What to Fuse and How to Fuse: Exploring Emotion and Personality Fusion Strategies for Explainable Mental Disorder Detection

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

Daniel Wiechmann
University of Amsterdam
d.wiechmann@uva.nl

Xiaofei Li
RWTH Aachen University
xiaofei.li1@rwth-aachen.de

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

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

Abstract

We present the results of conducting extensive experiments with three types of deep learning-based fusion strategies: (1) feature-level fusion, where a pre-trained masked language model for mental health detection (MentalRoBERTa) was infused with a comprehensive set of engineered features, (2) model fusion, where the MentalRoBERTa model was infused with hidden representations of other language models and (3) task fusion, where a multi-task framework was leveraged to learn the features for auxiliary tasks. In addition to exploring the role of different fusion strategies, we extend previous work by broadening the information infusion to include a second domain related to mental health, i.e. personality. We evaluate the performance of our models on two benchmark mental health datasets encompassing five conditions: Attention Deficit Hyperactivity Disorder, Anxiety, Bipolar Disorder, Depression, and Psychological Stress. The results of our experiments show that the task fusion strategy is most promising for the detection of ADHD, anxiety, and bipolar disorder, whereas feature-level fusion is most advantageous for the detection of psychological distress and depression. Moreover, the results indicate that both emotion and personality constitute valuable sources of information for predicting mental health.

1 Introduction

Mental health disorders (MHD) are increasingly prevalent worldwide and constitute one of the greatest challenges facing our healthcare systems and modern societies in general. In response to this societal challenge, there has been a surge in digital mental health research geared towards the development of new techniques for unobtrusive and efficient automatic detection of MHD. Within this area of research, natural language processing techniques are playing an increasingly important role, showing promising detection results from a variety of textual data. Recently, there has been a growing interest in improving mental illness detection from textual data by way of leveraging emotions: ‘Emotion fusion’ refers to the process of integrating emotion information with general textual information to obtain enhanced information for decision-making. However, while the available research has shown that MHD prediction can be improved through a variety of different fusion strategies, previous works have been confined to a particular fusion strategy applied to a specific dataset, and so is limited by the lack of meaningful comparability.

As a result, the clinical community is increasingly seeking new approaches to the early detection and monitoring of mental health problems that can greatly improve the effectiveness of interventions, reduce their cost, and prevent them from becoming chronic. In this context, Natural Language Processing (NLP) is recognized as having transformative potential to support healthcare professionals and stakeholders in the early detection, treatment and prevention of mental disorders (for comprehensive reviews, see Calvo et al., 2017; Zhang et al., 2022; Zhou et al., 2022). Data from social media are particularly appealing to the NLP research community due to their scope and the deep embeddedness in contemporary culture (Perrin). Research utilizing NLP techniques in combination with social media has yielded new insights into population mental health and shown promise for incorporating data-driven analytics into the treatment of psychiatric disorders (Chancellor and De Choudhury, 2020; Garg, 2023).

Recently, this line of research has developed a growing interest in improving NLP approaches to mental illness detection by leveraging information from related domains, in particular emotion (see Zhang et al., 2023, for a comprehensive review). Behavioral and psychological research has long established links between emotions and mental disorders: For example, individuals with depressive symptoms have difficulty regulating their emotions,

Findings of the Association for Computational Linguistics: ACL 2023, pages 8926–8940
July 9-14, 2023 ©2023 Association for Computational Linguistics

resulting in lower emotional complexity (Joormann and Gotlib, 2010; Compare et al., 2014). Disrupted emotion regulation has also been implicated in anxiety (Young et al., 2019). In the light of such links, information about emotions is useful in diagnosing mental disorders. 'Emotion fusion' refers to the process of "integrating emotion information with general textual information to obtain enhanced information for decision-making" (Zhang et al., 2023, p. 232). By the same rationale, information fusion approaches are likely to benefit from the inclusion of additional individual characteristics known to be associated with mental disorders, such as personality traits. Like emotion, personality has been linked to a diverse set of mental disorders based on genetic and behavioral evidence: For example, genome-wide association studies have demonstrated that genetic risk factors for depression are largely shared with the neuroticism peronality trait (Adams et al., 2019). Correlational studies comparing subjects diagnosed with Major Depressive Disorder (MDD) and healthy control subjects found that vulnerability to depression was associated with several personality dimensions, such that MDD subjects were characterized by high neuroticism and low extraversion, accompanied by low scores on openness and conscientiousness (Nikolic et al., 2020). Analyses language of use of Twitter users with self-disclosed depression and PTSD revealed that text-derived personality played s an important role in predicting the mental disorders (Preoţiuc-Pietro et al., 2015).

In addition to the question of 'what to fuse', information fusion approaches also raise the algorithmic question of 'how to fuse' the auxiliary information effectively. The available research has shown that MHD prediction can be improved through a variety of different fusion strategies. However, previous work has typically focused on a specific fusion strategy applied to a specific dataset, limiting their comparability.

In this work, we integrate and extend research on information fusion for mental disorder detection by conducting extensive experiments with three types of deep learning-based fusion strategies: (i) feature-level fusion, where a pre-trained masked language model for mental health detection (MentalRoBERTa, htt) was infused with a comprehensive set of engineered features, (ii) model fusion, where the MentalRoBERTa model was infused with hidden representations of other language models and (iii) task fusion, where a multi-task framework was leveraged to learn the features for auxiliary tasks. In addition to exploring the role of different fusion strategies, we expand on previous work by broadening the information infusion to include a second domain related to mental health, i.e. personality. We evaluate our model on data from two benchmark datasets, encompassing five mental health conditions: attention deficit hyperactivity disorder, anxiety, bipolar disorder, depression and psychological stress.[1]

The remainder of the paper is structured as follows: Section 2 presents a concise discussion of related work applying each of the three information fusion strategies. Section 3 introduces the datasets used to perform the mental health detection experiments. In Section 4, we describe our three mental status detection models that instantiate the three fusion strategies. Section 5 details the experimental setup including the specification of the fine-tuned MentalRoBERTa model baseline model. Section 6 presents and discusses the results of our experiments. Finally, we conclude with possible directions for future work in Section 7.

´

2 Related work

In this section we provide a concise discussion of selected works for each of the three fusion strategies. A comprehensive overview of work information fusion for mental illness detection from social media data has recently been provided by Zhang et al. (2023). One strand of recent work in the feature-level fusion approach is characterized by the integration of information from several groups of features extracted using NLP tools: Song et al. (2018) utilized a feature attention network (FAN) to combine indicators of mental disorders from four groups: (1) word-level features related to depressive symptoms taken from the Diagnostic and Statistical Manual of Mental Disorders (DSM-5, APA, 2013), (2) word-level sentiment scores of obtained from the SentiWordNet dictionary (Baccianella et al., 2010), (3) features related ruminative thinking, expressed as the amount of repetition of topics in a social media post (Nolen-Hoeksema et al., 2008) and (4) writing style features, measured in terms of the sequencing of part-of-speech in a social media. The FAN consists of four feature networks - one for each feature groups - fed into a post-level attention layer. The authors eval-

[1] Our code will be made available upon publication.

uated the performance of their approach on the Reddit Self-reported Depression Diagnosis dataset (RSDD, Yates et al. (2017)), a large scale general forum dataset contaning data from 9,210 users with an average of 969 posts for each user. Their model was competitive with a convolutional neural network baseline model, despite using a much smaller number of posts in training data (only 500 posts per user). A second strand of feature-fusion approaches combines emotion features extracted using NLP tools with textual embeddings from pre-trained language models, before feeding these into a CNN/LSTM structure to construct the MHC classification model. For example, Uban et al. (2021) used a hierarchical attention network with LSTM post-level and user-level encoders that combined multi-dimensional representations of texts. Specifically, their approach combined (i) content features, captured through word sequences encoded as 300-dimenional embeddings based on pre-trained GloVe vectors (Pennington et al., 2014), (ii) style features, expressed by numerical vectors representing stopword frequencies as bag-of-words, normalized by text lengths and usage of pronouns or other parts of speech, and (iii) emotion and sentiment features, represented by numerical vectors of word category ratios from two emotion- and sentiment-related lexicons, LIWC (Pennebaker et al., 2001) and NRC emotion (Mohammad and Turney, 2013). They evaluated the model on the eRisk Reddit datasets on depression, anorexia and self-harm (Losada et al., 2019), reaching competitive result across all three mental disorders, outperforming a strong RoBERTa baseline model in the detection of two of them (self-harm and depression).

Turning to the model fusion approach, Sawhney et al. (2020) presented a time-aware transformer based model for the screening of suicidal risk on social media. Their model, called STATENet, uses a dual transformer-based architecture to learn the linguistic and emotional cues in tweets. STATENet combines the 768-dimensional encoding obtained from Sentence BERT, capturing the language cues of the tweet to be assessed, with an aggregate representation of the emotional spectrum, obtained from a pre-trained BERT model fine-tuned on the the Emonet dataset (Abdul-Mageed and Ungar, 2017). This second model, referred to as the Plutchik Transformer, tokenizes each post and adds the [CLS] token at the beginning of each post. The authors then express the the aggregate representation of the emotional spectrum as the the final hidden state corresponding to this [CLS] token (768-dimensional encoding). They evaluated the STATENet models on the task of tweet-level prediction of suicide idation on the Twitter timeline dataset (Sinha et al., 2019), which contained 32,558 tweets. STATENet significantly outperforms competitive baselines models for suicidal risk assessment, demonstrating the utility of combining contextual linguistic and emotional cues for suicide risk assessment.

Recently, Turcan et al. (2021) explored the use of multi-task learning and emotion-infused language model finetuning for psychological stress detection. In this work, the authors introduced an innovative task fusion approach that utilized a multi-task learning setup to perform stress detection and emotion detection at the same time on the same input data. As currently available datasets for stress detection are not labeled for emotion, they first separately trained BERT models on different versions of the GoEmotions dataset (Demszky et al., 2020) and employed these to derive emotion labels for the stress detection dataset used in their experiments (Dreaddit, Turcan and McKeown, 2019). The authors then used these emotion labels as ‘silver data’ to train on them alongside stress in a multi-task learning setting with hard parameter sharing (Caruana, 1997). Their models achieved comparable performance to a state-of-the-art fine-tuned BERT baseline. Importantly, based on analyses designed to probe their models and discover what information they learn to use, the authors demonstrated that their task fusion approach improved the explainabilty of deep learning-sbased mental health prediction models. Specifically, by performing correlational analyses of the models predictions on each task, they were able to explore the usefulness of the emotion prediction layers in explaining stress classifications.

As can be seen from this overview, with the exception of Turcan et al. (2021), previous studies have focused on specific fusion strategies applied to a variety of mental health conditions. By applying different fusion strategies to five mental disorders (AHDH, anxiety, bipolar disorder, depression) and related symptomatology (psychological stress), we aim to facilitate the evaluation of current approaches to information fusion.

Mental Health	Dataset	Number	Avg. length	SD	Total	Avg. length	SD	Total
Condition		of posts	(words)	(words)	(words)	(chars)	(chars)	(chars)
ADHD	SMHD	5272	117.98	121.64	621992	638.60	677.77	3366710
Anxiety		4963	116.45	132.17	577925	619.73	711.21	3075701
Bipolar		3632	116.56	114.15	423342	622.31	624.05	2260240
Depression		7818	114.70	113.11	896735	610.82	608.08	4775377
Control		10000*	97.0	84.8	969580	525	522	5251129
Stress	Dreaddit	1857	93.0	35.3	172782	459.31	178.50	852949
Control		1696	85.5	29.9	145081	434.91	154.62	737622

Table 1: Count of posts, tokens and characters along with average post length for diagnosed and control users. ***NOTE**: In all binary classification tasks, the control set consisted of a randomly drawn subset of control users that matched the size of the respective positive class.

3 Data

Four datasets were used in the present work: The data used in the task of mental health detection were obtained from two publicly available social media datasets: (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 from data from Reddit, a social media platform consisting of individual topic communities called subreddits, including those relevant to MHC detection. The statistics of these datasets are provided in Table 1.

SMHD is a large dataset of social media posts from users with nine mental health conditions (MHC) corresponding to branches in the DSM-5, an authoritative taxonomy for psychiatric diagnoses (APA, 2013). 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 and 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.

Dreaddit is a dataset of 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 crowdsourced annotations aggregated as the majority vote from five annotators for each data point.

As the SMHD and Dreaddit datasets are labeled only with mental health status, two additional datasets were used to provide auxiliary information about personality and emotion. Following the approach used in Turcan et al. (2021), we first separately trained RoBERTa models on the GoEmotions dataset (Demszky et al., 2020) and the Kaggle MBTI dataset (Li et al., 2018) and used these models to predict emotion and personality labels for SMHD and Dreaddit. A table with dataset statistics for these resources is provided in the appendix.

GoEmotions is the largest available manually annotated dataset for emotion prediction. It consists of 58 thousand Reddit comments, labeled by 80 human raters for 27 emotion categories plus a neutral category. The authors provided a mapping of these 27 categories to Ekman's six basic emotions (anger, disgust, fear, joy, sadness, and surprise), which are assumed to be physiologically distinct (Ekman, 1992, 1999). Drawing on the results of experiments with different emotion mappings reported in Turcan et al. (2021), these six basic emotions are used in the present work.

The Kaggle MBTI dataset was collected through the PersonalityCafe forum[2] and thus provides a diverse sample of people interacting in an informal online social environment. It consists of samples of social media interactions from 8675 users, all of whom indicated their Myers–Briggs Type Indicator (MBTI) personality type (Meyers et al., 1990). The MBTI is a widely administered questionnaire that describes personality in terms of 16 types that result from combining binary categories from four dimensions: (a) Extraversion/Introversion (E/I) - preference for how people direct and receive their energy, based on the external or internal world, (b) Sensing/Intuition (S/N) - preference for how people take

[2] https://www.personalitycafe.com/

in information, through the five senses or through interpretation and meanings, (c) Thinking/Feeling (T/F) - preference for how people make decisions, relying on logic or emotion over people and particular circumstances, and (d) Judgment/Perception (J/P) - how people deal with the world, by ordering it or remaining open to new information.

3.1 Data preprocessing

For the SMHD dataset, we removed all posts with a length greater than 512 words, as these posts could not be processed by the large pre-trained models like RoBERTa and its variants. We then randomly sampled one post from each user and focused our analysis on the four most frequently attested mental health conditions. Furthermore, all dtasets were subjected to various standard pre-processing steps, including removal of HTML, URLs, extra spaces and emojis in the text, and the correction of inconsistent punctuation.

4 Models

We experiment with seven information-infusion models that differ (i) in the type of information to be infused (personality, emotion, both) and (ii) the fusion strategy applied to incorporate that information into the mental health detection models. The architectures of these models is shown in Figure 1.

4.1 Feature-level fusion

Our feature fusion model combines a MentalRoBERTa model (Ji et al., 2022) with a bidirectional long short-term (BiLSTM) network trained on 544 psycholinguistic features that fall into six broad categories: (1) features of morpho-syntactic complexity (N=19), (2) features of lexical richness, diversity and sophistication (N=52), (3) stylistic features (incl. register-based n-gram frequency features (N=57), (4) readability features (N=14), and (5) lexicon features designed to detect sentiment, emotion and/or affect (N=325). (6) Cohesion and Coherence features (N=77). 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-movement patterns during reading and Kerz et al. (2022) for detection of Big Five personality traits and Myers–Briggs types). Tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic PCFG parsing were performed using Stanford CoreNLP (Manning et al., 2014). The given text is fed to a pre-trained language model and its output is passed through a BiLSTM layer with 2 layers and hidden size of 512. The second part of the model is the PsyLin model which is a 3-layer BiLSTM with hidden size of 1024 which is further passed through a fully connected layer to obtain a 256 dimensional vector. The input to this model is a set of over 600 handcrafted features across 5 categories. We constructed the feature-level fusion models by (1) obtaining a set of 256 dimensional vector from the BiLSTM network and then (2) concatenating these features along with the output from the Mental RoBERTa model component. This is then fed into a 2-layer feedforward classifier. To obtain the soft labels (probabilities that a text belongs to the corresponding emotion label), sigmoid was applied to each dimension of the output vector.

4.2 Model fusion

In our model fusion approach, the MentalRoBERTa model was infused with hidden features of a fine-tuned RoBERTa emotion model and fine-tuned RoBERTa personality model (see also Section 3). Both these models are fine-tuned 'roberta-base' models with a linear classification layer on top of them. We use the output values obtained from this layer to provide the infused model information on emotion and/or personality. Specifically, we pass the output obtained from the MentalRoBERTa through a sequential layer consisting of two linear layers and concatenate the features with the second part. We finally pass this through a linear layer to obtain the soft predictions for the respective MHC. Similar to the previous model types, we train separate models for all five MHCs. For each MHC, we created three different binary classification models: one with just emotions (MentalRoBERTa + Emotion), one with just personality (MentalRoBERTa + Personality), and one with 'full infusion' (MentalRoBERTa + Emotion + Personality).

4.3 Task fusion

Our task fusion approach is an extended version of the multi-task learning setup used Turcan et al. (2021). Within this setup, we perform multiple tasks at the same time using the same input data. As the SMHD data is labeled only with MHC cate-

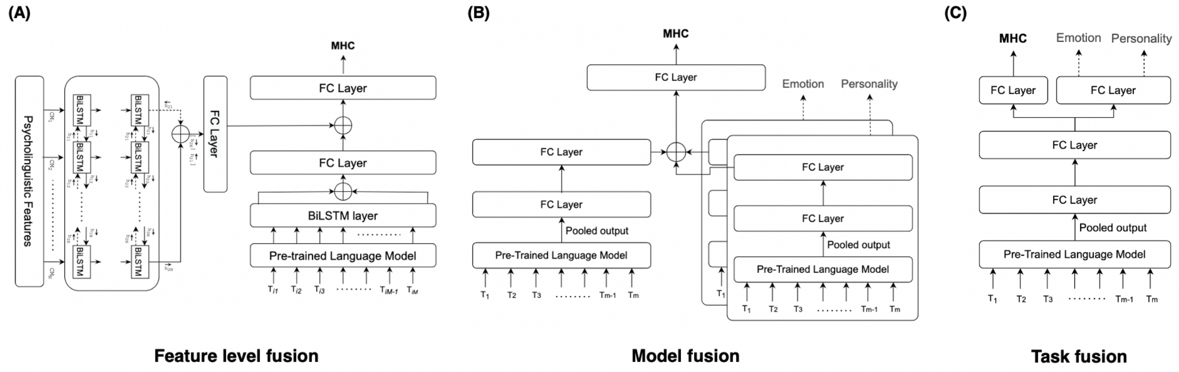
(A)
Psycholinguistic Features
CM_1
CM_2
CM_N
BiLSTM
BiLSTM
BiLSTM
BiLSTM
BiLSTM
BiLSTM
FC Layer
MHC
FC Layer
FC Layer
BiLSTM layer
Pre-trained Language Model
T_{i1} T_{i2} T_{i3} .... T_{iM-1} T_{iM}
Feature level fusion
(B)
MHC
FC Layer
FC Layer
FC Layer
Pooled output
Pre-Trained Language Model
T_1 T_2 T_3 .... T_{m-1} T_m
Emotion
Personality
FC Layer
FC Layer
Pooled output
Pre-Trained Language Model
T_1 T_2 T_3 .... T_{m-1} T_m
Model fusion
(C)
MHC
Emotion
Personality
FC Layer
FC Layer
FC Layer
FC Layer
Pooled output
Pre-Trained Language Model
T_1 T_2 T_3 .... T_{m-1} T_m
Task fusion

Figure 1: Information fusion architectures used in our experiments

Emotion	F1-score	Personality	F1-score
Anger	55	I/E	74
Disgust	39	N/S	83
Fear	61	T/F	73
Joy	81	P/J	63
Sadness	62	Macro Avg	73
Surprise	58		
Neutral	62		
Macro Avg	60		

Table 2: Performance of auxiliary models used to generate 'silver labels' for emotion and personality

gories and Dreaddit only has labels for stress, we followed the approach described in Turcan et al. (2021) to derive emotion and personality labels for the two datasets. To this end, we first separately trained RoBERTa models on the GoEmotions and Kaggle MBTI datasets and use them to generate 'silver labels' for emotion and personality. The performance of these models is presented in Table 2.

We then trained the model in a multi-task setup on two tasks (mental health detection and emotion recognition or personality detection) or on all three tasks. In each task fusion model, the loss is the weighted sum of the loss from MHC part and secondary task part, where the weights are tunable $L = W_{MHC} \times L_{MHC} + (1 - W_{MHC}) \times L_{SEC}$

Separate binary classification models were constructed for each of four self-reported diagnosed mental health conditions (MHC) from the SMHD dataset (ADHS, anxiety, depression, bipolar) and stress from the Dreaddit dataset. For each MHC we constructed an emotion-infused model (MentalRoBERTa + Emotion), a personality-infused model (MentalRoBERTa + Personality), and a 'full-infusion' model (MentalRoBERTa + Emotion + Personality).

5 Experimental Setup

5.1 Baseline

We compared our models against a fine-tuned MentalRoBERTa model. We used the pretrained 'MentalRoBERTa-base' models from the Huggingface Transformers library (Wolf et al., 2019). The models consist of 12 Transformer layers with hidden size 768 and 12 attention heads. We run experiments with (1) a linear fully-connected layer for classification as well as with (2) an intermediate bidirectional LSTM layer with 256 hidden units. The following hyperparameters are used for fine-tuning: a fixed learning rate of 2×10^{-5} is applied and $L2$ regularization of 1×10^{-6}. All models were trained for 8 epochs, with batch size of 4 and maximum sequence length of 512 and dropout of 0.2. We report the results from the best performing models.

5.2 Training details

We trained all the models using BinaryCrossEntropy loss and Adam optimizer (adamw). We set the learning rate as 2e-5 and weight decay of 1e-5. We train the different models with different batch sizes. The BiLSTM network component of the feature fusion model had a batch size of 128 and for training all the other models we set a batch size of 32. We trained that component model for 200 epochs and all the other models for 8 epochs and saved the best preforming models on validation set. We evaluated these models on the test set and report the performance in terms of macro-F1 scores.

We selected the hyperparameters based on the the macro F1 score obtained on the the development set. We used grid search for getting the optimal values for the following: (1) for task fusion models: loss weights for primary and secondary tasks (0.5,0.5), (0.6, 0.4), (0.7,0.3) with the best f1 scores attained at equal weights for both tasks; (2) for the feature fusion model: hidden size 128, 256, 512, 1024, number of LSTM layers 1,2,3,4, dropout 0.2,0.4, we found the best performance with hidden size of 512, 3 layers and 0.2 dropout.

6 Results and discussion

Table 1 provides a concise overview of the performance in detecting five mental disorders (ADHD, anxiety, bipolar disorder, depression, and stress) for three fusion strategies (feature-level fusion, model fusion, and task fusion) in comparison to the baseline MentalRoBERTa model. In general, it is shown that our fusion-models outperform the MentalRoBERTa baseline model for three of the five mental health conditions (ADHD, anxiety, bipolar disorder), and performed similarly to the baseline model for depression and stress. For the the ADHD condition the best performing model, the 'Task Fusion - emotion' model, achieved an improvement of 4% F1 over the MentalRoBERTa baseline model. For anxiety and bipolar the best performance was achieved by the 'Task Fusion - personality model', an improvement over the baseline of 2% F1. Overall, these results indicate that task fusion is the most effective fusion strategy for detecting these three mental health conditions. Task fusion models were able to learn the features for the auxiliary tasks (emotion classification and personality detection) and thereby improve the performance of the primary task (mental health detection) for three conditions. The results also suggest that both emotions and personality are important in the detection of specific mental health disorders: We observed that detection of ADHD benefited most from infusion of emotion information, whereas detection of anxiety and bipolar disorders benefited most from infusion of personality information. The finding that fusion model performed similarly to MentalRoBERTa baseline model for stress is consistent with the findings reported in Turcan et al (2021): Their emotion fusion models constructed for the task of binary stress prediction achieved comparable performances to a fine-tuned BERT baseline model (F1 BERT = 78.88, F1 Emotion fusion model with Ekman GoEmotions relabeling = 80.24). The F1 score of our baseline MentalRoBERTa model was 3.3% higher than that of their baseline BERT model. For stress and depression, the best performance was obtained with the feature-level fusion approach, which yielded slight improvements over the MentalRoBERTa baseline. At the same time, we observed that infusing only information from the most informative source was more effective than full infusion, i.e. emotion and personality. A possible reason for this finding is noise or erroneous hidden features generated by the the auxiliary models in the case of model fusion (see Zhang et al., 2023; Pan and Yang, 2010, for discussion). A potential reason for lower performance of the full infusion models in the task learning approach is competition among the auxiliary tasks with regard to providing evidence for the relevance of particular features (see Ruder, 2017, for a discussion of 'attention focusing' in multi task learning). We intend to explore these issues in future research.

Building upon the approach described in Turcan et al (2021), we go a step further to probe our full task fusion models and discover the exact nature of the information it learned to use, i.e. how the six basic emotion categories (anger, disgust, fear, joy, sadness, and surprise) and four personality dimensions (Extraversion/Introversion (E/I), Sensing/Intuition (S/N), Thinking/Feeling (T/F) and Judgment/Perception (J/P)) guided the prediction of mental health status. To this end, we calculated Pearson correlation coefficients between the predicted probabilities for each of the five mental health conditions and the probabilities for the four personality and six emotion categories. Table 4 presents an overview of the results of this analysis. A visualization of the results can be found in Figure 2 in the appendix. The results revealed that the full task fusion model learned moderate to strong correlations between specific mental health statuses and specific emotion and personality categories: More specifically, the ADHD condition was strongly associated with sadness and disgust and moderately associated with anger and anxiety, whereas it was strongly negatively correlated with joy. Anxiety was strongly linked to joy and moderately associated with sadness, while being strongly negatively correlated with disgust. Bipolar disorder is characterized by strong negative associations with fear and disgust, with tendencies towards anger and sadness. Depression was strongly linked to the negative emotions of fear, anger, disgust and sad-

Model	ADHD	Anxiety	Bipolar	Depression	Stress
MentalRoBERTa	64.28	71.50	71.83	71.34	82.22
Feature-level Fusion	64.24	71.09	71.36	**71.88**	**82.59**
Model Fusion - emotion	65.75	70.59	71.14	71.43	80.80
Model Fusion - personality	65.12	71.42	71.58	70.44	81.08
Model Fusion - emotion & personality	64.04	71.57	69.68	68.91	81.07
Task Fusion - emotion	**68.02**	72.32	71.49	70.18	81.01
Task Fusion - personality	66.99	**73.40**	**73.23**	68.33	80.19
Task Fusion - emotion & personality	65.35	72.36	72.14	71.42	82.03

Table 3: Results of information-fusion models in comparison to baseline models. F1 scores averaged over two runs

	Emotion							Personality			
MHC	Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral	Extrovert	Intuitive	Thinker	Judging
ADHD	0.35	0.53	0.36	-0.73	0.68	0.27	-0.96	0.19	-0.28	-0.97	0.86
Anxiety	0.22	-0.74	-0.01	0.88	0.50	-0.21	-0.91	-0.80	0.75	-1.00	-0.90
Bipolar	0.14	-0.60	-0.75	-0.19	-.14	-0.77	-0.98	0.41	0.85	-1.00	-0.90
Depression	0.49	0.69	0.65	0.98	0.74	-0.11	-0.97	0.62	-0.77	-1.00	-0.92
Stress	0.12	0.05	0.34	-0.35	0.24	0.02	-0.30	-0.04	0.00	-0.17	-0.11

Table 4: Pearson correlations between predicted values on the primary task (mental health prediction) and each category of the secondary tasks (emotion prediction and personality recognition)**Note**: Following Cohen (1988), we consider correlation coefficients with absolute values greater than 0.3 to be 'moderate' and greater than 0.5 to be 'strong'.

ness. In addition - like anxiety - it was positively related to with joy, which is somewhat unexpected. Stress exhibited the weakest correlations to emotional categories with moderate positive correlations with fear and negative ones with joy being the most salient. We note that the weaker correlations between stress and the emotional categories can explain the more modest gain in predictive accuracy of the fusion models compared to the fine-tuned transformer model in both the present study and in Turcan et al. (2021).

Turning to personality, the task fusion model learned that all mental health conditions are associated with the MBTI-T dimension, such that individuals with a preference for relying on emotions in decision making are more likely to have an MHC diagnosis. Bipolar depression, ADHD and anxiety were also associated with the MBTI-J dimension, such that individuals that are less open to new information are more likely to exhibit any of these MHCs. Anxiety and bipolar disorder were correlated with the MBTI-E dimension, such that these conditions were more likely for individuals with a preference for focusing on the future with an emphasis on patterns and possibilities. Anxiety was also strongly negatively correlated with the MBTI-N dimension, meaning that the condition was much more prevalent in introverted individuals, than in extraverted ones. At the same time, extraversion was associated with both depression and to a lesser extent with bipolar disorder.

In line with results from experimental and genome-wide association studies of mental health and personality (Adams et al., 2019; Nikolic et al., 2020), these results suggest that personality dimensions are important in understanding vulnerability to mental health disorders.

7 Conclusion

In this work, we presented the first comprehensive experimental evaluation of current deep learning-based fusion strategies (feature-level fusion, model fusion, task fusion) for the detection of mental disorders. We go beyond previous work by applying these approaches to five mental health conditions. The results of our experiments showed that the task fusion strategy is most promising for the detection of three of the five conditions (ADHD, anxiety, and bipolar disorder), while feature-level fusion is most advantageous for the detection of psychological distress and depression. We demonstrated that the prediction of mental health from textual data benefits from the infusion of two information sources related to mental disorders, i.e. emotion and personality. Furthermore, we show that information fusion models can improve the classification accuracy of strong transformer-based prediction models while enhancing their explainability.

In this paper, we focused on developing binary classifiers that aim to distinguish between individuals with a particular mental illness and control users. In future work, we intend to addresses the more complex problem of distinguishing between multiple mental health conditions, which is essential if we are to uncover the subtle differences among the statistical patterns of language use associated with particular disorders. We further intend to employ our approach to longitudinal data to gain valuable insights into the evolution of symptoms over time and extend it to languages beyond English, specifically German.

Limitations

We note that the datasets used in this work solely represent social media interactions from Reddit, which is known to have a demographic bias toward young, white, American males[3]. Furthermore, systematic, spurious differences between diagnosed and control users can prevent trained models from generalizing to other data. Future research on other social media and datasets is needed to determine to what extent the presented findings are generalizable to broader populations.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Mark J Adams, David M Howard, Michelle Luciano, Toni-Kim Clarke, Gail Davies, W David Hill, Daniel Smith, Ian J Deary, David J Porteous, Andrew M McIntosh, et al. 2019. Stratifying depression by neuroticism: revisiting a diagnostic tradition using gwas data. *bioRxiv*, page 547828.

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

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Rafael A. Calvo, David N. Milne, M. Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

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*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jacob Cohen. 1988. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates. *Hillsdale, NJ*, pages 20–26.

Angelo Compare, Cristina Zarbo, Edo Shonin, William Van Gordon, and Chiara Marconi. 2014. Emotional regulation and depression: A potential mediator between heart and mind. *Cardiovascular Psychiatry and Neurology*, 2014.

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*, pages 4040–4054, Online. Association for Computational Linguistics.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99 3:550–3.

Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Muskan Garg. 2023. Mental health analysis in social media posts: A survey. *Archives of Computational Methods in Engineering*, pages 1–24.

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.

Jutta Joormann and Ian H Gotlib. 2010. Emotion regulation in depression: Relation to cognitive inhibition. *Cognition and Emotion*, 24(2):281–298.

[3] https://social.techjunkie.com/demographics-reddit

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*, pages 182–194, Dublin, Ireland. Association for Computational Linguistics.

Charles Li, Monte Hancock, Ben Bowles, Olivia Hancock, Lesley Perg, Payton Brown, Asher Burrell, Gianella Frank, Frankie Stiers, Shana Marshall, Gale Mercado, Alexis-Walid Ahmed, Phillip Beckelheimer, Samuel Williamson, and Rodney Wade. 2018. Feature extraction from social media posts for psychometric typing of participants. In *Augmented Cognition: Intelligent Technologies*, pages 267–286, Cham. Springer International Publishing.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of eRisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer.

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, Baltimore, Maryland. Association for Computational Linguistics.

Isabel Briggs Meyers, Mary H McCaulley, and Allen L Hammer. 1990. *Introduction to Type: A Description of the Theory and Applications of the Myers-Briggs Type Indicator*. Consulting Psychologists Press.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.

Sanja Nikolic, Ivana Perunicic Mladenovic, Olivera Vukovic, Jasmina Barišić, Dragan Švrakić, and Srdjan Milovanović. 2020. Individual and gender differences in personality influence the diagnosis of major depressive disorder. *Psychiatria Danubina*, 32(1):97–104.

Susan Nolen-Hoeksema, Blair E. Wisco, and Sonja Lyubomirsky. 2008. Rethinking rumination. *Perspectives on Psychological Science*, 3(5):400–424. PMID: 26158958.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

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.

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.

A. Perrin. *Social Media Usage: 2005-2015: 65% of Adults Now Use Social Networking Sites–a Nearly Tenfold Jump in the Past Decade*. Pew Research Trust.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 21–30.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7685–7697, Online. Association for Computational Linguistics.

Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. 2019. # suicidal-a multipronged approach to identify and explore suicidal ideation in twitter. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 941–950.

Hoyun Song, Jinseon You, Jin-Woo Chung, and Jong C Park. 2018. Feature attention network: Interpretable depression detection from social media. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

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, Hong Kong. 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, Online. Association for Computational Linguistics.

Ana-Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, 124:480–494.

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)*, pages 5276–5290, Dublin, Ireland. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

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.

Katherine S Young, Christina F Sandman, and Michelle G Craske. 2019. Positive and negative emotion regulation in adolescence: links to anxiety and depression. *Brain Sciences*, 9(4):76.

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

Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. 2023. Emotion fusion for mental illness detection from social media: A survey. *Information Fusion*, 92:231–246.

Binggui Zhou, Guanghua Yang, Zheng Shi, and Shaodan Ma. 2022. Natural language processing for smart healthcare. *IEEE Reviews in Biomedical Engineering*.

A Appendix

Secondary	Dataset	Number	Avg. len	SD	Total	Avg. len	SD	Total
attribute		of posts	(words)	(words)	(words)	(chars)	(chars)	(chars)
Personality	Kaggle MBTI	8675	1309.11	327.11	11356602	6795.6	1676.95	58951828
ENFJ		190	1372	326	260724	7062	1651	1341841
ENFP		675	1344	315	907091	6902	1601	4658783
ENTJ		231	1299	304	300037	6809	1550	1572947
ENTP		685	1290	294	883676	6717	1529	4601132
ESFJ		42	1379	373	57905	7069	1908	296884
ESFP		48	1099	405	52753	5656	2084	271501
ESTJ		39	1312	315	51178	6740	1564	262870
ESTP		89	1242	337	110567	6374	1719	567266
INFJ		1470	1363	316	2003249	7061	1619	10379463
INFP		1832	1328	325	2432535	6858	1658	12564597
INTJ		1091	1274	334	1389940	6693	1732	7301709
INTP		1304	1281	321	1669835	6713	1667	8753488
ISFJ		166	1328	377	220413	6818	1922	1131708
ISFP		271	1217	360	329703	6269	1833	1698980
ISTJ		205	1297	348	265895	6692	1746	1371951
ISTP		337	1250	341	421101	6459	1744	2176708
Emotion	GoEmotion	52501	13.84	6.97	726668	67.69	36.60	3553890
Anger		7022	14.5	6.94	101980	71.8	36.7	504334
Disgust		1013	14.2	6.84	14388	71.1	35.8	72008
Fear		929	14.5	7.03	13507	71.6	36.3	66554
Joy		21733	13.6	6.91	296623	66.2	35.9	1438087
Neutral		17772	13.5	7.06	239784	66.3	37.5	1178538
Sadness		4032	15.0	6.80	60386	73.0	35.4	294369

Table 5: Count of posts, tokens and characters along with average post length of datasets used for secondary tasks

Model	**ADHD**		**Anxiety**		**Bipolar**		**Depression**		**Stress**	
Runs	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**1**	**2**
MentalRoBERTa	64.48	64.08	71.72	71.28	72.04	71.62	72.01	70.67	81.98	82.46
Feature-level Fusion	63.96	64.52	72.34	69.84	71.54	71.18	71.89	71.87	82.83	82.35
Model Fusion - emotion	65.64	65.86	70.40	70.78	71.63	70.66	71.73	71.13	81.08	80.52
Model Fusion - personality	65.03	65.21	71.33	71.51	71.99	71.17	70.63	70.25	81.14	81.02
Model Fusion - emotion & personality	64.16	63.92	71.66	71.48	69.97	69.39	69.33	68.49	81.52	80.62
Task Fusion - emotion	68.55	67.49	71.98	72.66	71.89	71.09	70.61	69.75	81.16	80.86
Task Fusion - personality	66.85	67.13	73.26	73.54	73.30	73.16	68.63	68.04	80.49	79.89
Task Fusion - emotion & personality	65.03	65.67	72.27	72.45	72.37	71.91	71.60	71.24	82.52	81.54

Table 6: Results of information-fusion models in comparison to baseline models

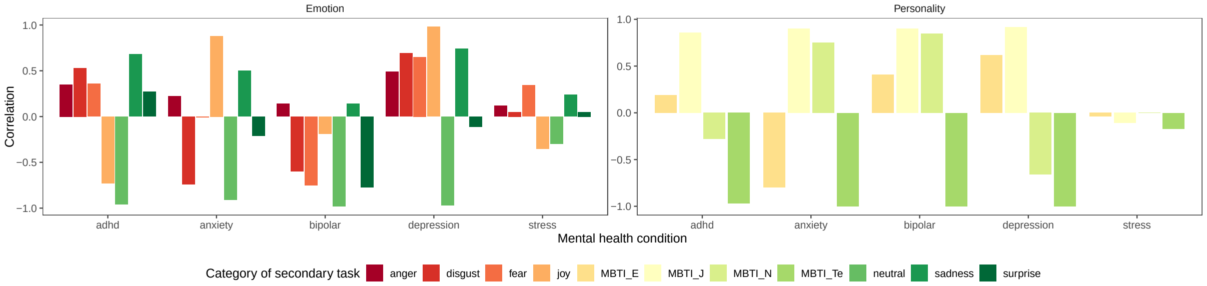

Figure 2: Pearson correlations between predicted values on the primary task (mental health prediction) and each category of the secondary tasks (emotion prediction and personality recognition)

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
Left blank.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.