The Best of Both Worlds: Combining Engineered Features with Transformers for Improved Mental Health Prediction from Reddit Posts

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

In recent years, there has been increasing interest in the application of natural language processing and machine learning techniques to the detection of mental health conditions (MHC) based on social media data. In this paper, we aim to improve the state-of-the-art (SotA) detection of six MHC in Reddit posts in two ways: First, we built models leveraging Bidirectional Long Short-Term Memory (BLSTM) networks trained on in-text distributions of a comprehensive set of psycholinguistic features for more explainable MHC detection as compared to black-box solutions. Second, we combine these BLSTM models with Transformers to improve the prediction accuracy over SotA models. In addition, we uncover nuanced patterns of linguistic markers characteristic of specific MHC.

1 Introduction

The last decade has seen a surge in digital mental health research aimed at using linguistic cues from social media data to predict mental health conditions (MHC). The ultimate goal of this research branch is to enable people to lead healthy lives and to support healthcare professionals in the diagnosis and treatment of mental disorders. Data from social media platforms are particularly compelling to the research community because of their scale and deep embedding in contemporary culture. The computational research on such data has yielded new insights into population mental health and has shown promise for incorporating data-driven analytics into the treatment of psychiatric disorders (see Guntuku et al. 2017; Thieme et al. 2020 for recent general overviews of this research; see Chancellor and De Choudhury 2020; Harrigian et al. 2021 for reviews focusing on methods and data used for mental health status on social media). The current state of the art (SotA) in MHC detection from text data in social media is achieved by approaches that employ deep learning techniques that use pre-trained word embeddings, in particular Transformer-based models. At the same time, there are increasing calls to move away from such models toward explainable models for critical industries (Rudin, 2019; Loyola-Gonzalez, 2019), and the black-box nature of such models is increasingly being recognized as a major hurdle in the use of AI models in clinical practice (Mullenbach et al., 2018). Effective supporting the diagnosis and treatment of mental disorders therefore requires both accurate and interpretable models.

In this paper, we make the following contributions to the dynamic area of research on detecting mental health conditions in social media texts: First, we present attention-based BLSTM models that leverage the within-text distributions ('textual contours') of a large array of engineered language features to predict the presence of six mental health conditions (MHC) from user posts on Reddit. We show that these models can perform competitively with SotA models (drop in F1 $\leq 3\%$) for three of the six MHC, while maintaining explainability. Second, we demonstrate that the accuracy of MHC detection can be increased considerably by integrating these models with Transformer-based models. In addition, we uncover nuanced patterns of linguistic markers characteristic of specific MHC.

2 Related work

In this section, we focus on the state-of-the-art approaches to MHC detection on the two Reddit datasets used in this paper, i.e SMHD (Cohan et al., 2018) and Dreaddit (Turcan and McKeown, 2019) (for details see Section 3.1). On the SMHD dataset the current state-of-the-art approach is based on a Hierarchical Attention Network (HAN) (Sekulic and Strube, 2019). The proposed model uses two layers of bidirectional GRU units with hidden size of 150, each of them followed by a 100 dimensional attention mechanism. The first layer was set up to encode posts, whereas the second one encodes

Proceedings of the 29th International Conference on Computational Linguistic, pages 197–202
October 12–17, 2022.

users in terms of a sequence of encoded posts. The input layer was initialized with 300 dimensional GloVe word embeddings (Pennington et al., 2014). The output layer is 50-dimensional fully connected network, with binary cross entropy as a loss function. Sekulic and Strube (2019) perform experiments with the number of posts per user available to the model (50, 100, 150, 200, 250) to examine the amount of data needed for reasonable performance. Their best-performing model achieved an average of 67.56% F1 across the five MHC. On the Dreaddit dataset the current state-of-the-art approach is based on a 'emotion-infused' BERT model (Turcan et al., 2021). In the best-performing model presented in the paper, the BERT representation was first augmented with emotion knowledge by fine-tuning it on the GoEmotions dataset (Demszky et al., 2020) before applying it to the stress detection task. The model reached 80.25% F1 in binary stress classification, performing slightly better than a BERT baseline model. It is interesting to note that both SotA papers highlight the importance of explainability in MHC detection and suggest ways to increase the interpretability of the predictions of their models: Sekulic and Strube (2019) perform attention weights analyses to identify posts, and words or phrases in those posts, that are relevant for classification, whereas Turcan et al. (2021) use LIME to identify the most important words used their models for stress classification and mapping these to LIWC categories (Pennebaker et al., 2015).

3 Experimental setup

3.1 Datasets

The data for this work comes from two recent corpora used for the detection of MHC: (1) the Self-Reported Mental Health Diagnoses (SMHD) dataset (Cohan et al., 2018) and (2) the Dreaddit dataset (Turcan and McKeown, 2019). Both SMHD and Dreaddit were constructed data from Reddit, a social media platform consisting of individual topic communities called subreddits, including those relevant to MHC detection. The length of Reddit posts makes them a particularly valuable resource, as it allows modeling of the distribution of linguistic features in the text (see the concept of 'text contours' in Section 3.2).

SMHD is a large dataset of social media posts from users with nine mental health conditions (MHC) corresponding to branches in the DSM-5 (APA, 2013), an authoritative taxonomy for psychiatric diagnoses. User-level MHC labels were obtained through carefully designed distantly supervised labeling processes based on diagnosis pattern matching. The pattern matching leveraged a seed list of diagnosis keywords collected from the corresponding DSM-5 headings and extended by synonym mappings. To prevent that target labels can be easily inferred from the presence of MHC indicating words/phrases in the posts, all posts made to mental health-related subreddits or containing keywords related to a mental health condition were removed from the diagnosed users' data.

Table 1: Datasets statistics (number of posts, means and standard deviations of post length (in words) across mental health conditions and control groups.

MHC	Dataset	N posts	M length	SD
Stress	Dreaddit	1857	91	35
Control	Dreaddit	1696	83.6	29.7
ADHD	SMHD	1849	91.4	57
Anxiety	SMHD	1846	91.7	56.3
Bipolar	SMHD	1848	93	57.7
Depression	SMHD	1846	92.4	58.7
PTSD	SMHD	1600	95.7	59.9
Control	SMHD	1805	78.8	48.6

Dreaddit is a dataset of lengthy social media posts from subreddits in five domains that include stressful and non-stressful text. For a subset of 3.5k users employed in this paper, binary labels (+/- stressful) were obtained from aggregated ratings of five crowdsourced human annotators.

Based on these two corpora, we constructed a dataset with the goal of obtaining sub-corpora of equal size for the six MHCs targeted in this paper. To this end, we downsampled SMHD to match the size of Dreaddit and to be balanced in terms of class distributions. The sampling procedure from the SMHD dataset was such that each post was produced by a distinct user. In doing so, we addressed a concerning trend described in recent review articles that points to the presence of a relatively small number of unique individuals, which may hinder the generalization of models to platforms that are already demographically skewed (Chancellor and De Choudhury, 2020; Harrigian et al., 2021). These constraints were met for five of the nine MHC in the SMHD dataset (ADHD, anxiety, bipolar, depression, PTSD). Statistics for these datasets are presented in Table 1.

3.2 Measurement of text contours of psycholinguistic features

A set of 435 psycholinguistic features used in our approach fall into five broad categories: (1) features of morpho-syntactic complexity (N=19), (2) features of lexical richness, diversity and sophistication (N=52), (3) register-based n-gram frequency features (N=25), (4) readability features (N=14), and (5) lexicon features designed to detect sentiment, emotion and/or affect (N=325). An overview of these features can be found in Table 3 in supplementary material. All measurements of these features were obtained using an automated text analysis system that employs a sliding window technique to compute sentence-level measurements. These measurements capture the within-text distributions of scores for a given psycholinguistic feature, referred to here as 'text contours' (for its recent applications, see e.g. Wiechmann et al. (2022) for predicting eye-moving patterns during reading and Kerz et al. (2022) for detection of Big Five personality traits and Myers–Briggs types). A visualization of these text contours is shown in Figure 1 in Supplementary material. Tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic PCFG parsing were performed using Stanford CoreNLP (Manning et al., 2014).

4 Modeling approach

We conducted experiments with a total of five models: (1) a fine-tuned BERT model (Devlin et al., 2019), (2) a fine-tuned RoBERTa model (Zhuang et al., 2021), (3) a bidirectional long short-term memory (BLSTM) classifier trained on measurements of psycholinguistic features described in Section 3.2, and (4) and (5) two hybrid models integrating BERT and RoBERTa predictions with the psycholinguistic features. For (1) and (2) we used the pretrained 'bert-base-uncased' and 'roberta-base' models from the Huggingface Transformers library (Wolf et al., 2020) each with an intermediate BLSTM layer with 256 hidden units (Al-Omari et al., 2020). For (3) - the model based solely on psycholinguistic features, we constructed a 4-layer BLSTM with a hidden state dimension of 1024. The input to that model is a sequence $CM_1^N = (CM_1, CM_2 \ldots, CM_N)$, where CM_i, the output of CoCoGen for the ith sentence of a post, is a 435 dimensional vector and N is the sequence length. To predict the labels of a sequence, we concatenate the last hidden states of the last layer in forward ($\overrightarrow{h_n}$) and backward directions ($\overleftarrow{h_n}$). The result vector of concatenation $h_n = [\overrightarrow{h_n}|\overleftarrow{h_n}]$ is then transformed through a 2-layer feedforward neural network, whose activation function is Rectifier Linear Unit (Agarap, 2018). The output of this is then passed to a Fully Connected Layer FC with ReLu activation function and dropout of 0.2 and it is finally fed to a final FC layer. The output is finally passed through sigmoid function and finally a threshold is used to determine the labels. We trained these models for 100 epochs, with a batch size of 256 and a sequence length of 5. The architecture of the hybrid classification models - models (4) and (5) - consists of two parts: (i) a pretrained Transformer-based model with a BLSTM layer and FC layer on top of it and (ii) the psycholinguistic features of the text fed into a BLSTM network and a subsequent FC layer. The FC layers of both parts take the concatenation of last hidden states of the last BLSTM layer in forward and backward direction. We concatenate the outputs of these layers before finally feeding them into a final FC layer with a sigmoid activation function. The model used to generate predictions for the test set was the RoBERTa-PsyLing hybrid model with the following configuration: 2-layer BLSTM, 256 hidden units and a dropout of 0.2; BLSTM-PsyLing: 3-layers, hidden size of 512 and dropout 0.2. We trained this model for 12 epochs, saving the model with the best performance (F1-Score) on the development set. The optimizer used is AdamW with a learning rate of 2e-5 and a weight decay of 1e-4. Structure diagrams of the model based solely on psycholinguistic features and the hybrid architectures are presented in Figures 2 and 3 in supplementary material. All models were trained using 5-fold CV of the training data as base classifiers and model stacking was performed using logistic regression as a meta-learner to adaptively combine the outputs of the base classifiers.

5 Results and Discussion

An overview of the results of our models in comparison to those reported in the previous studies reviewed above is presented in Table 2. The results show that our Psyling-BLSTM models outperform the Hierarchical Attention Networks that are based on 50 concatenated posts (HAN 50) across all five self-reported diagnosed MHC, with an average increase in performance of over 12% F1. This is significant considering that our Psyling-BLSTM models were trained in a much more challenging

Table 2: F1 scores averaged over five runs of the binary classification models across MHC.

	Stress	ADHD	Anxiety	Bipolar	Depression	PTSD
HAN (50) (Sekulic and Strube, 2019)	-	48	38	52	49	56
HAN (250) (Sekulic and Strube, 2019)	-	64.27	69.24	67.42	68.28	68.59
BERT (Turcan et al., 2021)	78.88	-	-	-	-	-
BERT-Emotion (Turcan et al., 2021)	80.25	-	-	-	-	-
BERT	76.24	60.82	64.52	65.11	62.91	69.40
RoBERTa	81.13	61.76	67.79	66.12	65.93	72.32
Psyling-BLSTM	70.20	58.2	61.19	61.19	59.54	65.18
Psyling-BLSTM-BERT	79.46	61.09	67.47	67.44	66.90	73.32
Psyling-BLSTM-RoBERTa	83.32	62.37	68.91	67.85	67.16	73.85

setting, namely the detection of the presence of an MHC based on a single post with an average length of less than 100 words. The Psyling-BLSTM models show a relative performance drop of -6.5% compared to a HAN that was trained on 250 times as much textual data. These results demonstrate that the Psyling-BLSTM approach is much more data-efficient than the HAN approach. Furthermore, our results show that the Psyling-BLSTM models are surpassed by the Transformer-based models, with BERT achieving on average a 3.9% higher F1 and RoBERTa achieving 6.6% higher F1. Importantly, the Psyling-BLSTM models perform competitively with BERT (drop in F1 $\approx 3\%$) for three of the six MHC (ADHD, anxiety, depression), indicating that strong detection models can be constructed for at least these MHC without compromising explainability. Another key finding is that the combination of Psyling-BLSTM models with Transformers consistently yielded the highest prediction accuracy for all six MHC, with the RoBERTa hybrid models outperforming the BERT hybrid models in all cases. The highest prediction accuracy was achieved in stress detection, with a +3.06% F1 improvement over the previous SotA, the emotion-loaded BERT model presented in Turcan et al. (2021). For diagnosed MHC, the RoBERTa hybrid showed a robust increase in F1 over a Transformer-only model by an average of 1.24%. For the BERT-based case, the improvement in F1 averaged +3.2% for stress and +2.7% for self-reported diagnosed MHC.

To further investigate the differences in the linguistic markers among MHC, we ran one-way ANOVAs comparing the features across the six MHC and the control group for each of the 435 linguistic features. Prior to being entered in the models, all feature scores were z-score normalized and Bonferroni correction was used to correct for family-wise type I errors. Follow up Tukey's HSD post-hoc test were run to measure the differences between all MHC pairs. The results revealed significant differences across groups for 337 of the linguistic features (see Table **??** in the appendix). We focused on the linguistic features with the greatest variance among the groups, as indicated by the highest F-statistics (N=68), and conducted hierarchical agglomerative cluster analyses over MHC and linguistic features to derive the linguistic profiles associated with the six MHC groups and the control group. The analyses show that the control group is distinctly separated from all MHC. Furthermore, while each of the six MHC were characterized by distinct patterns of language use, the linguistic markers of stress were clearly separated from those of the five self-reported diagnosed MHC. In light of space limitations, we present here only some selected general patterns (for details, see Figure 4 and Table 7 in supplementary material). For example, there were persistent differences between stressful and non-stressful Reddit posts, with stressful posts being generally characterized by much higher proportions of negative words and words related to sadness, fear, anxiety and anger. The language use of Reddit users diagnosed with depression was characterized by higher proportions of self-referencing words and lower lexical sophistication. Posts from users diagnosed with PTSD had significantly longer mean sentence lengths, greater syntactic complexity and lower readability, relative to controls, while the Reddit posts of users diagnosed with anxiety were characterized by high proportions of regular verbs and those of users diagnosed with bipolar disorder showed strong reliance on words signaling indifference and lower lexical diversity. Overall, our results show that mental health prediction from textual data benefits from the advancement of methods aimed at integrating Transformers with comprehensive sets of open- and closed-vocabulary features and general features of linguistic complexity and style.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Hani Al-Omari, Malak A. Abdullah, and Samira Shaikh. 2020. EmoDet2: Emotion detection in english textual dialogue using bert and bilstm models. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 226–232.

APA. 2013. Diagnostic and statistical manual of mental disorders. *American Psychiatric Association*, 21(21):591–643.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):1–11.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, Online. Association for Computational Linguistics.

Elma Kerz, Yu Qiao, Sourabh Zanwar, and Daniel Wiechmann. 2022. Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, Dublin, Ireland. Association for Computational Linguistics.

Octavio Loyola-Gonzalez. 2019. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical report.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Ivan Sekulic and Michael Strube. 2019. Adapting deep learning methods for mental health prediction on social media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Association for Computational Linguistics.

Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(5):1–53.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107. Association for Computational Linguistics.

Elsbeth Turcan, Smaranda Muresan, and Kathleen McKeown. 2021. Emotion-infused models for explainable psychological stress detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2895–2909.

Daniel Wiechmann, Yu Qiao, Elma Kerz, and Justus Mattern. 2022. Measuring the impact of (psycho-)linguistic and readability features and their spill over effects on the prediction of eye movement patterns. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland. Association for Computational Linguistics.

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.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, Huhhot, China. Chinese Information Processing Society of China.

A Supplementary material

Supplementary material can be found here https://bit.ly/3LgGYla.