SMHD-GER: A Large-Scale Benchmark Dataset for Automatic Mental Health Detection from Social Media in German

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

Mental health problems are a challenge to our modern society, and their prevalence is predicted to increase worldwide. Recently, a surge of research has demonstrated the potential of automated detection of mental health conditions (MHC) through social media posts, with the ultimate goal of enabling early intervention and monitoring population-level health outcomes in real time. Progress in this area of research is highly dependent on the availability of high-quality datasets and benchmark corpora. However, the publicly available datasets for understanding and modeling MHC are largely confined to the English language. In this paper, we introduce SMHD-GER (Self-Reported Mental Health Diagnoses for German), a large-scale, carefully constructed dataset for MHC detection built on high-precision patterns proposed for English. We provide benchmark models for this dataset to facilitate further research and conduct extensive experiments. These models leverage engineered (psycho-)linguistic features as well as BERT-German. We also examine nuanced patterns of linguistic markers characteristics of specific MHC.

1 Introduction

Mental health is a major challenge in healthcare and in our modern societies at large (Rehm and Shield, 2019; Santomauro et al., 2021). The World Health Organization estimates that 970 million people worldwide suffer from mental health conditions[1][2], with the rate of undiagnosed mental disorders estimated to be as high as 45% (La Vonne et al., 2012).

The enormous societal impact of mental health conditions (MHC) requires prevention and intervention strategies that focus primarily on screening and early detection. The last decade has seen a surge in digital mental health research, an interdisciplinary line of research that brings together insights from computational linguistics, cognitive psychology and computational social sciences to understand the relationship between patterns of language use and mental health conditions (D'Alfonso, 2020; Schindler and Domahidi, 2022). Natural language processing, in particular, is increasingly recognized as having transformative potential to support healthcare professionals in the diagnosis and treatment of mental disorders and enable people to lead healthy lives (see Guntuku et al. 2017; Thieme et al. 2020; Chancellor and De Choudhury 2020; Zhang et al. 2022 for recent overviews of this research).

Progress in this area of research is highly dependent on the availability of high-quality datasets and benchmark corpora. Social media has emerged as an increasingly vital resource for obtaining such data, as it is now a central place for individuals to participate in discussions, share information, and seek advice. Based on data drawn from platforms such as Twitter and Reddit, recent work has developed scalable methods for constructing mental health datasets based on self-reported diagnoses or grouping individuals based on activity patterns (Coppersmith et al., 2015; Yates et al., 2017; Cohan et al., 2018; Kumar et al., 2015). However, recent reviews on the state of data used for mental health status on social media show that the vast majority of the publicly available datasets for understanding and modeling MHC are on the English language: For example, of the 102 datasets reviewed in Harrigian et al. (2021) 83% were on English with the remaining 17% distributing over five other languages: Chinese (9.8%), Japanese (3.9), Korean (1.9), Spanish and Portuguese (each $< 1\%$). Zhang

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[2] 'Mental disorders' can also be referred to as 'mental health conditions'. The latter is sometimes used as a broader term encompassing mental disorders, psycho-social disabilities, and mental conditions, include different types of depression, bipolar disorder, schizophrenia, anxiety disorders, chronic stress etc.. In this work, the two terms are used interchangably.

Findings of the Association for Computational Linguistics: EACL 2023, pages 1526–1541
May 2-6, 2023 ©2023 Association for Computational Linguistics

et al. (2022) report that 81% of all the datasets are in English, followed by datasets in Chinese (10%), and Arabic (1.5%). While an overwhelming focus on English data is a theme throughout the NLP community, it is a specific concern in this domain where culture often influences the presentation of mental health disorders (De Choudhury et al., 2017; Loveys et al., 2018). Thus, there is an urgent need for publicly available, high-quality mental health datasets and benchmark models to support early detection of MHC in languages other than English.

The main contributions of this work are threefold: (1) We introduce SMHD-GER (Self-reported Mental Health Diagnosis for German), a new large dataset of social media posts for mental health detection in the German language, and make it publicly available; (2) We provide benchmark models for the detection of four mental health conditions based on a comprehensive set of text-based features that pertain to multiple levels of language use, the German BERT-based model, and hybrid models that combine the two; and (3) We uncover nuanced patterns of linguistic markers characteristic of specific mental health conditions.

The remainder of this paper is organized as follows: In Section 2 we briefly review available social media datasets and NLP classification methods for MHC detection. Section 3 details the construction of the SMHD-GER dataset along with an ethics and privacy statement. Section 4 presents the results of a analysis of linguistic markers of specific MHC. In Section 5, we describe the modeling approach for our benchmark models, and in Section 6, we present and discuss the results. Finally, we conclude with directions for future work in Section 6.

2 Related work

In this section, we provide a concise overview of some of the most widely used self-disclosure social media datasets along with the classification methods used in the detection of mental health conditions. The self-disclosure approach to obtaining labeled data from social media was introduced in Coppersmith et al. (2014) and further refined in consecutive work (Yates et al., 2017; Cohan et al., 2018). In this approach, public self-reports of mental health diagnoses are identified through the use of carefully designed 'diagnosis patterns' combined with 'diagnosis keywords' mapped to particular mental health conditions: A user is included for a specific MHC if one of the condition keywords occurs within a certain distance of the diagnosis pattern. Coppersmith et al. (2014) originally applied this approach to Twitter data and identified approximately 1,200 users with four MHC (bipolar, depression, PTSD, SAD) by matching diagnosis patterns in their tweets (e.g., "I was diagnosed with depression"). This dataset was employed in the shared task at the 2nd Computational Linguistics and Clinical Psychology Workshop (CLPsych 2015) that focused on identifying depression and PTSD users on Twitter (Coppersmith et al., 2015). Submissions to the task used traditional (shallow) classification models trained on unigram vectors, character language models, closed-vocabulary approaches (e.g. LIWC, Pennebaker et al., 2001) and supervised topic models. The leading systems reached average precision rates over 85% for both MHC. However, the dataset had a balanced distribution between the classes, rather than one that accurately reflect the user population. This hampered the reliable estimation of actual false alarm rates, as the number of false alarms in the general population is estimated to be 7-15 times higher than in the CLPsych 2015 test sample (Coppersmith et al., 2015).

The text content of a Tweet can contain up to 280 characters or Unicode glyphs. Thus, this format presents a barrier to capturing mental health related language signals. Recent work on compiling datasets for mental health is increasingly turning to Reddit for long-form content that can provide additional linguistic insights[3]: Yates et al. (2017) applied the self-disclosure approach to create the Reddit Self-reported Depression Diagnosis (RSDD) dataset, which contains 9,210 users with depression and 107,274 control users. Apart from increasing the dataset size by an order of magnitude – 969 posts per user with mean post length of 148 words, the RSDD dataset displays a realistic number of control users matched with each diagnosed user.

The main limitation of the RSDD dataset is its focus on a single mental health condition, depression. In what is to our our knowledge the most comprehensive, carefully constructed mental health dataset based on the self-disclosure approach, Cohan et al. (2018) expand on RSDD by including for eight additional MHC: The Self-reported Mental Health Diagnoses (SMHD) dataset, whose design

[3]Reddit (https://www.reddit.com/) is a social news aggregation, content rating, and discussion website without any length constraints.

underlies the current work, comprises 20,406 diagnosed users and 335,952 matched controls. Diagnosed users were identified using a refined version of the high precision diagnosis patterns used in RSDD, which incorporated synonyms in matching patterns from two synonym mapping ontologies (MedSyn, Yates and Goharian, 2013, Behavioral, Yom-Tov et al., 2013). Control users were selected based on a similar Reddit posting activity, i.e. each daignosed user was matched with an average of 9 control users with a similar number of posts and a similar range of subreddits they posted in. Importantly, SMHD does not contain any posts that contain any mental health terms or that have been posted in a mental health-related subreddits. The detection of MHC can thus not be based on terms associated with specific mental health conditions. Along with the dataset itself, Cohan et al. (2018) provided benchmarks for both binary (MHC vs. control) and multi-class classification settings. The classification methods included several traditional (shallow) machine learning models (logistic regression, XGBoost (Chen and Guestrin, 2016), support vector machine with linear kernel) trained on tf-idf bag-of-words features, a shallow neural net model trained on character ngrams (Supervised FastText, Joulin et al., 2016), and a Convolutional neural network trained on ngram sequences represented by the FastText embeddings. Subsequent work has improved MHC detection accuracy using Hierarchical Attention Networks (Sekulic and Strube, 2019) and attention-based model using BERT representations (Jiang et al., 2020). Recently, (Zanwar et al., 2022) leveraged transformer language models (BERT Devlin et al., 2019 and RoBERTa Liu et al., 2019) in combination with attention-based BLSTM models trained on engineered language features for MHC detection.

3 Data

3.1 Data construction

In this section we describe the construction and characteristics of the SMHD-GER dataset. SMHD-GER comprises data on seven mental health conditions that correspond to branches in the DSM-5 (APA, 2013): Five conditions are top-level DSM-5 disorders: schizophrenia spectrum disorders (schizophrenia), bipolar disorders (bipolar), depressive disorders (depression), anxiety disorders (anxiety), obsessive-compulsive disorders (OCD). The remaining two conditions are one rank lower: post-traumatic stress disorder (ptsd) is classified under trauma- and stress-related disorders, and attention-deficit/hyperactivity disorder (ADHD) under neurodevelopmental disorders. The construction of the dataset is an adaptation of the general procedure underlying the construction of the SMHD dataset described in Cohan et al. (2018): The textual data were obtained from Reddit using the Pushshift.io API Wrapper by searching for all posts mentioning any mental health (MH) terms, such as the name of a condition. The list of MH-terms was derived from the corresponding materials used for the SMHD dataset using DeepL translator[4] followed by manual inspection and editing. We then filtered these posts to keep only those that were in German using the 'langdetect' the Python library.[5].

Diagnosed users were identified using high precision diagnosis patterns as in Cohan et al. (2018): Reddit users received a positive label for a specific MHC if and only if at least one of their posts explicitly states that they suffer from a specific condition or are engaging in behaviors indicative of it. These were triangulated with specific expressions, such as "Ich wurde diagnostiziert mit X" ("I was diagnosed with X"), where X would be filled with a specific MH-term (e.g. "Depression"). Like the MH-terms, the diagnosis patterns were derived from the corresponding materials used for the SMHD dataset using DeepL translator followed by manual inspection and editing. We then collected all posts and comments for the users with a positive label and filtered these to keep only those that (i) were in German, (ii) had no mentions of any of the MH-terms and (iii) were not posted in a subreddit related to mental health (MH-subreddit).

Control users: To compile the data used for control we collected 1049202 posts from 24981 users from r/de[6] subreddit, and filtered out those users who (i) had used any MH-term in any of their posts or (ii) had posted in a MH-subreddits. For all remaining users we collected all the available posts and comments in German. All Reddit posts were made between August 14, 2009, and October 2, 2022 (inclusive). This procedure yielded a dataset containing 5,611 diagnosed users and 22,426 control users. On average each user in the dataset contributed 16.23 posts with a mean post length of 69 word tokens (see Table 1).

[4]https://www.deepl.com/translator
[5]https://pypi.org/project/langdetect/
[6]r/de is a reddit community for german speakers

MHC	#users	#posts	mean #posts/user	mean #words/post	sd #words/post
ADHD	1055	19212	18.21	59.50	119.35
Anxiety	14	277	19.79	263.85	557.50
Bipolar	1424	23711	16.65	46.46	84.76
Control	22426	361670	16.13	42.08	56.97
Depression	975	15654	16.06	48.12	110.92
OCD	257	3881	15.10	44.02	111.54
Other	1072	17591	16.41	46.67	86.47
PTSD	728	11684	16.05	44.25	74.39
Schizophrenia	86	1380	16.05	44.64	66.60

Table 1: Means (standard deviations) and counts of posts, tokens and characters for diagnosed and control users.

3.2 Ethics and privacy

Although we rely solely on publicly available Reddit data, mental health remains a sensitive issue, and measures to avoid risks to individuals in social media research should always be considered (Hovy and Spruit, 2016; Šuster et al., 2017; Cohan et al., 2018). Following the data handling procedures of the original SMHD (Cohan et al., 2018), we do not publish excerpts from the data, we did not attempt to contact users, and we did not attempt to identify or link users to other social media accounts. We also replace usernames with random identifiers to prevent users' identities from being revealed without external information. The SMHD-GER dataset is made available through a data usage agreement (DUA) that protects user privacy. Specifically, the DUA specifies that no attempt may be made to publish any part of the dataset (which could lead to user identification), contact users, identify them, or link them to other user information.

An ethical issue raised by an anonymous reviewer concerns the annotation of positive mental health conditions through self-disclosure of users, as those who choose to disclose them might differ from the population of individuals living with such conditions without disclosing them. Another ethical issue concerns the use of psychometric evaluation of large text corpora leveraging LIWC-like features alone, as this approach may lack precision: Since LIWC's diagnostic scores are based on both computational correlation and human judgment (in determining the system's dictionaries and word categories), the outcomes may reflect evaluative biases grounded in the context of social, historical, and cultural development (Stark, 2018).

4 Analysis of Linguistic Markers

In this section, we address the exploration of nuanced patterns of linguistic markers that are indicative of specific MHC. We first obtained measurements of 117 engineered language features that can be roughly divided into five groups: (1) features related to morphological and syntactic structural complexity (N=5), (2) features related to lexical sophistication, variety, and richness (N=8), (3) word-level ngram features related to register-specific language use (N=20), (4) features covering the German version of the LIWC (Linguistic Inquiry and Word Count) dictionary (N=68), and (5) word-level dictionary features from three lexicons related to emotion, affect and sentiment (N=16). An overview of these features can be found in Table 5 in the appendix.

The first group of includes surface features related to the length of production units, such as the average length of clauses and sentences, and the type and frequency of embedded structures, such as mean length of sentence or number of dependent clauses per sentence (Lu, 2010). This group also includes an information-theoretic feature based on the Deflate algorithm (Deutsch, 1996).

The second group of features probing lexical density features, such as the ratio of the number of lexical (as opposed to grammatical) words to the total number of words in a text, lexical variation, i.e. the range of vocabulary as manifested in language use as captured by text-size (corrected) type-token ratio (Lu, 2012).

The third group comprises register-based n-gram frequency features that take into account both frequency rank and the number of word n-grams ($n \in [1, 5]$). The latter were derived from four corpora compiled as to represent language use in four language registers (academic, fiction, news, spoken; see Table 6.

The fourth feature group is based on the German version of the LIWC dictionary (Linguistic Inquiry and Word Count) (Pennebaker et al., 2001).

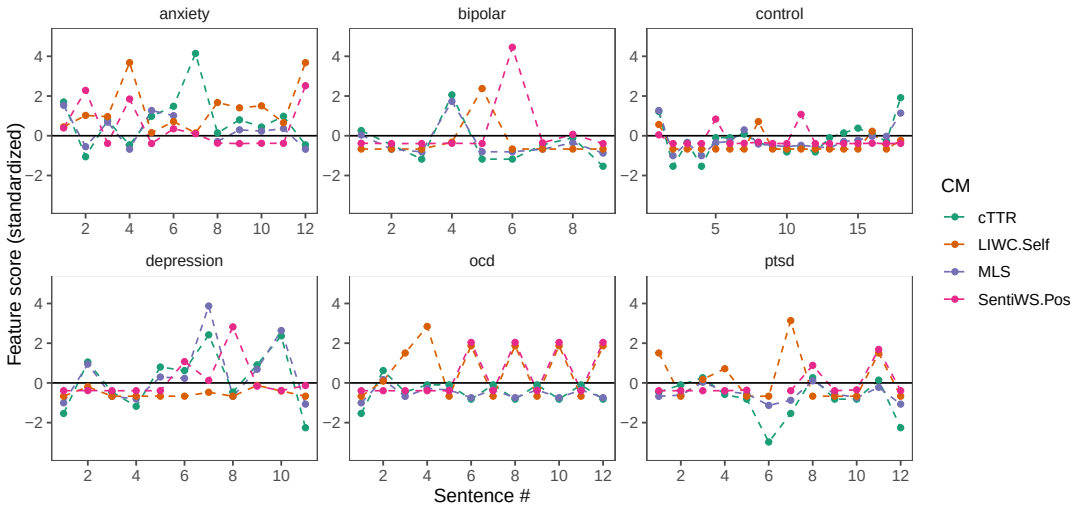

Figure 1: Within-text distributions of four textual features for six randomly selected Reddit post from as many MHC groups (cTTR = corrected Type-Token Ratio, LIWC.Self = self-focused language, MLS = mean length of sentence, SentiWS.Pos = words with positive semantic orientation). All features were z-standardized with 0 representing the corpus average.

The creators of the German version of the LIWC validated this version and demonstrated that the German LIWC categories have a high degree of equivalence to their English counterparts (Wolf et al., 2008). Building on the results of previous studies using LIWC categories for MHC detection in English (e.g. Cohan et al., 2018), we expect that subcategories of particular interest for the MHC classification task will include words with positive or negative emotions, words related to social processes (family/friends/society), pronouns that can capture inclusive (we, us) or exclusionary (you, they, them) language use, and words related to how the person feels (sad, anxious).

The fifth group includes features from three lexicons: MEemoLon (Buechel et al., 2020) is a lexicon comprising eight emotional variables with more than 100k lexical entries for eight emotional variables: Valence, Arousal, Dominance, and Joy, Anger, Sadness, Fear, and Disgust. ANGST is the German adaptation of the Affective Norms for English Words (Schmidtke et al., 2014). It comprises 1,003 German translations of the ANEW material that were were rated on a total of six dimensions: the three original scales for valence, arousal, and dominance plus three additional arousal ratings on an adapted scale. SentiWS (Remus et al., 2010) is a dictionary containing 3,468 sentiment bearing German words (1,650 negative and 1,818 positive) across four word classes (adjectives, adverbs, nouns and verbs) along with their weighted sentiment scores.

All measurements of these features were obtained using an automated text analytics system that employs a sliding window technique to compute measurements at the level of individual sentences. These measurements capture the within-text distributions of scores for a given feature (for recent applications, see e.g. Wiechmann et al., 2022 or Kerz et al., 2022). Tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic PCFG parsing were performed using Stanford CoreNLP (Manning et al., 2014). Examples of these within-text distributions is shown in Figure 1. Each of panels in Figure 1 shows the distributions of four of the 117 textual features for one of six randomly selected texts representing different MHC groups. We note that the distribution of feature values is generally not uniform, but shows large fluctuations over the course of the text. The six texts are characterized by different patterns of spikes of specific features: For example, the bipolar text exhibits a large spike in the SentiWS.Pos feature, which refers to words with positive semantic orientation. The OCD text is characterized by regular peaks of the LIWC.Self feature, which captures self-focused language. The anxiety text displays frequent spikes of high values (>2 standard deviations from corpus average) for three of the four features. In comparison, the control text shows less fluctuation with features scores being closer to the corpus average values. The classification mod-

els described in Section 5.1 are designed to detect and exploit these fluctuations for the detection of specific MHCs. The average scores of all features across all groups are provided in Table 8 in the appendix.

To identify profiles of language use that are characteristic of particular MHC, we compare these feature scores across users in each MHC group using factorial analyses of variance (ANOVA). We focus on those features that display significant differences across groups (N=16, for alpha = 0.05). Figure 2 presents a cluster heatmap visualizing the patterns in the data matrix with the MHC groups and the 16 most significant language features.

The results of these analyses revealed some interesting patterns of differential language use: We find that the control group is situated at the margin of the clustering, indicating that the patterns of language use of diagnosed MHC are distinguishable from this baseline.

The language use of anxiety is distinctly different from all other MHC. It is characterized by very high feature scores on five LIWC dimensions related words referring to self-reference, death and sadness. They are further characterized by high scores on the top feature cluster, comprising words referring to anger, fear, disgust, sadness, arousal and negative emotions.

The language use of schizophrenia, is similar to anxiety in that it too displays a larger proportion of words indicating negative emotions. However, it is characterized by low scores on LIWC dimensions related words referring to self-reference. They are also characterized by low scores on the n-gram frequency features, indicating dependence on conventional phrases from specific speech registers.

A striking feature of obsessive-compulsive disorder (OCD) is its heavy reliance on such terms. A characteristic feature of (unipolar) depression is a markedly increased use of words with positive semantic orientation, in stark contrast to bipolar depression, which has significantly lower scores on this dimension. This is intriguing in light of the fact that distinguishing between bipolar disorder and recurrent unipolar depression is a major clinical challenge (de Almeida and Phillips, 2013). In general, conditions of depression and bipolar disorder, attention-deficit/hyperactivity disorder (ADHD) and post traumatic stress disorder (PTSD) display similar patterns of language use.

These findings reflect evidence in the psychiatric literature indicating that there is considerable overlap in clinical symptoms and pathophysiological processes and that depressive symptoms may also occur in the context of another psychiatric disorder (e.g., bipolar disorder) (Baldwin et al., 2002). Furthermore, psychiatric data suggest that depressive disorders (i.e., major depressive disorder and dysthymia) are highly comorbid with other common mental disorders (Rohde et al., 1991; Gold et al., 2020).

MHC	# posts	mean # words	mean # chars
ADHD	1052	168.78	805.78
Bipolar	1421	150.50	853.66
Depression	974	153.89	872.87
PTSD	728	150.57	902.34
Control	12789	158.55	848.23

Table 2: Description statistics of the data used in benchmark experiments. Note: The size of the control data used in the binary MHC classification tasks were adopted to outnumber the positive cases by a factor of 9. The descriptive statistics of the control categories are based on the entire control corpus.

5 Experiments

5.1 Experimental Setup

In this section, we describe MHC detection experiments performed to obtain benchmark models for the SMHD-GER dataset. We conduct binary classification experiments for the top four most frequently attested MHC in the dataset, namely ADHD, bipolar, depression and PTSD. For each MHC, we use a 1:9 ratio of positive cases to controls to create a more realistic unbalanced classification setting. The the size of the textual input to the models was constrained to fall between 110 words, which corresponds to the median number of words all all posts, and 512 words, which represents a upper limit to the BERT models. In case no single post of a given user satisfied these constraints, we concatenated several posts from that user so that their total amount fell withing the specified boundaries (Figure 5 in the appendix presents a decision tree of the selection method). Table 2 presents the descriptive statistics of the dataset used in classification experiments.

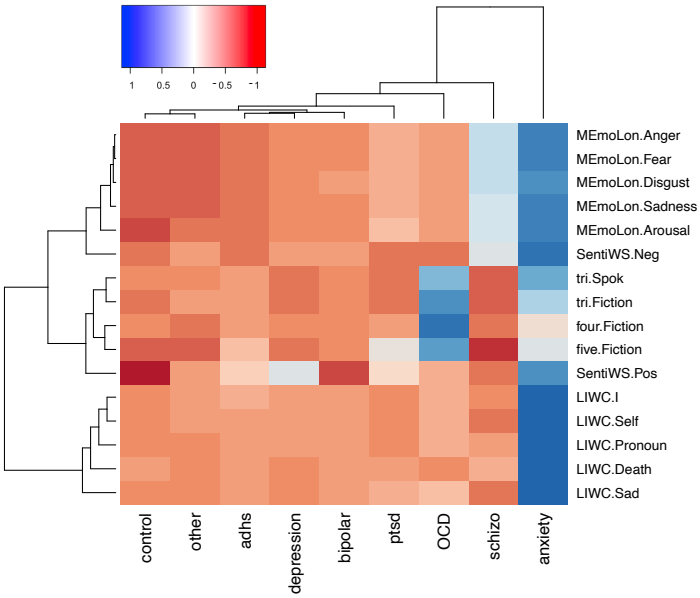

Figure 2: Heatplot of the language profiles of the nine MHC categories (on the x-axis) based on the top-16 language features (on the y-axis). Columns and rows are ordered according to the results of hierarchical clustering, with the dendrograms at the margins showing the groupings of MHC categories and features. Features with the prefix 'MEemoLon' refer to emotion categories from the lexicon of the same name. SentiWS.Neg refers to the category of negative words from the SentiWS dictionary. The terms 'tri', 'four' and ´five' refer to the size of word n-gram features from the spoken ('Spok') and fiction ('Fiction') reference corpora. Features with the prefix 'LIWC' refer to categories from the lexicon of the same name. Colors denote z-standardized feature scores.

5.2 Classification models

We performed experiments with three benchmark models: (1) a fine-tuned German BERT model (GBERT), (2) a bidirectional long short-term memory (BLSTM) classifier trained on measurements of linguistic features described in Section 4, and (3) hybrid model integrating GBERT predictions with the engineered language features introduced in Section 4. For (1) we used the pretrained 'German bert-base-uncased' (GBERT, Chan et al., 2020) model from the Huggingface Transformers library (Wolf et al., 2020) with an intermediate BLSTM layer with 256 hidden units (Al-Omari et al., 2020). For (2) - the model based solely on linguistic features, we constructed a 5-layer BLSTM with a hidden state dimension of 512. The input to that model is a sequence $CM_1^N = (CM_1, CM_2 \ldots, CM_N)$, where CM_i, the output of our text analytics system for the ith sentence of a post, is a 117 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, a sequence length of 5 and learning rate of 1e-3. The architecture of the hybrid classification model - model (3) - consists of two parts: (i) a pre-trained Transformer-based model with a BLSTM layer and FC layer on top of it and (ii) the linguistic 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 specified as follows: 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

Model	Metric	ADHD	Bipol.	Depr.	PTSD
Majority	Pre	44.85	44.85	44.84	44.82
Class	Rec	50.00	50.00	50.00	50.00
Baseline	F1	47.29	47.28	47.28	47.27
	Pre	50.63	50.36	49.74	50.57
GBERT	Rec	50.26	51.02	48.13	46.68
	F1	50.44	50.68	48.92	48.54
	Pre	**56.12**	**54.78**	50.41	50.15
PsyLing	Rec	52.45	**55.31**	50.18	49.86
	F1	**53.22**	**53.97**	50.26	49.92
	Pre	51.29	51.85	**51.62**	**53.20**
Hybrid	Rec	**53.47**	52.03	**50.38**	**52.44**
	F1	53.08	51.91	**50.89**	**53.03**

Table 3: Results of MHC prediction experiments (all values of performance metrics are macro averages)

Model	MHC	TN	FP	FN	TP
GBERT	ADHD	1734	96	187	23
	Bipolar	2466	112	157	20
	Depression	1374	136	361	17
	PTSD	1168	96	132	14
PsyLing	ADHD	1764	66	192	18
	Bipolar	2344	127	253	31
	Depression	1333	276	231	48
	PTSD	939	268	162	41
Hybrid	ADHD	1500	330	161	49
	Bipolar	2289	260	182	24
	Depression	1633	96	138	21
	PTSD	1160	104	127	19

Table 4: Confusion matrices of the three benchmark models (TN: True Negatives, FP: False Positive, FN: False Negative, TP: True Positive)

decay of 1e-4. Structure diagrams of the model based solely on linguistic features and the hybrid architectures are presented in Figures 4 and 3 in appendix. 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.

6 Results and Discussion

Table 3 gives an overview of the results of the MHC prediction experiments. All three baseline models displayed significant improvements in macro F1 scores over the majority baseline for all four MHC. Our PsyLing model consistently outperformed the GBERT baseline in terms of precision, recall and F1 (average improvement F1 = +2.37%; average improvement precision = 2.54%; average improvement recall = +2.93%). This result demonstrates that strong, interpretable mental health detection systems can be built if and when they make full use of the linguistic signals. The PsyLing model achieves highest performance in two of the four MHC, ADHD and bipolar disorder, with improvements over the hybrid model of +2.06% F1 for bipolar and +0.14% F1 for ADHD. However, the hybrid model improves on the performance of the PsyLing model by +3.11% F1 for PTSD and +0.63% F1 for depression.

The results of error analyses shown in Table 4 revealed that these performances were related to the divergent behaviors of the GBERT and PsyLing models for different MHCs: For Depression and PTSD the PsyLing model has a high false alarm rate, i.e. it classified users as being diagnosed, when they are in fact not (Depression: FP_{GBERT}=136 , $FP_{PsyLing}$=276; PTSD: FP_{GBERT}=96, $FP_{PsyLing}$=268). On the other hand, it also correctly identified a much higher proportion of diagnosed users (Depression: TP_{GBERT}=17 , $TP_{PsyLing}$=48; PTSD: TP_{GBERT}=14, $TP_{PsyLing}$=41). Our results thus indicate that the hybrid model improves on the PsyLing model for depression and PTSD by leveraging the lower false alarm rate of GBERT for these MHC. These results demonstrate that the NLP systems designed to support the diagnosis of mental disorders benefit from employing both interpretable and hybrid approaches.

7 Conclusion and Future Work

We introduced SMHD-GER, a large dataset of Reddit users with diverse mental health conditions and matched control users. The dataset was created using adaptations of the high-precision diagnostic patterns developed for the original English version (Cohan et al., 2018). Furthermore, we investigated the differences in language use between users with mental health conditions and control groups, as measured by a large set of linguistic and psychological cues. We provided strong benchmark models designed to identify diagnosed users for the four most frequently attested MHC. We found that BLSTM networks trained on within-text distributions of interpretable linguistic features consistently outperformed a Transformer-based model based on GBERT. A hybrid model combining the two approaches proved to be the most effective

method for two of the four conditions. We make our dataset available to the community in the hope that it will encourage further research into these problems and improve the reproducibility of suggested approaches.

8 Limitations

In this work, we have framed mental health detection as a binary classification task that aims to distinguish between individuals with a particular mental disorder and control users. In future work, we intend to frame it as a multi-class classification task to determine the extent to which individual mental disorders can be distinguished from one another.

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.

David S Baldwin, Dwight L Evans, RM Hirschfeld, and Siegfried Kasper. 2002. Can we distinguish anxiety from depression? *Psychopharmacology Bulletin*, 36:158–165.

Sven Buechel, Susanna Rücker, and Udo Hahn. 2020. Learning and evaluating emotion lexicons for 91 languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217, Online. Association for Computational Linguistics.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. *arXiv preprint arXiv:2010.10906*.

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.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

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.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 31–39.

Jorge Renner Cardoso de Almeida and Mary Louise Phillips. 2013. Distinguishing between unipolar depression and bipolar depression: current and future clinical and neuroimaging perspectives. *Biological psychiatry*, 73(2):111–118.

Munmun De Choudhury, Sanket S Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and cross-cultural differences in social media disclosures of mental illness. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 353–369.

Peter Deutsch. 1996. Rfc1951: Deflate compressed data format specification version 1.3.

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.

Simon D'Alfonso. 2020. Ai in mental health. *Current Opinion in Psychology*, 36:112–117.

Stefan M Gold, Ole Köhler-Forsberg, Rona Moss-Morris, Anja Mehnert, J Jaime Miranda, Monika Bullinger, Andrew Steptoe, Mary A Whooley, and Christian Otte. 2020. Comorbid depression in medical diseases. *Nature Reviews Disease Primers*, 6(1):1–22.

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.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Zheng Ping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020. Detection of mental health from reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

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.

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM conference on Hypertext & Social Media*, pages 85–94.

A Downey La Vonne, Leslie S Zun, and Trena Burke. 2012. Undiagnosed mental illness in the emergency department. *The Journal of emergency medicine*, 43(5):876–882.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 78–87.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

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.

Justus Mattern, Yu Qiao, Elma Kerz, Daniel Wiechmann, and Markus Strohmaier. 2021. Fang-covid: A new large-scale benchmark dataset for fake news detection in german. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 78–91.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Jürgen Rehm and Kevin D Shield. 2019. Global burden of disease and the impact of mental and addictive disorders. *Current psychiatry reports*, 21(2):1–7.

Robert Remus, Uwe Quasthoff, and Gerhard Heyer. 2010. SentiWS - a publicly available German-language resource for sentiment analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Paul Rohde, Peter M Lewinsohn, and John R Seeley. 1991. Comorbidity of unipolar depression: Ii. comorbidity with other mental disorders in adolescents and adults. *Journal of abnormal psychology*, 100(2):214.

Damian F Santomauro, Ana M Mantilla Herrera, Jamileh Shadid, Peng Zheng, Charlie Ashbaugh, David M Pigott, Cristiana Abbafati, Christopher Adolph, Joanne O Amlag, Aleksandr Y Aravkin, et al. 2021. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *The Lancet*, 398(10312):1700–1712.

Max Schindler and Emese Domahidi. 2022. The computational turn in online mental health research: A systematic review. *New Media & Society*, page 14614448221122212.

David S Schmidtke, Tobias Schröder, Arthur M Jacobs, and Markus Conrad. 2014. Angst: Affective norms for german sentiment terms, derived from the affective norms for english words. *Behavior research methods*, 46(4):1108–1118.

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.

Luke Stark. 2018. Algorithmic psychometrics and the scalable subject. *Social Studies of Science*, 48(2):204–231.

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. *arXiv preprint arXiv:1703.10090*.

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.

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.

Markus Wolf, Andrea B Horn, Matthias R Mehl, Severin Haug, James W Pennebaker, and Hans Kordy. 2008. Computergestützte quantitative textanalyse: äquivalenz und robustheit der deutschen version des linguistic inquiry and word count. *Diagnostica*, 54(2):85–98.

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.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Andrew Yates and Nazli Goharian. 2013. ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In *European Conference on Information Retrieval*, pages 816–819. Springer.

Elad Yom-Tov, Evgeniy Gabrilovich, et al. 2013. Postmarket drug surveillance without trial costs: discovery of adverse drug reactions through large-scale analysis of web search queries. *Journal of medical Internet research*, 15(6):e2614.

Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2022. The best of both worlds: Combining engineered features with transformers for improved mental health prediction from reddit posts. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 197–202.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):1–13.

A Appendix

Table 5: Overview of the 117 features investigated in the work.

Feature group	Number of features	Features	Example/Description
Morpho-syntactic	5	MLC	Mean length of clause (words)
		MLS	Mean length of sentence (words)
		C/S	Clauses per Sentence
		CoordP/C	Coordinate phrases per clause
		BaseKolDef	Kolmogorov Complexity
Lexical richness	8	MLWc	Mean length per word (characters)
		LD	Lexical density
		NDW	Number of different words
		cNDW	Corrected number of different words
		TTR	Type-Token Ratio (TTR)
		cTTR	Corrected TTR
		rTTR	Root TTR
		log TTR	Logarithmic TTR
Register-based N-gram	20	Spoken ($n \in [1, 5]$)	Frequencies of uni-, bi-
		Fiction ($n \in [1, 5]$)	tri-, four-, five-grams
		News ($n \in [1, 5]$)	from four reference corpora
		Academic ($n \in [1, 5]$)	(see appendix Table 6)
LIWC	68	LIWC-German	Pennebaker et al. (2001)
Emotion Lexicon	2	SentiWS	Remus et al. (2010)
	6	ANGST	Schmidtke et al. (2014)
	8	MEmoLon	Buechel et al. (2020)

Table 6: Text corpora used to derive register-specific n-gram frequencies

Register	**Corpus**	**Size**		
		Vocab	**# Words**	**Items**
Academic	Papers from top 100 German publications	477876	12M	2524 papers
Fiction	Gutenberg project German books	907656	49M	2063 books
News	News articles from FANG-Covid corpus (authentic news) (Mattern et al., 2021)	487841	21M	28056 articles
Spoken	OpenSubtitle dataset	1209934	218M	

Table 7: Descriptive statistics of feature groups 1-5 across MHC.

Feature	**Control**	**ADHD**	**Anxiety**	**Bipol**	**Depres.**	**OCD**	**PTSD**	**Schiz.**	**other**
LD	0.56	0.56	0.51	0.56	0.56	0.56	0.55	0.55	0.56
TTR	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
cTTR	4.27	4.27	4.57	4.28	4.26	4.23	4.30	4.29	4.31
logTTR	0.93	0.94	0.97	0.93	0.93	0.93	0.94	0.94	0.93
rTTR	2.92	2.92	3.17	2.91	2.91	2.89	2.94	2.97	2.94
NDW	18.11	17.90	20.05	17.98	17.98	17.48	18.13	18.45	18.42
cNDW	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
MLW	5.26	5.30	5.27	5.28	5.30	5.29	5.31	5.42	5.32

Continued on next page

Table 7 – continued from previous page

Feature	Control	ADHD	Anxiety	Bipol	Depres.	OCD	PTSD	Schiz.	other
CoordPpC	0.25	0.25	0.18	0.24	0.24	0.25	0.25	0.25	0.24
MLC	7.09	7.07	7.07	7.03	6.99	6.91	7.19	7.27	6.99
MLS	20.00	19.69	21.81	19.76	19.95	19.74	19.83	20.26	20.36
ClpS	2.52	2.47	3.14	2.48	2.54	2.53	2.50	2.57	2.61
KD	1.00	1.00	0.91	1.00	1.00	1.01	0.99	0.99	1.00
uni.Acad	108.74	105.51	129.07	107.04	107.04	109.27	106.57	108.90	109.16
bi.Acad	8.88	8.53	10.98	8.73	8.61	9.39	8.64	8.78	8.80
tri.Acad	0.20	0.20	0.23	0.22	0.21	0.23	0.23	0.20	0.21
four.Acad	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01
uni.Fiction	137.06	133.33	165.51	135.23	135.20	139.62	134.44	139.88	137.58
bi.Fiction	19.30	18.52	27.95	18.95	18.55	23.01	18.19	18.87	19.06
tri.Fiction	0.88	0.83	1.26	0.86	0.81	1.47	0.81	0.78	0.87
four.Fiction	0.03	0.03	0.05	0.03	0.03	0.07	0.03	0.03	0.03
uni.News	135.13	131.42	159.84	133.30	133.25	135.85	132.93	137.82	135.76
bi.News	19.17	18.35	27.74	18.79	18.49	21.03	18.19	19.13	18.87
tri.News	0.92	0.85	1.25	0.90	0.84	1.08	0.83	0.86	0.88
four.News	0.04	0.04	0.04	0.06	0.04	0.05	0.04	0.03	0.04
uni.Spok	160.79	156.36	194.64	158.58	159.34	163.27	157.73	163.88	161.56
bi.Spok	27.26	26.07	41.32	26.61	26.43	32.79	25.65	26.10	26.89
tri.Spok	1.78	1.63	3.30	1.73	1.62	3.19	1.59	1.52	1.69
four.Spok	0.12	0.10	0.22	0.15	0.10	0.26	0.10	0.07	0.11
five.Spok	0.01	0.01	0.04	0.03	0.01	0.02	0.01	0.00	0.01

Table 8: Descriptive statistics of feature scores across MHC.

Feature	Control	ADHD	Anxiety	Bipol	Depres.	OCD	PTSD	Schiz.	other
LIWC.Pronoun	0.08	0.08	0.13	0.08	0.08	0.08	0.08	0.08	0.08
LIWC.I	0.04	0.04	0.08	0.04	0.04	0.04	0.04	0.04	0.04
LIWC.Self	0.04	0.04	0.08	0.04	0.04	0.05	0.04	0.04	0.04
LIWC.You	0.01	0.01	0.02	0.01	0.01	0.01	0.02	0.02	0.01
LIWC.Other	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
LIWC.Negate	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02
LIWC.Assent	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.00	0.01
LIWC.Article	0.08	0.08	0.07	0.08	0.08	0.08	0.09	0.09	0.08
LIWC.Preps	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
LIWC.Affect	0.06	0.05	0.07	0.05	0.05	0.05	0.05	0.06	0.05
LIWC.Posemo	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
LIWC.Posfeel	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LIWC.Optim	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
LIWC.Negemo	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.01
LIWC.Sad	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00
LIWC.Cogmech	0.10	0.10	0.11	0.10	0.10	0.10	0.10	0.10	0.10
LIWC.Cause	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.02
LIWC.Insight	0.03	0.03	0.04	0.03	0.03	0.03	0.02	0.03	0.03
LIWC.Discrep	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
LIWC.Inhib	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
LIWC.Tentat	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.02

Continued on next page

Table 8 – continued from previous page

Feature	Control	ADHD	Anxiety	Bipol	Depres.	OCD	PTSD	Schiz.	other
LIWC.Certain	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
LIWC.Social	0.07	0.07	0.09	0.07	0.07	0.07	0.07	0.07	0.07
LIWC.Comm	0.02	0.02	0.03	0.02	0.02	0.02	0.02	0.02	0.02
LIWC.Othref	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
LIWC.Friends	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
LIWC.Humans	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01
LIWC.Time	0.04	0.04	0.07	0.04	0.04	0.04	0.04	0.04	0.04
LIWC.Past	0.03	0.03	0.04	0.03	0.03	0.02	0.03	0.03	0.03
LIWC.Present	0.08	0.08	0.11	0.08	0.08	0.08	0.08	0.08	0.08
LIWC.Future	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01
LIWC.Space	0.07	0.07	0.06	0.07	0.07	0.07	0.07	0.07	0.07
LIWC.Up	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
LIWC.Incl	0.05	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.05
LIWC.Excl	0.02	0.03	0.02	0.03	0.03	0.03	0.03	0.02	0.03
LIWC.Motion	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
LIWC.Occup	0.05	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.06
LIWC.School	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
LIWC.Job	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02
LIWC.Achieve	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
LIWC.Leisure	0.01	0.02	0.02	0.01	0.02	0.02	0.02	0.01	0.02
LIWC.Home	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01
LIWC.Money	0.01	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.01
LIWC.Metaph	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
LIWC.Physcal	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
LIWC.Body	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.00	0.01
Angst.AROANEW	0.15	0.15	0.14	0.15	0.15	0.13	0.14	0.15	0.15
Angst.AROBAWL	0.08	0.08	0.08	0.08	0.08	0.07	0.08	0.08	0.08
Angst.DOM	0.18	0.17	0.16	0.18	0.17	0.16	0.17	0.16	0.17
Angst.IMA	0.14	0.13	0.12	0.13	0.13	0.12	0.13	0.12	0.13
Angst.POT	0.17	0.17	0.16	0.17	0.17	0.15	0.17	0.17	0.17
Angst.VAL	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02
MEmoLon.Anger	1.38	1.38	1.43	1.38	1.38	1.38	1.39	1.41	1.38
MEmoLon.Arousal	3.70	3.71	3.81	3.71	3.71	3.72	3.73	3.76	3.71
MEmoLon.Disgust	1.37	1.37	1.42	1.38	1.37	1.38	1.38	1.40	1.37
MEmoLon.Dominance	5.06	5.07	5.18	5.07	5.07	5.08	5.10	5.10	5.06
MEmoLon.Fear	1.40	1.40	1.45	1.40	1.40	1.40	1.41	1.43	1.40
MEmoLon.Joy	1.99	1.99	2.05	1.99	1.99	1.99	2.00	1.99	1.99
MEmoLon.Sadness	1.33	1.33	1.39	1.34	1.34	1.34	1.34	1.36	1.33
MEmoLon.Valence	4.98	4.98	5.09	4.98	4.99	5.00	5.01	4.99	4.98
SentiWS.Pos	0.06	0.06	0.07	0.06	0.07	0.06	0.06	0.06	0.06
SentiWS.Neg	0.06	0.06	0.11	0.06	0.06	0.06	0.06	0.09	0.06

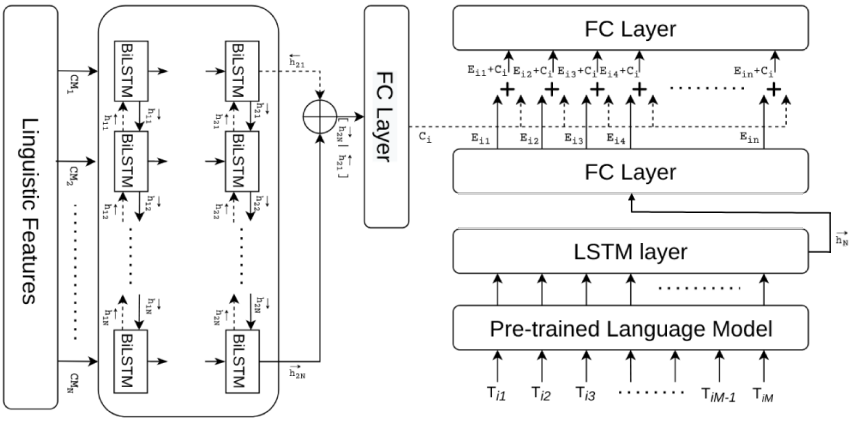

Figure 3: Hybrid model structure

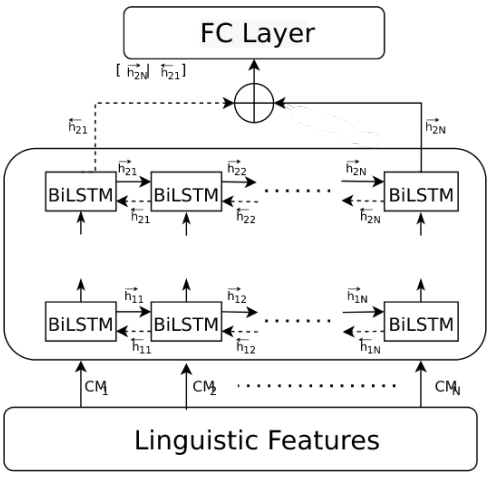

Figure 4: PsyLin model structure

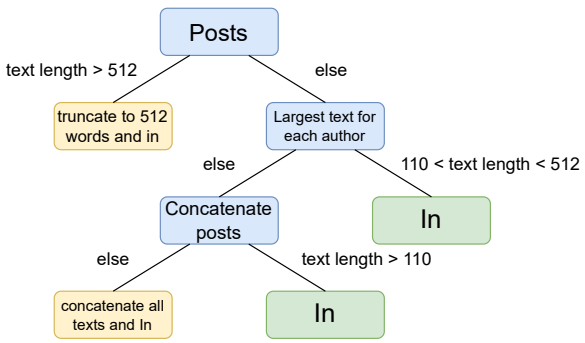

Figure 5: Decision tree for selecting experimental data.

Table 9: Results of MHC prediction experiments (micro scores)

		Mental Health Condition			
Model type	Metric	ADHD	Bipolar	Depression	PTSD
GBERT	Pre	19.12	15.14	11.10	13.24
	Rec	11.43	11.32	4.50	10.07
	F1	14.28	13.29	7.34	11.20
PsyLing	Pre	20.88	19.67	14.81	13.26
	Rec	9.49	11.03	17.18	20.18
	F1	13.34	14.28	15.90	16.00
Hybrid	Pre	13.22	8.45	17.95	15.37
	Rec	22.76	11.84	13.21	12.78
	F1	17.52	10.92	15.72	14.55