

Algorithmic Classification of Psychiatric Disorder Related Spontaneous Communication Using Large Language Model Embeddings: Algorithm Development and Validation

Ryan Allen Shewcraft, John Schwarz, Mariann Micsinai Balan

Submitted to: JMIR AI
on: October 09, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 4

Supplementary Files..... 14

 Figures 15

 Figure 1..... 16

 Figure 2..... 17

 Figure 3..... 18

 Figure 4..... 19

Algorithmic Classification of Psychiatric Disorder Related Spontaneous Communication Using Large Language Model Embeddings: Algorithm Development and Validation

Ryan Allen Shewcraft¹ PhD; John Schwarz¹ PhD; Mariann Micsinai Balan¹ PhD

¹ Bristol Myers Squibb Lawrence Township US

Corresponding Author:

Ryan Allen Shewcraft PhD

Bristol Myers Squibb

Lawrence Township

US

Abstract

Background: Language, which is a crucial element of human communication, is influenced by the complex interplay between thoughts, emotions, and experiences. Psychiatric disorders have an impact on cognitive and emotional processes, which in turn affect the content and way individuals with these disorders communicate using language. The recent rapid rise in performance of large language models (LLMs) suggests that leveraging them for quantitative analysis of language usage has the potential to become a useful method for objectively diagnosing and distinguishing between various psychiatric disorders.

Objective: Here, we present a novel application of LLMs to analyze patterns of language usage from spontaneous communication related to psychiatric disorders.

Methods: We utilized embeddings derived from the GRIT-7b LLM for posts originating from subreddits dedicated to seven common conditions: schizophrenia (Scz), borderline personality disorder (BPD), depression (Dep), attention deficit hyperactivity disorder (ADHD), anxiety (Anx), post-traumatic stress disorder (PTSD) and bipolar disorder (Bip). Using these embeddings, we trained a cross-validated, multi-class XGBoost classifier to label which subreddit each post came from.

Results: The 10-fold cross-validated classifier has weighted average precision, recall, F1, and accuracy scores were of 0.73, 0.68, 0.70 and 0.73, respectively, for the multi-class objective. In a one-versus-rest task, individual category AUCs ranged from 0.89-0.97, with an micro-average one-versus-rest AUC of 0.95.

Conclusions: Our results suggest that LLMs can provide valuable insights into the linguistic patterns that distinguish between different psychiatric disorders, and that these patterns can be used to develop more objective, efficient, and patient-centered strategies for assessment, monitoring, and research.

(JMIR Preprints 09/10/2024:67369)

DOI: <https://doi.org/10.2196/preprints.67369>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in a JMIR journal, I will be able to make the full manuscript available to all users.

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Algorithmic Classification of Psychiatric Disorder Related Spontaneous Communication Using Large Language Model Embeddings: Algorithm Development and Validation

Ryan A. Shewcraft, John Schwarz, Mariann Micsinai Balan

Abstract

Background: Language, which is a crucial element of human communication, is influenced by the complex interplay between thoughts, emotions, and experiences. Psychiatric disorders have an impact on cognitive and emotional processes, which in turn affect the content and way individuals with these disorders communicate using language. The recent rapid rise in performance of large language models (LLMs) suggests that leveraging them for quantitative analysis of language usage has the potential to become a useful method for objectively diagnosing and distinguishing between various psychiatric disorders.

Objective: Here, we present a novel application of LLMs to analyze patterns of language usage from spontaneous communication related to psychiatric disorders.

Methods: We utilized embeddings derived from the GRIT-7b LLM for posts originating from subreddits dedicated to seven common conditions: schizophrenia (Scz), borderline personality disorder (BPD), depression (Dep), attention deficit hyperactivity disorder (ADHD), anxiety (Anx), post-traumatic stress disorder (PTSD) and bipolar disorder (Bip). Using these embeddings, we trained a cross-validated, multi-class XGBoost classifier to label which subreddit each post came from.

Results: The 10-fold cross-validated classifier has weighted average precision, recall, F1, and accuracy scores were of 0.73, 0.68, 0.70 and 0.73, respectively, for the multi-class objective. In a one-versus-rest task, individual category AUCs ranged from 0.89-0.97, with an micro-average one-versus-rest AUC of 0.95.

Conclusions: Our results suggest that LLMs can provide valuable insights into the linguistic patterns that distinguish between different psychiatric disorders, and that these patterns can be used to develop more objective, efficient, and patient-centered strategies for assessment, monitoring, and research.

Keywords: Psychiatric disorders; large language models; speech; language; spontaneous communication; social media

Introduction

Psychiatric disorders encompass a diverse range of conditions affecting an individual's thoughts, emotions, and behaviors. These disorders are characterized by complex and heterogeneous symptomatology, making it difficult to establish precise diagnostic criteria and monitor disease progression over time [1]. While standardized clinical interviews and questionnaires are used, they rely on subjective assessments and can be time-consuming or insensitive to subtle changes leading to potential misdiagnosis and delayed intervention [2].

Language, as a fundamental aspect of human communication, reflects the intricate interplay between thoughts, emotions, and experiences. Quantitative analysis of language usage has emerged as a valuable tool for providing objective measures for diagnosing and differentiating between different psychiatric disorders. Studies have shown that language-based features, such as syntactic complexity,

semantic coherence, and emotional valence, can serve as reliable markers for differentiating between psychiatric disorders. For instance, individuals with schizophrenia often exhibit disturbances in their speech patterns, characterized by disorganized syntax and impaired semantic coherence [3]. Similarly, individuals with borderline personality disorder have higher levels of overall expressive language impairment, as well as decreased syntactic and lexical complexity [4].

Furthermore, quantitative analysis of language usage can aid in tracking disease progression and treatment response. Longitudinal studies have demonstrated that changes in linguistic patterns over time can be indicative of disease progression and treatment outcomes. For example, changes in language usage have been shown to correlate with changes in current depression symptoms [5]. Additionally, “tentativeness”, as measured by a higher degree of uncertainty reflected in use of language by individuals with anxiety, is correlated with quantitative levels of symptoms measure by the GAD-7 scale [6]. These findings underscore the potential of quantitative language analysis as a sensitive and objective measure for monitoring disease trajectories and treatment efficacy.

Recent advancements in large language models (LLMs) have opened up exciting possibilities for quantitative assessment of neurological diseases [7]. As a result of their transformer architecture, LLMs project strings of text (sentences, paragraphs, etc.) onto a high dimensional embedding space which represents the semantic and syntactic relationships between words and phrases. In this embedding space, linguistically similar texts are likely to geometrically co-located. Therefore, we hypothesize that, given the differences in patterns of speech by individuals across psychiatric disorders, spontaneous use of language will occupy diagnosis-specific subspaces in the LLM embedding space. To test this hypothesis, we specifically focus on utilizing embeddings derived from the GRIT-7b LLM [8] to classify posts originating from subreddits dedicated to six common conditions: schizophrenia, borderline personality disorder, depression, attention deficit hyperactivity disorder, anxiety, post-traumatic stress disorder, and bipolar disorder [9].

This investigation represents an innovative application of LLMs in the field of psychiatric disorders. By utilizing LLM embeddings to analyze spontaneous use of language in online discussion data, we hope to contribute to the development of more objective, efficient, and patient-centered strategies for assessment, monitoring, and research.

Methods

Data Collection

The dataset we used in this study was obtained from the publicly available dataset provided by Low et al. [9], which can be accessed on Zendodo. The dataset consists of posts from seven subreddits related to psychiatric disorders: r/adhd, r/anxiety, r/bipolarreddit, r/bpd, r/depression, r/ptsd, and r/schizophrenia collected between December 2018 and December 2019. These subreddits were selected to represent a diverse range of psychiatric disorders. The individual posts were inspected for regular expressions directly related to the subreddit title, and any posts containing these references were removed from the dataset (Table 1).

Table 1. Regex codes used to remove posts with potentially revealing words.

Subreddit	Regex terms for cleaning
r/adhd	adhd attention hyperact
r/anxiety	anxiety
r/bipolarreddit	bipolar
r/bpd	borderline bpd

r/depression	depress
r/pstd	ptsd post-traumatic post traumatic
r/schizophrenia	schiz

Embedding Generation

We generated embeddings using the GritLM-7B model [8] (Fig. 1). The GritLM-7B model is based on Mistral 7B [10] that was fine-tuned using both representational instruction tuning and generative instruction tuning, resulting in a model that performs at state-of-the-art levels for both generative and embedding tasks. For embedding tasks, GRITLM employs bidirectional attention followed by mean pooling of the final hidden state to generate the final representation. Model inference was performed on a single GPU (NVIDIA A10G) Amazon Web Services EC2 instance.

Figure 1: Embeddings of the input reddit posts obtained from GritLM-7B are used as features in an XGBoost classifier to predict subreddit of origin.

2-D visualization of embedding space

We began by standardizing the embedding matrix to ensure that each feature had a mean of zero and a unit variance. This standard scaling process is crucial for normalizing the data and mitigating the effects of differing scales among features. Following this, we employed Uniform Manifold Approximation and Projection (UMAP) to reduce the dimensionality of the scaled data, transforming it into a two-dimensional representation. UMAP is a powerful non-linear dimensionality reduction technique that preserves the local and global structure of the data. For our UMAP transformation, we set the number of neighbors to 50 and the minimum distance parameter to 0.05, chosen by visual inspection to optimize the balance between local and global data structure preservation. The number of neighbors parameter determines the size of the local neighborhood UMAP considers when learning the manifold structure of the data. A higher number of neighbors allows UMAP to capture more of the global structure of the data, ensuring that the overall shape and relationships between clusters are well-preserved. This is particularly important in our context, where understanding the relationships between different conditions may provide insights into which conditions have features of spontaneous communication that overlap with other conditions. The minimum distance parameter controls how tightly UMAP packs points together in the low-dimensional space. A smaller minimum distance value allows points to be closer together, which can help in preserving the local structure and making clusters more distinct. Decreasing this parameter enhances the separation between different clusters, making it easier to identify and interpret distinct groups within a label that may reflect heterogeneity of a particular condition. Finally, we visualized the resulting 2D UMAP representation by creating a scatter plot, where each point was colored according to its corresponding label, facilitating the identification of patterns and clusters within the high-dimensional data.

Classification Model

For the classification task, we employed the XGBoost algorithm using a multiclass classifier with a softmax objective function to simultaneously predict the class labels for the posts from the seven psychiatric disorder subreddits. To account for potential biases due to class imbalance, we utilized balanced class weighting, which assigns weights to each class that are inversely proportional to their frequencies.

Performance Evaluation

To evaluate the performance of the classification model, we employed 10-fold cross-validation. We calculated class-wise precision, recall, and F1 scores to assess the performance for each psychiatric

disorder associated subreddit class. Additionally, we computed macro and weighted average scores across all classes to provide an overall evaluation of the model's performance.

Results

Data Description

The data used in this study comes from seven different subreddits: r/adhd, r/anxiety, r/bipolarreddit, r/bpd, r/depression, r/ptsd, and r/schizophrenia. Following removal of posts containing text that would be revealing of subreddit (see Methods), there was a nearly 7-fold difference between the total number of posts in each subreddit. The r/depression subreddit had the greatest number of posts (11513 posts, 11483 unique users) and the r/bipolarreddit had the least number of posts (1711 posts, 1633 unique users) (Table 2). Overall, 36102 out of 37195 (97.1%) of posts were made by unique users. Two users made five posts, seven users made four posts, and 54 users made three posts. The remaining 36039 (99.8%) made only one or two posts. Furthermore, no user made posts in more than one subreddit.

Table 2: Number of posts and unique users for each subreddit following post cleaning.

Subreddit	No. posts	Unique users
r/adhd	7568	7319
r/anxiety	6391	5990
r/bipolarreddit	1711	1633
r/bpd	5849	5699
r/depression	11513	11483
r/ptsd	1954	1769
r/schizophrenia	2209	2209

Relationships between categories

We utilized the UMAP (Uniform Manifold Approximation and Projection) algorithm to visualize a two-dimensional representation of the embedding space and visualize relations between subreddits (Fig. 2). This low-dimensional visualization reveals several qualitative features of the dataset. First, posts from r/anxiety subreddit lie in the middle of the projection, neighboring all the other categories. This may suggest that the language of anxiety has some presence in all of the subreddits. Second, while some subreddits (r/ptsd, r/bipolarreddit, r/adhd, and r/schizophrenia) have distinct clusters, the other three subreddits (r/anxiety, r/bpd, and r/depression) have more overlapping point clouds, suggesting greater linguistic similarity across the latter three subreddits compared to the other categories.

Figure 2. A two-dimensional representation of posts in the high-dimensional embeddings space using the UMAP algorithm. The posts are colored by their corresponding subreddit.

Classifier performance

In addition to visualizing the subreddit categories embedding space using UMAP, we employed a 10-fold cross-validated XGBoost classifier to predict the subreddit for each post. Across all categories, the weighted average precision, recall, and F1 scores were found to be 0.73, 0.68, and 0.70, respectively. Furthermore, the overall accuracy of the model was measured to be 0.73. At the individual category level, the model performed best on r/adhd (f1 score = 0.86) and r/depression (f1 score = 0.77) and worst on r/bipolarreddit (f1 score = 0.63) and r/bpd (f1 score = 0.58) (Table 3). These results indicate that the XGBoost classifier demonstrated moderate predictive performance in

identifying the subreddit from which a post originated.

Table 3. Performance metrics for XGBoost classifier using a 0.5 threshold.

Subreddit	Precision	Recall	f1 score
r/adhd	0.86	0.86	0.86
r/anxiety	0.66	0.69	0.67
r/bipolarreddit	0.68	0.50	0.58
r/bpd	0.65	0.61	0.63
r/depression	0.73	0.81	0.77
r/ptsd	0.78	0.63	0.70
r/schizophrenia	0.76	0.62	0.69

For each subreddit, we used the classifier probabilities to estimate the area under the curve (AUC) for a one-vs-rest classification task (Fig. 3). Overall, the AUC values were high (0.89-0.97), indicating that posts within a subreddit are highly distinguishable from the others. The r/adhd subreddit had the highest AUC (0.97) suggesting that this topic is most dissimilar from other topics. Meanwhile, the r/bpd subreddit post had the lowest AUC (0.89) and thus may share more linguistic features with the other subreddits.

Figure 3. ROCs and AUCs for comparing each class against all other classes. The thick red line shows the micro-averaged ROC of all the individual ROCs.

We computed a confusion matrix to assess the performance of the multiclass classifier at the individual category level and across pairs of categories (Fig. 4). This allows us to explore the misclassification rates between different categories and gain insights into the specific types of errors made by the classifier. In particular, we find the following to be four most common true-predicted category confusions: (r/bpd-r/depression, r/anxiety-r/depression, r/bipolarreddit-r/depression, r/bipolarreddit-r/bpd). Notably, the model often misclassifies posts as coming from r/depression. This may that result from users in r/bpd, r/anxiety, and r/bipolarreddit using language that is more similar to users in r/depression. Interestingly, the confusion rates between r/anxiety, r/bpd, and r/bipolarreddit are less than 0.13, suggesting that linguistic overlap between these subreddits is less than each with r/depression. Additionally, this result may be partially explained by the overrepresentation of the r/depression class in the dataset that is incompletely offset by class weighting when training the model.

Figure 4. Rates of correct and incorrect classification for each class, normalized by the total number of posts in each class.

Finally, we compared using LLM-based embeddings to the current state of the art methods for generating sentence embeddings: Siamese BERT-networks [11]. On all measures, using GritLM-7b embeddings as features in the classifier outperformed S-BERT embeddings (Table 4).

Table 4. Comparison of performance of classifier using LLM and S-Bert embeddings.

Accuracy	AUC	Marco Average		
		Precision	Recall	F1-score

GritLM-7b	0.73	0.95	0.73	0.68	0.7
S-BERT	0.54	0.86	0.47	0.45	0.46

Discussion

Principal Results

The use of large language models (LLMs) offers a novel approach to analyzing patterns of language usage from spontaneous, patient-generated communication. In the field of psychiatric disorders this has the potential to revolutionize the way we diagnose and monitor these conditions. By analyzing the spontaneous use of language in online discussion data, we can gain valuable insights into the linguistic patterns that distinguish between different psychiatric disorders. In this study, we utilized embeddings derived from the GRIT-7b LLM to classify posts originating from subreddits dedicated to seven common conditions: schizophrenia, borderline personality disorder, depression, attention deficit hyperactivity disorder, anxiety, post-traumatic stress disorder, and bipolar disorder.

We found that the XGBoost classifier using features generated from GRIT embeddings has high, but not perfect performance. Symptoms of psychiatric disorders span a number of different modalities, such as sleep, appetite, and motor activity [12-14]. Therefore, in addition to standard reasons for misclassification such as overfitting and noisy or limited data, posts may be misclassified due to similarities in presentations across the different psychiatric disorders, or high rates of comorbidity between disorders. In our study, the pair-wise misclassification rates may reveal disorders where the presentation of symptoms that are likely to be revealed by spontaneous speech are more similar. For example, posts coming from r/bpd are likely to be mis-classified as coming from r/depression. This is consistent with previous studies showing similarities in word usage [15] and high rates of comorbidity for bipolar disorder and depression [16].

We observe a range of performance in the one-versus-rest classifiers across different disorders. Notably, Borderline Personality Disorder (BPD) has the lowest AUC (0.89). This may be because BPD shares many symptoms with other conditions and is often comorbid with them[17]. A key differentiating factor between BPD and related conditions is the temporal nature of symptoms. In BPD, core symptoms are typically more variable and transient compared to their presentation in related conditions. The dataset used in this study consists mostly of single social media posts from each user, providing a cross-sectional view rather than a longitudinal one. This cross-sectional data does not capture within-subject symptom variability which may explain the reduced performance of the BPD classifier. In contrast, ADHD has the highest AUC (0.97). This could be because ADHD is a neurodevelopmental disorder, unlike the other conditions examined in this work. As a result, spontaneous communication related to ADHD may be more distinct from other conditions, leading to higher classifier performance.

Limitations

One limitation of analyzing posts from mental health disorder subreddits is that they may not necessarily originate from officially diagnosed patients. The individuals participating in these online communities may self-identify with a particular disorder without having received a formal diagnosis from a healthcare professional or may be care-givers supporting a diagnosed individual. As a result, the content may not be directly clinically relevant. Therefore, the insights gained from analyzing these posts should be interpreted with caution and may not fully represent the experiences and perspectives of clinically diagnosed individuals. However, future research could apply this

methodology to free speech data generated by verified patients to derive further clinical findings.

Additionally, since subreddits posts are unprompted. Therefore, the embedding space may separate the different categories by variations in the topics or themes across subreddits, in addition to features related to the way language is used in the posts. Nevertheless, this could still hold clinical significance as specific themes, like paranoia in schizophrenia or anhedonia in depression, have diagnostic relevance.

Conclusions

This study demonstrates the potential of using LLMs with free speech to diagnose psychiatric disorders and track their status in individuals. The results suggest that LLMs can provide valuable insights into the linguistic patterns that distinguish between different psychiatric disorders, and that these patterns can be used to develop more objective, efficient, and patient-centered strategies for assessment, monitoring, and research. However, further research is needed to validate these findings and explore the limitations of this approach.

Acknowledgements

The project was conceived and designed by R.S., J.S. and M.M.B. Funding for computational resources was provided by Bristol Myers Squibb Company. Analyses and interpretation were done by R.S. Data visualization was performed by R.S. Supervision and scientific direction were provided by J.S. and M.M.B. The manuscript was written by R.S., J.S. and M.M.B. All authors provided critical review of the manuscript.

The authors thank Joan Buenconsejo and Kapil Sen for helpful comments.

Conflicts of Interest

R.S., J.S. and M.M. B. are current employees of and report equity ownership in Bristol Myers Squibb. Funding for data generation, processing, and storage was provided by Bristol Myers Squibb.

Abbreviations

LLM: Large language model
AUC: Area under the ROC curve
ROC: Receiver operating characteristic
Scz: Schizophrenia
BPD: Borderline personality disorder
ADHD: Attention deficit hyperactivity disorder
Bip: Bipolar disorder
Anx: Anxiety
PTSD: Post-traumatic stress disorder
Dep: Depression

References

1. Newson JJ, Hunter D, Thiagarajan TC. The Heterogeneity of Mental Health Assessment. *Front Psychiatry*. 2020;11:76. PMID: 32174852. doi: 10.3389/fpsy.2020.00076.
2. Aboraya A. Use of Structured Interviews by Psychiatrists in Real Clinical Settings: Results of an Open-question Survey. *Psychiatry (Edgmont)*. 2009 Jun;6(6):24-8. PMID: 19724758.

3. Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB, et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. *NPJ Schizophr*. 2015;1:15030. PMID: 27336038. doi: 10.1038/npjrschz.2015.30.
4. Carter PE, Grenyer BF. Expressive language disturbance in borderline personality disorder in response to emotional autobiographical stimuli. *J Pers Disord*. 2012 Jun;26(3):305-21. PMID: 22686220. doi: 10.1521/pedi.2012.26.3.305.
5. Kelley SW, Gillan CM. Using language in social media posts to study the network dynamics of depression longitudinally. *Nat Commun*. 2022 Feb 15;13(1):870. PMID: 35169166. doi: 10.1038/s41467-022-28513-3.
6. O'Dea B, Boonstra TW, Larsen ME, Nguyen T, Venkatesh S, Christensen H. The relationship between linguistic expression in blog content and symptoms of depression, anxiety, and suicidal thoughts: A longitudinal study. *PLoS One*. 2021;16(5):e0251787. PMID: 34010314. doi: 10.1371/journal.pone.0251787.
7. Agbavor F, Liang H. Artificial Intelligence-Enabled End-To-End Detection and Assessment of Alzheimer's Disease Using Voice. *Brain Sci*. 2022 Dec 23;13(1). PMID: 36672010. doi: 10.3390/brainsci13010028.
8. Muennighoff N, Su H, Wang L, Yang N, Wei F, Yu T, et al. Generative Representational Instruction Tuning2024 February 01, 2024:[arXiv:2402.09906 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2024arXiv240209906M>.
9. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. *J Med Internet Res*. 2020 Oct 12;22(10):e22635. PMID: 32936777. doi: 10.2196/22635.
10. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Singh Chaplot D, de las Casas D, et al. Mistral 7B2023 October 01, 2023:[arXiv:2310.06825 p.]. Available from: <https://ui.adsabs.harvard.edu/abs/2023arXiv231006825J>.
11. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*. 2019.
12. Cosgrove KT, Burrows K, Avery JA, Kerr KL, DeVille DC, Aupperle RL, et al. Appetite change profiles in depression exhibit differential relationships between systemic inflammation and activity in reward and interoceptive neurocircuitry. *Brain Behav Immun*. 2020 Jan;83:163-71. PMID: 31604141. doi: 10.1016/j.bbi.2019.10.006.
13. Benca RM. Sleep in psychiatric disorders. *Neurol Clin*. 1996 Nov;14(4):739-64. PMID: 8923493. doi: 10.1016/s0733-8619(05)70283-8.
14. Depp CA, Bashem J, Moore RC, Holden JL, Mikhael T, Swendsen J, et al. GPS mobility as a digital biomarker of negative symptoms in schizophrenia: a case control study. *NPJ Digit Med*. 2019;2:108. PMID: 31728415. doi: 10.1038/s41746-019-0182-1.
15. Molendijk ML, Bamelis L, van Emmerik AA, Arntz A, Haringsma R, Spinhoven P. Word use of outpatients with a personality disorder and concurrent or previous major depressive disorder. *Behav Res Ther*. 2010 Jan;48(1):44-51. PMID: 19819423. doi: 10.1016/j.brat.2009.09.007.
16. Zanarini MC, Frankenburg FR, Dubo ED, Sickel AE, Trikha A, Levin A, et al. Axis I comorbidity of borderline personality disorder. *Am J Psychiatry*. 1998 Dec;155(12):1733-9. PMID: 9842784. doi: 10.1176/ajp.155.12.1733.
17. (UK) NCCfMH. National Institute for Health and Care Excellence: Guidelines. Borderline Personality Disorder: Treatment and Management. Leicester (UK): British Psychological Society (UK)

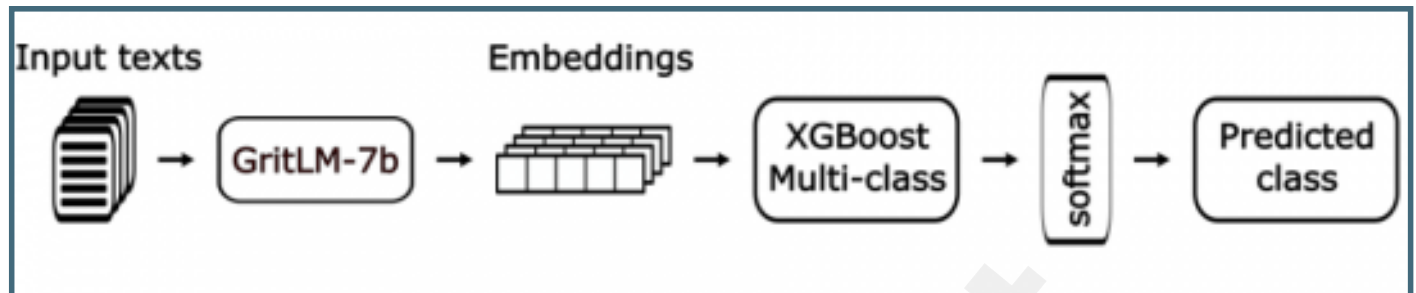
Copyright © 2009, The British Psychological Society & The Royal College of Psychiatrists.; 2009.



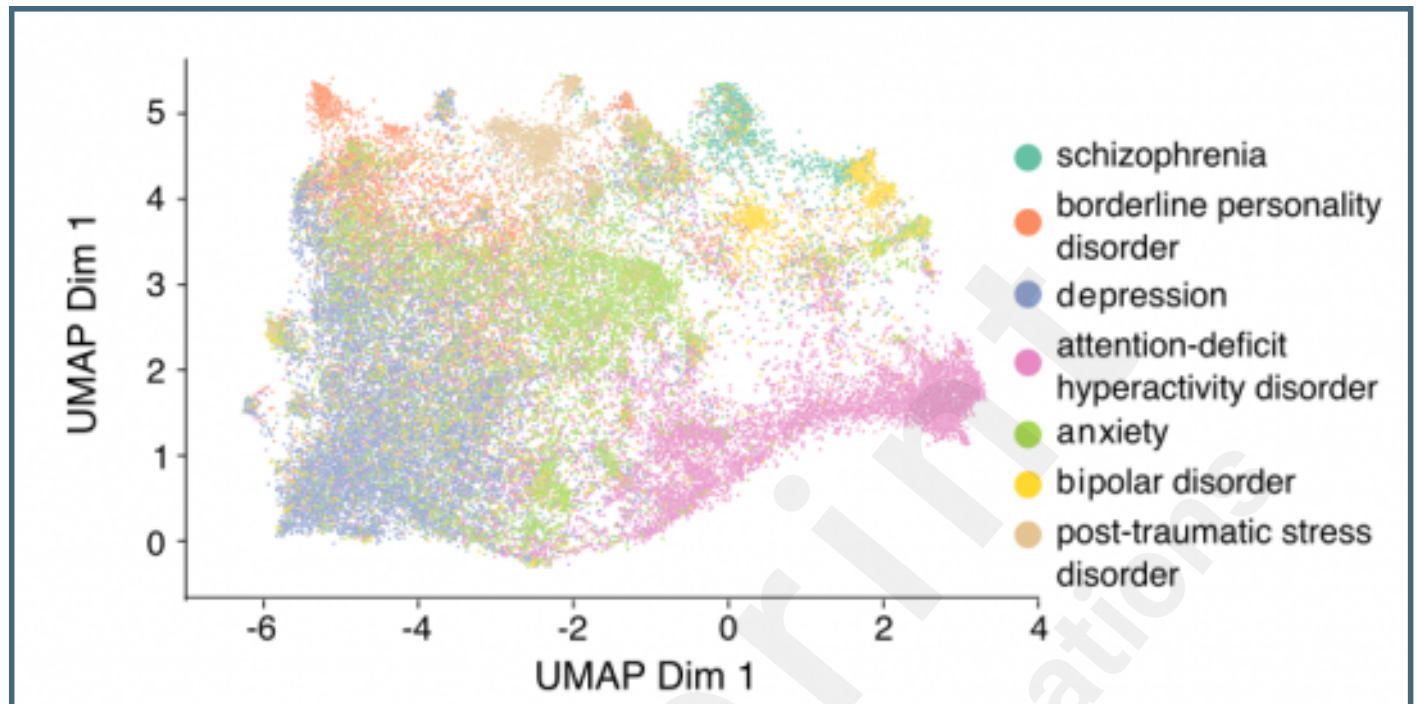
Supplementary Files

Figures

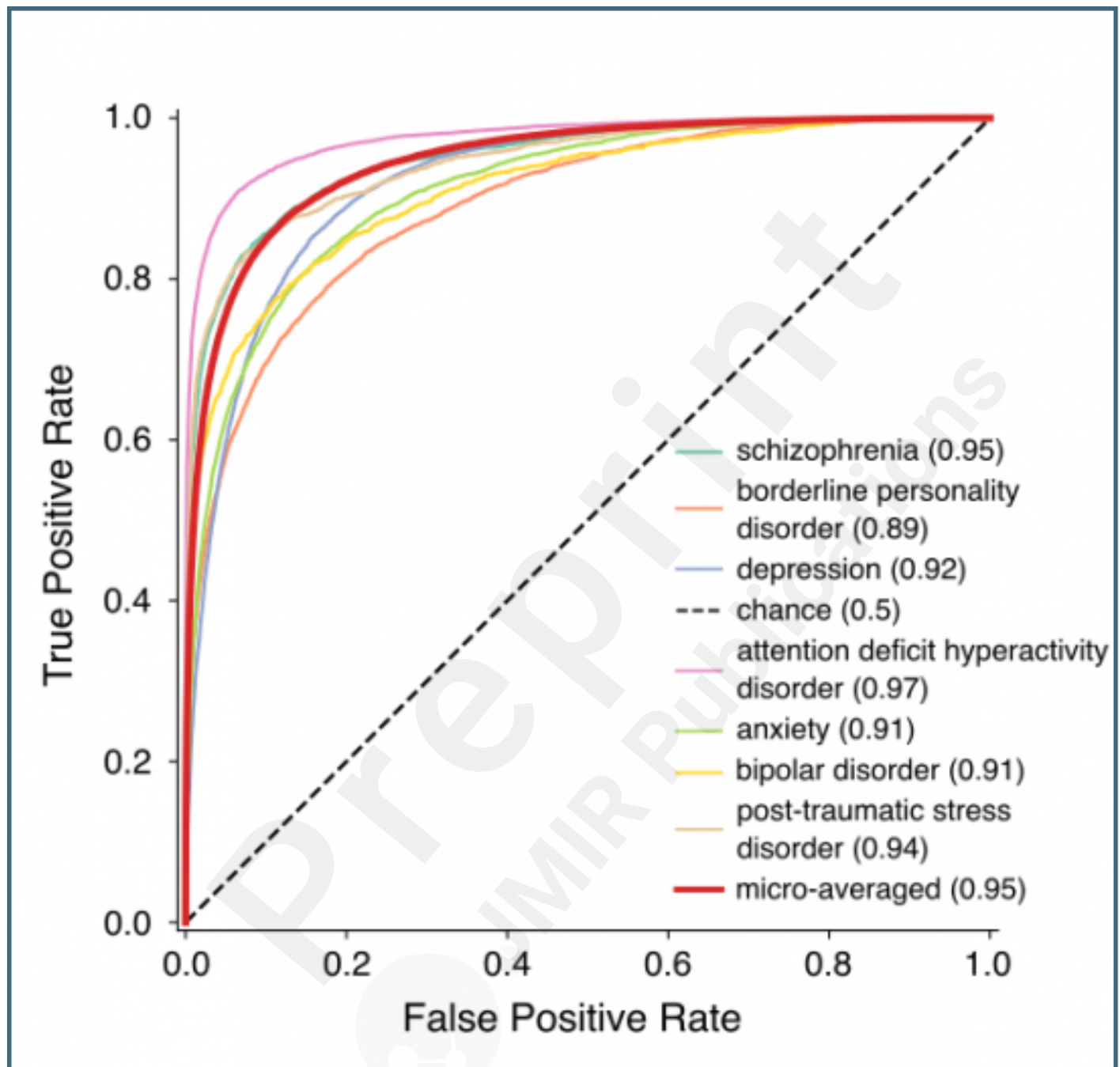
Embeddings of the input reddit posts obtained from GritLM-7B are used as features in an XGBoost classifier to predict subreddit of origin.



A two-dimensional representation of posts in the high-dimensional embeddings space using the UMAP algorithm. The posts are colored by their corresponding subreddit.



ROCs and AUCs for comparing each class against all other classes. The thick red line shows the micro-averaged ROC of all the individual ROCs.



Rates of correct and incorrect classification for each class, normalized by the total number of posts in each class.

