/eow002 (2016).
51. Solé, E., Garriga, M., Valentí, M. & Vieta, E. Mixed features in bipolar disorder. *CNS Spectrums* **22**, 134–140. https://doi.org/10.1017/S1092852916000869 (2017).
52. Zun, L. S. & Nordstrom, K. Mood disorders. In Rosen's Emergency Medicine: Concepts and Clinical Practice (eds Rosen, P. et al.), chap. 101, 1346–1352.e1 (Elsevier, 2020).
53. Clark, L. A. & Watson, D. Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *J. Abnorm. Psychol.* **100**, 316–336. https://doi.org/10.1037/0021-843X.100.3.316 (1991).
54. Rutter, L. A. et al. Negative affect variability differs between anxiety and depression on social media. *PLOS ONE* **19**, e0272107. https://doi.org/10.1371/journal.pone.0272107 (2024).
55. Zimmerman, M. & Morgan, T. A. The relationship between borderline personality disorder and bipolar disorder. *Dialogues Clin. Neurosci.* **15**, 155–169. https://doi.org/10.31887/DCNS.2013.15.2.mzimmerman (2013).
56. Nazari, M.-J. et al. A machine-learning approach for differentiating bipolar ii disorder and borderline personality disorder using eeg and cognitive abnormalities. *PLOS ONE* **19**, e0303699. https://doi.org/10.1371/journal.pone.0303699 (2024).
57. Ahad, A. A., Sanchez-Gonzalez, M. & Junquera, P. Understanding and addressing mental health stigma across cultures for improving psychiatric care: A narrative review. *Cureus* **15**, e39549. https://doi.org/10.7759/cureus.39549 (2023).

Acknowledgements

Work partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - #NGEU; by project "Exploring Social Media for understanding Borderline Personality Disorder" under the PhD in Cognitive Computational and Social Neuroscience, approved by the Joint Research Ethics Committee of Scuola Normale Superiore and Scuola Superiore Sant'Anna Pisa with resolution no. 48 of 20 November 2024.

Author contributions

All the authors designed the research, wrote, and reviewed the manuscript. I.S.R. and J.B. performed the research and analyzed data. I.S.R, M.P. and F.Pa. contributed to the conceptualization of the work and to the study design. F.Pi. contributed to the methodology and to the explainability analyses. E.R. and P.P. contributed to funding acquisition. M.P. coordinated the whole team.

Funding
Work partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - #NGEU; by project "Exploring Social Media for understanding Borderline Personality Disorder" under the PhD in Cognitive Computational and Social Neuroscience and project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-35918-3.

Correspondence and requests for materials should be addressed to I.S.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.