

# MANTIS at #SMM4H 2023: Leveraging Hybrid and Ensemble Models for Detection of Social Anxiety Disorder on Reddit

**Sourabh Zanwar**

RWTH Aachen University  
sourabh.zanwar@rwth-aachen.de

**Daniel Wiechmann**

University of Amsterdam  
d.wiechmann@uva.nl

**Yu Qiao**

RWTH Aachen University  
yu.qiao@rwth-aachen.de

**Elma Kerz**

RWTH Aachen University  
elma.kerz@ifaar.rwth-aachen.de

## Abstract

This paper presents our system employed for the Social Media Mining for Health 2023 Shared Task 4: Binary classification of English Reddit posts self-reporting a social anxiety disorder diagnosis. We systematically investigate and contrast the efficacy of hybrid and ensemble models that harness specialized medical domain-adapted transformers in conjunction with BiLSTM neural networks. The evaluation results outline that our best performing model obtained 89.31% F1 on the validation set and 83.76% F1 on the test set.

## 1 Introduction

According to the Anxiety & Depression Association of America<sup>1</sup>, anxiety disorders rank as the most prevalent mental illnesses in the United States. An estimated 40 million adults, constituting 19.1% of the population aged 18 and above, grapple with these conditions annually. This challenge is compounded by the scarcity of accessible mental health care services and the frequent occurrence of misdiagnoses, often causing individuals to unknowingly endure these disorders (Kasper, 2006).

Natural Language Processing in combination with Machine Learning is increasingly recognized as having transformative potential to support healthcare professionals and stakeholders in the early detection, treatment and prevention of mental disorders (Zhang et al., 2022). In this paper, we report on our participation in The Social Media Mining for Health Applications (#SMM4H) 2023 workshop, which aims to promote automated methods for mining social media data for health informatics. We chose to participate in the Shared Task 4 competition, which was to improve social anxiety detection in Reddit posts (Klein et al., 2023). We approached this task by developing hybrid and ensemble models combining domain-matched transformers with

Bidirectional Long Short-Term Memory (BiLSTM) networks trained on a comprehensive set of engineered linguistic features. This set encompasses measures of morpho-syntactic complexity, lexical sophistication/diversity, readability, stylistics measures (register-specific ngram frequencies) and sentiment/emotion lexicons.

## 2 Data

The data for Task 4 consisted of 8117 Reddit posts written by users aged between 12 and 25 years. These data were split into training (75%), validation (8.4%), and testing sets (16.6%).

In preparation for model training, all texts were subjected to preprocessing procedures including eliminating HTML, URLs, excessive spaces, and emojis from the text, as well as rectifying inconsistent punctuation.

## 3 System Description

Our systems leveraged three domain-adapted Transformer-based pretrained language models (PLM), a BiLSTM trained on engineered features and their combination forming into hybrid and ensemble models. Domain-adapted pretrained language models include: (1) PsychBERT (Vajre et al., 2021) (2) Mental RoBERTa (Ji et al., 2022) and (3) Clinical BERT (Alsentzer et al., 2019). All PLMs were obtained from the Huggingface (Wolf et al., 2020), choosing the uncased, where applicable, base versions. We constructed a BiLSTM trained on 168 features that fall into six categories. All measurements of these features were obtained using a system that employs a sliding window technique to compute sentence-level measurements. The BiLSTM model is formulated as:

$$h_N^{(L)} = \text{BiLSTM}_H^L(\text{CM}_1^N), h_N^{(L)} = h_{f,N}^{(L)} \oplus h_{b,N}^{(L)} \\ \hat{y} = \text{Softmax}(Wf c_2(f c_1(h_N^{(L)})) + b)$$

where  $f c_i(x) = \text{ReLU}(W_i x + b_i)$ ,  $\text{BiLSTM}_H^L(\cdot)$  is a  $L$  layer BiLSTM with hidden size of  $H$ .  $\text{CM}_i^N$  represents the linguistic features for the  $i$ th sentence

of a post consisting of  $N$  sentences. The last hidden representation of the last layer in forward and backward directions are denoted by  $h_{f,N}^{(L)}$  and  $h_{b,N}^{(L)}$ .  $\oplus$  denotes the concatenation operator.

The hybrid model combines a Mental RoBERTa model with above BiLSTM.

$$\begin{aligned} S_1^M &= \text{Mental\_RoBERTa}(T_1^M) \\ s_M^{(L_1)} &= \text{BiLSTM}_{H_1}^{L_1}(S_1^M), s_M^{(L_1)} = s_{f,M}^{(L_1)} \oplus s_{b,M}^{(L_1)} \\ h_N^{(L_2)} &= \text{BiLSTM}_{H_2}^{L_2}(CM_1^N), h_N^{(L_2)} = h_{f,N}^{(L_2)} \oplus h_{b,N}^{(L_2)} \\ \hat{y} &= \text{Softmax}(Wfc_3[f c_1(s_M^{(L_1)}) \oplus f c_2(h_N^{(L_2)})] + b) \end{aligned}$$

where  $T_1^M$  is the sequence of tokens from a post.

We constructed three distinct ensemble models using the stacking method: ensemble model (1) composed of instances from the hybrid model, which emerged as the most accurate base model (M6), (2) combining hybrid models with fine-tuned PsychBERT models (M7), and (3) consisting of Mental RoBERTa models, PsychBERT models, and BiLSTM models (M8). The resulting models represent homogeneous ensemble (HOE), intermediate and heterogeneous ensemble (HEE) approaches (Ganaie et al., 2022). As meta-learners, Support Vector Classifier, Logistic Regression, Gradient Boosting and Ridge Regression and XGBoost were used.

Further details are provided in the supplementary material (<https://shorturl.at/epuF3>).

## 4 Results and Evaluation

The results of our models on the validation set and test set are presented in Table 1. The Mental RoBERTa model achieved the highest performance (F1=86.59%) among the PLMs, outperforming the PsychBERT and ClinicalBERT models by 4% and 14.73%, respectively. This finding indicates that the detection of anxiety on Reddit sees a marked improvement from pretraining the PLM on mental health-related subreddits, as opposed to pretraining on clinical text. The hybrid models consistently outperformed the standalone PLM across all model iterations, yielding an average increase in F1 scores of 0.3%. The use of model stacking enhanced classification outcomes with performance boosts ranging between 1.86% and 2.44% in F1 score. The highest balanced classification score was achieved by the HEE model (M8). A variant of this model using ridge regression as a meta-learner (M12) achieved the best performance on the test set (F1 = 83.76%, mean<sub>all teams</sub> = 79.3%, median<sub>all teams</sub> = 82.4%). The HOE model (M6) achieved the second-highest performance and the best precision among all models examined. This suggests that

Table 1: Results on the validation set (top) and test set (bottom). For each ensemble model, we report results of the best performing meta-learner.

Detection model	F1	P	R
<i>Pretrained Language Models</i>			
M1: Mental RoBERTa	86.59	81.85	91.91
M2: PsychBERT	82.59	79.46	85.98
M3: ClinicalBERT	71.86	69.36	74.55
M4: BiLSTM	59.01	60.03	58.92
M5: Hybrid Model	86.87	83.30	90.76
<i>Ensemble Models</i>			
M6: HOE M5 (GB)	88.80	85.57	92.28
M7: HEE M5+M2 (GB)	88.73	84.90	92.92
M8: HEE M1+M2+M4 (GB)	89.31	85.02	94.06
<i>Test set</i>			
M6: HOE: M5 (GB)	83.63	<b>81.07</b>	86.35
M12: HEE M1+M2+M4 (Ridge)	<b>83.76</b>	80.6	<b>87.17</b>

both ensemble approaches can produce beneficial, albeit distinct, impacts on the detection of social anxiety disorder.

## References

- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics.
- Mudasir A Ganaie, Minghui Hu, AK Malik, M Tanveer, and PN Suganthan. 2022. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.
- Siegfried Kasper. 2006. Anxiety disorders: underdiagnosed and insufficiently treated. *International Journal of Psychiatry in Clinical Practice*, 10(sup1):3–9.
- AZ Klein, JM Banda, Y Guo, JI Flores Amaro, R Rodriguez-Esteban, A Sarker, AL Schmidt, D Xu, and G Gonzalez-Hernandez. 2023. Overview of the eighth social media mining for health applications (#smm4h) shared tasks at the AMIA 2023 annual symposium. *Proceedings of the Eighth Social Media Mining for Health Applications (#SMM4H) Workshop and Shared Task*.
- Vedant Vajre, Mitch Naylor, Uday Kamath, and Amarda Shehu. 2021. Psychbert: a mental health language model for social media mental health behavioral analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1077–1082. IEEE.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

T. Zhang, A Schoene, and S. Ananiadou. 2022. Natural language processing applied to mental illness detection: A narrative review. *NPJ Digital Medicine*, 5:46.