Explainable Depression Symptom Detection in Social Media

Eliseo Bao Souto
Information Retrieval Lab, CITIC
Universidade da Coruña
A Coruña, Galicia, Spain
eliseo.bao@udc.es

Anxo Pérez
Information Retrieval Lab, CITIC
Universidade da Coruña
A Coruña, Galicia, Spain
anxo.pvila@udc.es

Javier Parapar
Information Retrieval Lab, CITIC
Universidade da Coruña
A Coruña, Galicia, Spain
javier.parapar@udc.es

ABSTRACT

Users of social platforms often perceive these sites as supportive spaces to post about their mental health issues. Those conversations contain important traces about individuals' health risks. Recently, researchers have exploited this online information to construct mental health detection models, which aim to identify users at risk on platforms like Twitter, Reddit or Facebook. Most of these models are centred on achieving good classification results, ignoring the explainability and interpretability of the decisions. Recent research has pointed out the importance of using clinical markers, such as the use of symptoms, to improve trust in the computational models by health professionals. In this paper, we propose using transformer-based architectures to detect and explain the appearance of depressive symptom markers in the users' writings. We present two approaches: *i*) train a model to classify, and another one to explain the classifier's decision separately and *ii*) unify the two tasks simultaneously using a single model. Additionally, for this latter manner, we also investigated the performance of recent conversational LLMs when using in-context learning. Our natural language explanations enable clinicians to interpret the models' decisions based on validated symptoms, enhancing trust in the automated process. We evaluate our approach using recent symptom-based datasets, employing both offline and expert-in-the-loop metrics to assess the quality of the explanations generated by our models. The experimental results show that it is possible to achieve good classification results while generating interpretable symptom-based explanations.

KEYWORDS

Explainability, Interpretability, Depression detection, Social Media

1 INTRODUCTION

Mental health is a crucial component of overall health and well-being. The most recent figures estimate mental disorders prevalence in adults over 20% [48]. The World Health Organization (WHO) estimates that approximately 332 million people globally are affected by depression [50]. Early intervention in mental disorders is crucial in mitigating their impact, especially for young individuals [54]. However, due to the stigma associated with depression, more than 60% of individuals with the condition do not seek professional support [21]. To address this problem, computational researchers have a growing interest in assisting in the early detection and diagnosis of depression, thereby mitigating its societal impact [14].

In this scenario, researchers have found that the writings posted by individuals on social media platforms are valuable evidence for looking for early signs of depression. Individuals experiencing depression find comfort in expressing their thoughts and emotions on these platforms, motivated by factors such as privacy or anonymity [9]. Consequently, social media provides a complementary opportunity to access valuable information about individuals' state of mind beyond traditional professional therapy. The combined use of computational linguistic techniques and the vast amount of data from social networking has led to significant advancements in detecting signs of depression [59]. The field's critical nature motivated much effort in creating curated experimental benchmarks [52, 79], which allowed the development and evaluation of many new predictive models. Traditional efforts used engineered features such as word counts, posting activity or emotion levels [8, 60, 75]. Due to the rise of transformer-based language models, many researchers used these deep learning models as classifiers to identify users at risk of depression and similar disorders in online environments [2, 5, 7].

However, researchers in this field do not plan to replace mental health professionals but rather offer support to their work. Licensed clinicians play a crucial role in validating the predictions made by computational models and taking appropriate actions with individuals when necessary. Computational models may only be used carefully to extend those professionals' reach and facilitate their workflow. The existing models, however, present many limitations for achieving that goal [70]. One significant barrier is their limited ability to explain their predictions resulting in scepticism by the professionals. Reliable interpretation of models' decisions is mandatory for professionals to understand and trust these models and use them in their daily work [25]. One way to pursue that is by designing new models incorporating trustworthy and reliable explanations [19]. Following that path, recent research has explored the utilization of symptoms collected from validated clinical questionnaires. Most of these proposals, in the field of depression, used the markers from the Beck Depression Inventory-II (BDI-II) [4] or the 9-Question Patient Health Questionnaire (PHQ-9) [30], which encompass a range of depressive symptoms such as irritability, pessimism or sleep problems. The utilization of such symptom markers has been shown to improve the explainability, generalization and overall performance of depression detection models [45, 55, 76, 77].

With that motivation, we aim to develop models that categorise whether or not social media posts exhibit validated depressive symptoms. Accurately detecting symptom information along a user's vast amount of writing is the first step in developing explainable depression detection models. We go beyond classifying posts for depressive symptoms by providing a explanation for the decisions.

The terms 'interpretability' and 'explainability' can be challenging to define in our specific context, as there is some discrepancy in previous literature. Some authors use these terms interchangeably, referring to the ability to explain or present systems in a manner understandable to humans [17]. However, other authors consider them as distinct concepts. In this perspective, interpretability relates to the system's capacity to be understood by humans, while explainability encompasses being true to reality [39]. In our work, we adopt a simple definition: interpretability/explainability refers to the extent to which humans can understand the reasons behind a decision [40].

In this paper, we introduce a text-to-text pipeline designed to achieve two main objectives: classifying the relevance of social media publications to depressive symptoms and providing explanations for the classification decisions. To implement this pipeline, we explore the effectiveness of state-of-the-art transformer-based models. Figure 1 illustrates the two approaches considered: Part *a*) employs text-to-text models that perform classification and explanation simultaneously. Part *b*) corresponds with two-step approaches, utilizing separate models for classification and explanation. Additionally, we evaluate the capabilities of recent Large Language Models (LLMs) utilizing recent conversational models such as GPT-3.5 Turbo [49] and Vicuna-13B [12]. We utilize two datasets, PsySym [76] and BDI-Sen [53], for model training and evaluation. These datasets consist of short sentences from social media posts associated with depressive symptoms. As we will discuss later, these sentences can be used as explanations for a post to explain its relevance to a symptom. To assess the quality of the generated explanations, we employ a combination of offline and expert-in-the-loop metrics. The experimental results demonstrate the effectiveness of our methods. By prioritizing explainability, our approach bridges the gap between automated predictions and human understanding, facilitating more informed clinical decision-making[1].

Our study aims to address the following research questions: **RQ1** Can we train transformer-based models to accurately classify the presence of depressive symptoms in social media posts, while providing explanations for their decisions? **RQ2** How many training examples, i.e. hand-labelled posts by domain experts, are needed to generate good explanations? **RQ3** Does a unified model, designed for both classification and explanation, significantly differ in performance from using two separated models trained for each task? **RQ4** How do recent conversational LLMs perform in detecting and explaining the presence of depressive symptoms in social media content?

2 RELATED WORK

Social media data has gained increasing attention for developing depression detection models in recent years [13, 22, 63]. The evaluation results of those methods show promising accuracy numbers [2, 5, 7, 59, 72]. Such platforms offer an opportunity to identify disorders at an early stage [62], a step that is crucial to reduce negative impacts and their associated costs [23, 54].

However, depression detection field still needs to overcome the cross-cutting problem of many machine learning applications: explainability. There are two common approaches when explaining models' decisions. On the one hand, we may use intrinsically explainable models, such as decision trees or linear models [20]. On the other hand, we can opt for model-agnostic, typically post hoc, explainability methods [58, 66], which can be applied to any supervised machine learning model, regardless of its architecture. Moreover, the explanations can be local or global. Local explanations focus on individual decisions, allowing users to understand why the model produces a particular decision [3, 57]. These local explanations, such as LIME (Local Interpretable Model-Agnostic Explanations) [57] and SHAP (Shapley Additive exPlanations) [36], have already been used in the social psychology field [35, 46]. Alternatively, global explanations consider the whole machine learning model behaviour and its predictions altogether [32, 37, 68]. Most traditional detection methods used engineered features (e.g., word counts or sentiment analysis) and based their global explanations on feature importance metrics [8, 59]. These measures quantify the impact of each feature on the predictions, providing insights into the relative importance of each one.

With the rise of transformer-based models [69], researchers have leveraged these models for sentence classification tasks to detect signs of depression in social media users, achieving impressive results in various experimental benchmarks [52, 79]. The encoder-decoder architecture of transformers effectively captures contextual information from input sequences and generates corresponding output sequences. However, maintaining the contextual integrity of each token in extensive texts presents a challenge. The attention mechanism tackles this by allocating weights to tokens relative to all others, allowing the model to focus on relevant parts of the input. Prior works have explored the attention mechanism's use as an explainability tool [1, 42, 64]. Despite its utility, attention mechanisms also have limitations, particularly in interpreting the weights assigned to input features. There is an ongoing discussion in the community about the utility of attention weights for interpretability purposes, with some research advocating its benefits [67, 73], while others point out its constraints [27, 61].

More recently, an alternative strategy for building interpretable models has emerged: the use of generative natural language explanations. This technique has several benefits: *i*) they are readily comprehensible to end users, *ii*) human annotators can more easily work with natural language, simplifying data collection, and *iii*) it may be feasible to extract natural language explanations from large datasets of domain-expert data, a promising prospect for future research [11]. In this paper, our goal is to cross that bridge following prior efforts in training text-to-text models in order to produce natural language explanations for the depression detection problem. To further enhance the interpretability for health professionals, we ground all our explanations in clinical symptoms following the BDI-II questionnaire. By formulating explanations in the format of human-readable text, our method makes the model's rationale behind its predictions transparent to clinicians. This level of interpretability is of critical importance to making informed decisions for depression detection [70], promoting both the efficacy and the trustworthiness of our models.

3 OUR PROPOSAL

Here we describe our proposal for producing explainable decisions for depression symptom detection in users' posts. The task is defined as follows: given a user post, the model classifies it as either indicative (positive) or non-indicative (negative) for depression

[1] Our implementation is available at: *coming soon.*

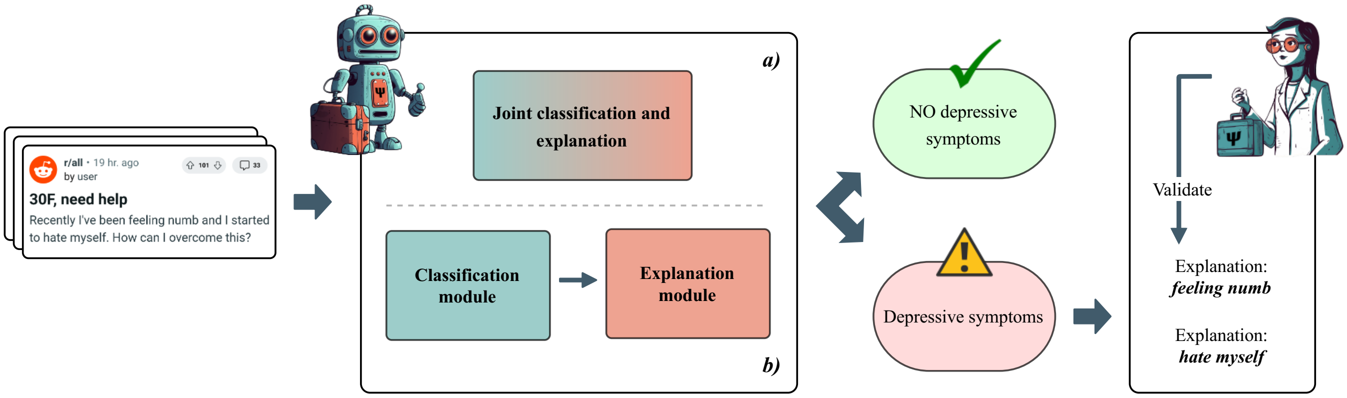

Figure 1: Overall pipeline of our proposals for the classification and generation of natural language explanations for the presence of depression symptom information in social media posts.

symptoms. A positive post implies that the post shows information for the user on one or more of the 21 symptoms described in the BDI-II questionnaire. Furthermore, we aim to explain the model's decision by generating natural language explanations, explaining the reasons for a positive classification. We undertake this task in two different manners: either as a single output (i.e. train together classification plus explanation), or in two separate steps (i.e. classification followed by post hoc explanation). These two scenarios are illustrated in Figure 1. Part *a)* is representative of the single-step approach, explained in Subsection 3.2.1, while part *b)* corresponds to the two-step approach, detailed in Subsection 3.2.2. If the pipeline identifies a post as having potential symptom risk, it provides corresponding explanations. Looking at the Figure, we can see that the explanations for the decision are *'feeling numb'* and *'hate myself'*. Health professionals can then validate these explanations.

3.1 Explanations

Natural language explanation generation methods can be either extractive or abstractive [10]. In the *extractive* case the model is asked to point out the parts of the original text that led to the decision. On the other hand, in the case of *abstractive explanations*, the model is trained to explain the reasons for the decision in a free format. We trained our methods for producing extractive explanations that include token spans from the input text. We decided to do so for data availability and because this approach allows the calculation of overlap statistics concerning the ground truth, thus providing a quantitative measure of explanation quality.

Another decision we adopted is to explain only the positive decisions. The models do not explain when they decide there is no trace of depression symptom markers. We decided to do so for two reasons. First, the explanation of negative cases is considered useless, as clinicians are concerned with positive cases. Secondly, for extractive explanations, it should be some constant phrase like *"there is no evidence in the post about any depression symptom"*.

3.2 Finetuning Text-to-text Models

Raffel et al. [56] proposed the idea of transforming text-related tasks to follow the sequence-to-sequence (seq2seq) format [65], which is referred to as the text-to-text framework. This diverges from conventional methods like BERT-based approaches, which train models to yield a probability distribution over predefined output classes [15]. In contrast, text-to-text models are trained to generate textual sequences. Consequently, these models may generate an unexpected output, which is considered a prediction error.

In this context, the input text sequence, which we will call $\tilde{x}$, is defined as "`<task_prefix>: <input_text>`", where `<task_prefix>` identifies the task to be performed by the model. On the other hand, the output text sequence $\tilde{y}$ will be of the form "`<target>`", where `<target>` corresponds to the desired output. If we take a sentiment analysis task as example, the sequence would be: $\tilde{x}$ = "`sentiment analysis: terrible product`" and $\tilde{y}$ = "`negative`". The Text-to-Text Transfer Transformer (T5) model is one of the most popular models under this paradigm [56], and it achieved state-of-the-art results in many NLP tasks [71]. While there exist other text-to-text models, such as BART [33], prior research indicates that it performs worse than T5 in terms of explanation generation [18]. These models are pre-trained, providing us with a robust base that we can further finetune under either the single-step or two-steps approaches discussed below for generating explanations.

3.2.1 Single-step. Narang et al. [43] explored how to teach text-to-text models to produce both classification and explanation. For that, they introduced an extension of T5 called WT5 ("Why T5?"). To use this method, the keyword "`explain`" is simply added to the input $\tilde{x}$, preceded by "`<task_prefix>`". The target $\tilde{y}$ is appended with the phrase "`explanation: <explanation>`". Using this template, we adapted it to our task resulting in the new input/output format:

$$\tilde{x} = \text{“}\texttt{explain symptom post: <post\_text>}\text{”}$$

$$\tilde{y} = \text{“}\texttt{<target> [explanation: <explanation}_1\texttt{>] ...}$$
$$\texttt{... [explanation: <explanation}_N\texttt{>]}\text{”}$$

The hard brackets denote potentially multiple explanation sentences. An example of input could be "`explain symptom post: I absolutely hate myself (...) And I hate how I feel the need to burden other people with this. I am so whiny, so disgustingly insensitive (...)`" [2]. Its corresponding target output would be

[2]Full-length post is paraphrased and redacted for clarity and space reasons.

"`positive explanation: I absolutely hate myself explanation: I am so whiny, so disgustingly insensitive`". Thus, in one step, users' posts can be both classified and extractive explained according to whether they indicate depressive symptoms or not. Part *a)* in Figure 1 illustrates this approach, which we use to build two of our systems: **WT5** and **WBART**. Our WT5 system replicates the research of Narang et al. that we just briefly discussed, while our WBART system extends the same idea but to the BART model [18].

3.2.2 Two-steps pipeline. Alternatively, rather than training a single model to execute both classification and explanation, we propose a two-steps pipeline. In this approach, one model is initially employed for classification, followed by the use of a different model for explanation. We develop three of our systems using this methodology: **T5 + T5**, **BERT + T5**, and **MBERT + T5**. This process is depicted in part *b)* of Figure 1. Next, we define the stages of classification and explanation as they occur in this pipeline.

Classification. In alignment with recent studies, we conduct experiments with various types of pre-trained language models to investigate their capabilities in terms of classification [45, 76, 77]. These models constitute the classification module and their sole responsibility is to determine whether a post is indicative of a depressive symptom. Firstly, we fine-tune a **T5** model with the aim of generating classification labels in a text format [56]. Secondly, we explore the utility of BERT-based architectures, utilizing the pre-trained **BERT** base uncased model [15]. Additionally, we finetuned MentalBERT (**MBERT**), a model purposely pre-trained for mental health applications using data compiled from various subreddits related to mental disorders [28].

Explanation. For the explanation module, we finetuned a **T5** model responsible for explaining the evidence for those posts deemed positive by the classification models. As previously commented, we only trained this model to generate explanations for positive posts. Negative posts are discarded and do not enter in the second stage of the two-steps pipeline.

3.3 In-context Learning with LLMs

In-context learning offers a valuable alternative to finetuning LLMs without directly modifying their parameters, instead guiding them to specific behaviours within a given context [6]. We employed this technique to instruct recent conversational LLMs to perform following our proposed single-step approach (*classify + explain*). More specifically, we used **Vicuna-13B** [12] and **GPT-3.5 Turbo** [49]. Using specific guidelines, we instructed the models to act as expert annotators responsible for detecting signs of depressive symptoms in user posts [3]. Additionally, they must justify their decisions by quoting relevant excerpts from the original text. This approach aims to replicate the methodology discussed earlier, which outlined the process of finetuning text-to-text models for explanation.

4 EXPERIMENTS

In this section, we describe the experiments conducted to evaluate the effectiveness of our approaches for symptom detection and explanation over social media content. We present the datasets utilized, followed by a description of the experimental configurations and training details. Subsequently, we explore the metrics employed to measure the performance of our methods, encompassing both offline evaluations and those involving human experts.

Table 1: Statistics of the positive instances of our explainability dataset. Average length of posts/explanations are expressed as tokens.

	BDI-Sen	PsySym
Number of posts	357	752
Number of explanations	546	764
Avg. number of expls. per post	1.53	1.02
Avg. length of post (in tokens)	127	515
Avg. length of explanation	13.76	13.44

Table 2: Settings considered in our experiments by combining the BDI-Sen and PsySym datasets.

Setting	Training			Test		
	Dataset	⊕/⊖	Avg. expls.	Dataset	⊕/⊖	Avg. expls.
B-B	BDI-Sen	285/285	1.48	BDI-Sen	72/359	1.72
B-P	BDI-Sen	285/285	1.48	PsySym	151/753	1.02
P-P	PsySym	601/601	1.01	PsySym	151/753	1.02
P-B	PsySym	601/601	1.01	BDI-Sen	72/359	1.72
M-M	Mix	886/886	1.16	Mix	223/1112	1.25

4.1 Datasets

In this work, we considered the symptoms of the BDI-II clinical questionnaire [4]. The BDI-II covers 21 recognized symptoms, such as pessimism, sleep problems or self-dislike. We have decided to align with the BDI-II symptoms by its widespread adoption in clinical practice [26] and its presence in prior literature concerning depression detection on the Internet [13, 44, 52, 55, 77]. We obtained the symptom evidence sentences from the BDI-Sen [53] and PsySym [76] resources. Both datasets consist of symptom-annotated sentences for depression sourced from the Reddit platform. BDI-Sen comprises relevant sentences related to the 21 symptoms covered in the BDI-II, while PsySym includes 14 main symptoms covered within the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [47]. Table 1 presents the statistics of the positive posts from the datasets. In addition to positive posts, we also included control instances in the experiments. We selected random control sentences from the datasets, totalling 1998 control posts. In Section 7, we comment on the anonymization and ethical use of the data.

The datasets comprise two different data sources. Thus, for our experiments, we can define five different settings for train and test sets by combining both resources. Initially, for both BDI-Sen and PsySym, we split their data 80-20 into training and test sets. Then, we fill these splits with control samples from PsySym with ratios 1:1

[3]Our guidelines are available at: *coming soon.*

for training set and ≈ 1:5 for the test set[4]. Table 2 illustrates the five different settings we consider, where we test each dataset on itself, and also using as test the other one to evaluate if it can generalize correctly. In the last setting (*M-M*), we merge both datasets.

4.2 Experimental Settings

4.2.1 Single-step. We finetuned two different text-to-text models for the single-step variant (**WT5** and **WBART**), as well as the two previously described conversational LLMs (**Vicuna-13B** and **GPT-3.5**). For WT5, we adhere to the training procedure proposed in the WT5 framework by Narang et al. [43]. We conducted 40 epochs of fine-tuning to adapt the pre-trained `t5-large` configuration for our *classify + explain* task, with the AdaFactor optimizer and a constant learning rate of 0.001. Sequence lengths were set to 2048 and 512 tokens for the inputs and the outputs, respectively. Regarding WBART, we also pre-trained the `bart-large` model. In this case, we used the default values for the learning rate and the optimizer hyperparameters. The maximum lengths of the source and target sequences were halved compared to WT5. This is due to computational constraints. We used NVIDIA A100-SXM4 80GB GPUs to train and test these models. For Vicuna-13B, we hosted an instance on a NVIDIA RTX A6000 48GB. We consumed GPT-3.5 via API calls. The latter two models were instructed using the in-context learning strategy described in Subsection 3.3 with 30 positive and 30 control samples randomly chosen.

4.2.2 Two-steps. First, we used two BERT-based models for text classification: BERT base, and MBERT base, following the existing implementations from the HuggingFace library without any additional hyperparameter tuning. The learning rate was $2e^{-5}$ during 20 epochs and a batch size of 32. We also finetuned a T5 on its `t5-large` configuration for text classification by generating labels in textual form. For explanations, we used another `t5-large` to explain the symptoms presence. The parameters used to finetune the T5 models were the same as those used for WT5, with same sequence lengths. As a result, we constructed three different variants within this approach: **T5 + T5**, **BERT + T5**, and **MBERT + T5**.

4.3 Evaluation

4.3.1 Classification. To evaluate the classification performance of our systems, we consider micro-averaged F1 due to the unbalanced nature of the datasets. By using micro-F1, we account for the importance of each instance, giving more weight to the minority class (positive). Additionally, we include the number of true positives (TPs) to gain insight into the actual number of correctly identified positive instances. This information is particularly significant since our systems only generate explanations for the positive samples.

4.3.2 Explanation. LLMs can generate text unsupported by the input and inaccurate, a phenomenon called "hallucination" [29]. In the clinical domain, where the validity and truthfulness of the explanations extracted are crucial [70], the generated texts' reliability is essential. To ensure the trustworthiness of the generated explanations, we complement classical offline metrics with *expert-in-the-loop* evaluation performed with three domain experts.

[4]We include more control sentences in the test set to replicate real scenarios, where there are many more negatives examples than positive ones.

Offline. We employed three offline metrics: ROUGE-L-F1 [34], Corpus BLEU [51], and Token F1 [16], each offering different evaluation perspectives. All these metrics compare the generated explanations with the text reference considered as golden truth. ROUGE emphasizes content overlap, ensuring that the generated output captures essential information from the references. On the other hand, BLEU focuses on fluency and adequacy, assessing the linguistic quality and alignment with references. Since these metrics effectively measure the quality of the generated text by comparing it against reference hypotheses, ROUGE and BLEU can evaluate the quality of our explanations while still allowing some paraphrasing. This flexibility can be beneficial when the explanation text slightly differs from the ground truth yet conveys the same meaning. Additionally, Token F1 measures which input tokens labelled as an explanation in the ground truth are present in the generated one. Thus, Token F1 score is computed through a token-by-token analysis, identifying if the tokens generated are the same as the references.

Expert-in-the-loop. We also performed an online evaluation using three domain experts to assess the practical utility and clinical relevance of the generated explanations. The evaluators consisted of two psychologists and a speech therapist. To ensure the consistency of judgments, we organized training sessions with the evaluators during which we discussed the criteria for relevance and established a consensus on samples of both positive and negative explanations. An explanation was considered relevant if it successfully provided evidence for the presence of a depressive symptom in the post. We provided the experts with explanations generated by our WT5 model variant under the *M-M* setting (see Table 3), leading to a total of 209 explanations being evaluated. We chose this setting because it includes both datasets and selected WT5 since it was the best performing model in this scenario.

5 RESULTS AND DISCUSSION

In this section, we present the results of the experiments designed to address the four research questions posed in the introduction.

5.1 RQ1: Can we train transformers models to detect and explain depressive symptoms?

5.1.1 Symptoms classification. The results presented in Table 3 confirm that we may train transformer architectures to detect traces of depression symptoms in social media posts. Across the different settings, the lowest-performing model achieved 0.86 of classification accuracy, and our finetuned models outperformed the instructed conversational LLMs (Vicuna and GPT-3.5). We observe the best classification figures in the *P-P* setting (training and testing in PsySym dataset), with all those models achieving over 0.97 F1 values. Table 3 also provides insights into how the models generalise when tested on a different dataset (the *B-P* and *P-B* settings). We observe bad generalisation in the case of the *P-B* setting. For instance, the MBERT + T5 model achieved an F1 of 0.98 when trained and tested on the PsySym dataset (*P-P* setting). Meanwhile, its performance dropped to 0.90 when tested on the BDI-Sen dataset (*P-B* setting). However, in the *M-M* setting, the trained models obtained an F1 greater than 0.94.

Table 3: Results of our methods in different settings. All metrics considered are in the range 0-1, except for true positives, which are an integer. Higher values indicate better performance.

Setting	System	F1	TPs	ROUGE	BLEU	TF1
B-B	WT5	0.91	**68**	0.62	0.53	0.48
	WBART	0.88	**68**	0.60	0.53	0.48
	T5 + T5	0.93	67	0.69	0.55	0.54
	BERT + T5	**0.95**	58	**0.75**	**0.61**	**0.60**
	MBERT + T5	0.92	65	0.72	0.58	0.58
	Vicuna-13B	0.73	41	0.42	0.22	0.34
	GPT-3.5	0.82	62	0.55	0.42	0.45
B-P	WT5	0.87	125	0.31	0.22	0.10
	WBART	0.86	134	0.22	0.15	0.10
	T5 + T5	0.86	79	**0.35**	**0.23**	**0.12**
	BERT + T5	0.91	79	0.34	0.22	0.12
	MBERT + T5	**0.92**	144	0.29	0.18	0.10
	Vicuna-13B	0.78	140	0.14	0.06	0.11
	GPT-3.5	0.87	**151**	0.17	0.08	0.10
P-P	WT5	**0.98**	143	0.53	0.43	0.47
	WBART	**0.98**	149	**0.56**	**0.53**	**0.51**
	T5 + T5	**0.98**	145	0.45	0.34	0.38
	BERT + T5	0.97	150	0.47	0.36	0.38
	MBERT + T5	**0.98**	149	0.45	0.35	0.38
	Vicuna-13B	0.69	150	0.15	0.08	0.09
	GPT-3.5	0.94	**151**	0.21	0.11	0.11
P-B	WT5	0.89	35	0.61	0.61	0.46
	WBART	0.89	35	**0.71**	**0.68**	**0.51**
	T5 + T5	0.89	40	0.61	0.60	0.45
	BERT + T5	**0.91**	48	0.57	0.57	0.42
	MBERT + T5	0.90	42	0.61	0.58	0.45
	Vicuna-13B	0.62	**53**	0.36	0.19	0.23
	GPT-3.5	0.85	49	0.53	0.34	0.36
M-M	WT5	0.95	209	**0.57**	**0.54**	**0.50**
	WBART	0.94	**211**	0.53	0.47	0.48
	T5 + T5	0.95	206	0.54	0.51	0.46
	BERT + T5	0.95	**211**	0.54	0.51	0.46
	MBERT + T5	**0.96**	**211**	0.54	0.51	0.45
	Vicuna-13B	0.59	195	0.22	0.10	0.14
	GPT-3.5	0.91	200	0.28	0.17	0.19

To further analyse that generalisation problem, we illustrate the distribution of true and false predictions of the systems in Figure 2. As observed in the confusion matrices [5], the models have a high ratio of true positives, with the majority identifying over 90% of the positive instances. Furthermore, the ratio of false negatives is remarkably small. However, we can again observe significant deviation in these trends in the matrix for the *P-B* setting. Here we can see the reason for that problem: the PsySym dataset covers 14 symptoms of depression, while BDI-Sen includes all 21 symptoms of the BDI-II. In the *P-B* setting, not all symptoms of BDI-Sen have been seen during training, leading to a substantially higher false positive number. This finding highlights the potential generalisation issues these models may encounter regarding the data used for training. False negatives are a crucial metric for clinical applications. In clinical practice, false negatives carry great significance. Failing to identify an individual at risk has far more severe implications than falsely identifying a healthy person (false positives).

[5] Due to clarity and space constraints, not all systems have been shown.

5.1.2 Symptoms explanation. Regarding explanation quality, offline metrics and human evaluation results indicate promising results. Once again, the trained models outperform the most recent LLMs. The figures for explanation quality indicate that generating explanations is more challenging than the classification task. As previously stated (§3.1), it is crucial to bear in mind that we generate explanations only for positive cases. Hence, the explanation quality numbers have to be jointly considered with the number of true positives (refer to the TPs column in Table 3). When considering the offline metrics (ROUGE, BLEU, and TF1), the fairest comparison can be made in the *M-M* setting, since it is the largest setting and provides a similar number of TPs cases for all models. Here, WT5 emerges as the top performer, achieving a ROUGE-L F1-score of 0.57. The two-step models also perform closely to the single-step models (WT5 and WBART) in nearly all settings. In the *B-B* setting, where the numbers of true positives are also roughly similar, they outperform the single-step models.

In terms of human evaluation, we presented the explanations generated by our WT5 model in the *M-M* setting to three domain experts, with 209 explanations provided to each of them. Table 4 provides four examples of positive predictions along with their generated explanations. The first three rows correspond to relevant explanations, since the three human assessors considered them as relevant. The last row shows a non-relevant explanation. In this case, while it appears to be topically related to a symptom of depression (*Suicidal ideas*), the experts classified it as non-relevant, as it does not provide any information about the symptom. Overall, the three assessors found 73%, 53%, and 77% of the explanations relevant and clinically useful, resulting in an average of 68%.

A more granular inspection of the assessors' annotations indicates that, out of 209 explanations, 154 were deemed clinically useful by at least two of the three assessors. In terms of total consensus, which is an important factor as it unambiguously confirms or rejects the practical utility of an explanation, we find that assessors agreed on the non-relevance of 25 explanations and the relevance of 86. Achieving high agreement in the mental health domain is challenging [38, 41, 53, 76]. Even so, in our expert evaluation we can observe similar trends to the results of the annotation process of the BDI-Sen dataset [53]. The presence of varying opinions among mental health professionals regarding certain explanations underlines the necessity for professional interpretation of the model outputs.

Table 4: Each row presents an example of positive predictions in the *M-M* setting using the WT5 model. Explanations are highlighted as judged by the assessors. The *Relevant* column displays the relevance labels after the human assessment. Samples have been paraphrased for privacy and shortened for clarity.

Input post	Relevant
(. . .) I want to start a business one moment, then pay out my IRA and travel throughout Europe. I do not comprehend who I am. **My short-term memory is terrible, and I can not concentrate** . I'm unsure of what to do. You guys are going to advise some really fantastic actions for me to pursue, but ultimately I lack the willpower or energy to carry out your advice.	1
Recently, I have been having a lot of difficulty with this. **I have been depressed, worried, or ill since I was a child** . Like my youth has been taken from me. As a last-ditch effort to feel like a person once more, I am actually considering seeing a naturopathic physician. Anyway, I hope you soon feel better.	1
I abhor myself to the core. Even reading back through Reddit postings I posted a few days ago makes me want to commit suicide. **I am such a disgusting waste of life — useless, unproductive, and with a future that is already in uncertainty** . And the fact that Im feeling this way on spring break is something I detest so fiercely. (. . .)	1
(. . .) how EMTs and first responders are looking after them, and how those individuals should persevere to witness another day. **Nonetheless, I find myself unable to avoid fall into thoughts about someone's death** . This triggers memories of my own experiences and how I might find myself in that person's position (. . .)	0

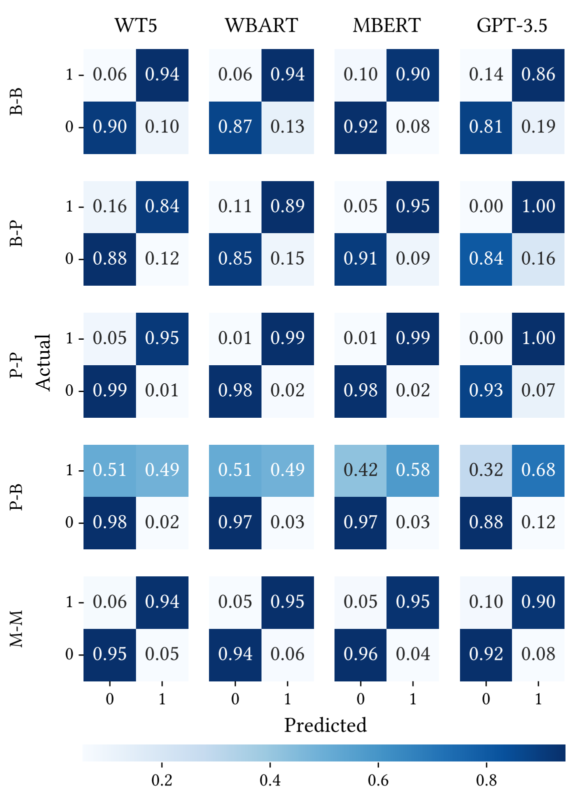

Figure 2: Confusion matrices showing the predictions accuracy for the proposed settings and WT5, WBART, MBERT and GPT-3.5 systems.

5.2 RQ2: How many labels are needed to generate good explanations?

Data labelled by domain experts is scarce and onerous, especially in the mental health domain [24, 38]. Therefore, we decided to investigate the impact of the number of labelled examples on the quality of explanations for depressive symptom detection. For that, we trained different models with a gradually larger number of examples in the *M-M* setting. With 886 explanations available in the training set, we compared the models' performance using subsets of 100, 200, 400, and 800 training samples. Figure 3 presents the results for the three evaluation metrics (ROUGE, BLEU, and TF1). As shown in the figure, there is a quasi-linear improvement in the number of training instances for all metrics and models. For instance, with only 100 explanations, the models achieve approximately half the performance compared to the entire training split. For completeness, we also investigated the performance of the classification task. We observed much robust behaviour for it. From 200 samples onwards, the F1 is higher than 0.9 for every model. This supports the idea that labels' quantity (and quality) are much more critical for explanation than classification.

5.3 RQ3: Single-step vs Two-steps

Through our experimentation, we conducted a comparative analysis of models trained to perform the *classify + explain* tasks in a single step (WT5 and WBART) against those operating in two separate steps (T5 + T5, BERT + T5, MBERT + T5). Regarding the classification task, the two-step models exhibited a slight advantage. Specifically, the BERT-based models consistently outperformed the other models, with MBERT achieving the best classification performance in three of the five settings. This performance improvement is reasonable, given that MBERT was the only model specifically pre-trained on corpora related to mental health domain [28]. Regarding explanation, to ensure a fair comparison, we focused on

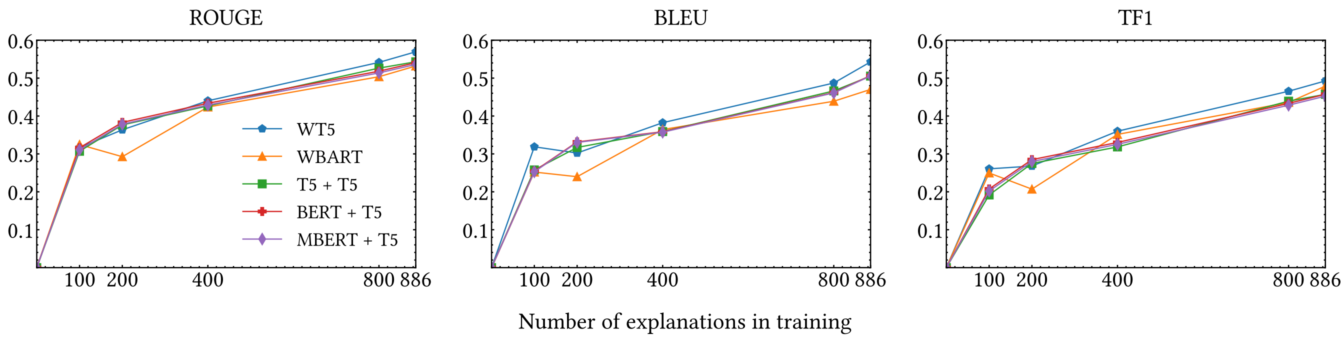

Figure 3: Offline metrics in relation to the quantity of explanations kept in the training data for the M-M setting.

settings with a similar number of true positives (TPs), which led us to consider the *P-P* and *M-M* settings. In both cases, the models with the best explanations were WT5 and WBART. These findings align with previous research, suggesting that classifying and explaining in the same step can contribute to higher-quality explanations [43].

5.4 RQ4: Conversational LLMs performance

We evaluated the performance of two recent conversational LLMs, Vicuna and GPT-3.5, in detecting and explaining depressive symptoms. The overall results presented in Table 3 indicate that Vicuna and GPT-3.5 achieved lower performance than the fine-tuned systems in classification and explanation tasks. Specifically, Vicuna obtained F1 values significantly below the rest of the models for classification and explanation. On the other hand, GPT-3.5's classification performance is closer to the other models. The quality of explanations generated by both LLMs is still inferior, with GPT-3.5 achieving better results than Vicuna. Despite these results, it is important to consider that we only used 30 positive and negative examples to guide these two systems through in-context learning (we have context token limits). This limited exposure to examples for the new task might have contributed to their relatively lower performance. Moreover, recent studies have shown the variability in the performance of these models depending on the prompts quality [74, 78].

6 CONCLUSIONS AND FUTURE WORK

This paper presents promising findings on the use of a text-to-text pipeline to classify and explain the presence of depressive symptoms on social media. Our models demonstrate high accuracy in their classification. Furthermore, the expert-in-the-loop evaluations further underscore the practicality and readability of the models' generated explanations. A comparison of the single-setp approach versus the two-steps (separate classification and explanation) method reveals that a unified framework is a viable and effective solution. Our study enriches the expanding field of explainable AI in mental health applications. By offering healthcare systems models capable of explaining their decision-making process, we allow clinicians to understand the automated reasoning and build trust in the outputs of these models.

A limitation of this study is that the data examined comes from a single social media platform. As future work, it would be valuable to apply our approach across diverse platforms, languages, and cultures to evaluate its cross-cultural applicability. While we presented an extractive approach for generating explanations, our discussions with domain experts have revealed the potential to create abstract explanations, matching the way one expert would explain the existence of depressive symptoms to another. Moreover, it might be advantageous to identify the precise symptom present in positive posts, leaving behind the actual binary paradigm and defining a new multi-class approach. We believe the creation of models capable of explaining the detection of clinical markers is a critical first step towards developing sophisticated natural language explanations for depression detection models.

7 ETHICAL AND ENVIRONMENTAL ASPECTS

The data used in our study were obtained from publicly accessible sources, adhering to the exempt status under title 45 CFR §46.104. The use of both the BDI-Sen and PsySym datasets was accomplished in full compliance with their respective data usage policies. To maintain privacy, we implemented measures to ensure that any personal information was unidentifiable and all users remained anonymous. The data used were sourced from Reddit, and we strictly conformed to all terms specified by this platform. For the human evaluation, even though the assessors were domain experts, they were not subjected to any imposed time restrictions, and reported no adverse effects post-evaluation. Importantly, it must be emphasized that the systems described in this study are intended to complement the work of healthcare professionals, not replace them. The development of such technologies necessitates a cautious approach, with a continuous emphasis on their ethical use and a firm respect for patient privacy and autonomy.

The experiments for this study were conducted using our private infrastructure, with a carbon efficiency of 0.432 $kgCO_2eq/kWh$, which reflects the OECD's 2014 yearly average. The resources utilised included 10 hours of computation on an RTX A6000 device (with a TDP of 300W) and 50 hours on an A100 PCIe 40/80GB device (with a TDP of 250W). Total emissions are estimated to be 6.7 $kgCO_2eq$. To provide some perspective, this is equivalent to driving an average car for 27 kilometres. These figures were determined with the assistance of the MachineLearning Impact Calculator [31].

REFERENCES

[1] Hessam Amini and Leila Kosseim. 2020. Towards Explainability in Using Deep Learning for the Detection of Anorexia in Social Media. In *Natural Language Processing and Information Systems*, Elisabeth Métais, Farid Meziane, Helmut Horacek, and Philipp Cimiano (Eds.). Springer International Publishing, Cham, 225–235.

[2] Mario Aragon, Adrian Pastor Lopez Monroy, Luis Gonzalez, David E. Losada, and Manuel Montes. 2023. DisorBERT: A Double Domain Adaptation Model for Detecting Signs of Mental Disorders in Social Media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 15305–15318.

[3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to Explain Individual Classification Decisions. *J. Mach. Learn. Res.* 11 (aug 2010), 1803–1831.

[4] Aaron T. Beck, R. A. Steer, and G. Brown. 1996. Beck Depression Inventory–II.

[5] Amin Beheshti, Vahid Moraveji-Hashemi, Shahpar Yakhchi, Hamid Reza Motahari-Nezhad, Seyed Mohssen Ghafari, and Jian Yang. 2020. Personality2vec: Enabling the Analysis of Behavioral Disorders in Social Networks. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 825–828.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901.

[7] Ana-Maria Bucur, Adrian Cosma, Paolo Rosso, and Liviu P Dinu. 2023. It's Just a Matter of Time: Detecting Depression with Time-Enriched Multimodal Transformers. In *European Conference on Information Retrieval*. Springer, 200–215.

[8] Fidel Cacheda, Diego Fernandez, Francisco J Novoa, and Victor Carneiro. 2019. Early detection of depression: social network analysis and random forest techniques. *Journal of medical Internet research* 21, 6 (2019), e12554.

[9] Amy Callahan and Kay Inckle. 2012. Cybertherapy or psychobabble? A mixed methods study of online emotional support. *British Journal of Guidance & Counselling* 40, 3 (2012), 261–278.

[10] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2023. A survey on XAI and natural language explanations. *Information Processing & Management* 60, 1 (2023), 103111.

[11] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc.

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

[13] Fabio Crestani, David E Losada, and Javier Parapar. 2022. *Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the ERisk Project*. Vol. 1018. Springer Nature.

[14] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, Vol. 7. 128–137.

[15] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[16] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4443–4458.

[17] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning.

[18] Karl Fredrik Erliksson, Anders Arpteg, Mihhail Matskin, and Amir H. Payberah. 2021. Cross-Domain Transfer of Generative Explanations Using Text-to-Text Models. In *Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, Proceedings* (Saarbrücken, Germany). Springer-Verlag, Berlin, Heidelberg, 76–89.

[19] Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16.

[20] Riccardo Guidotti and Salvatore Ruggieri. 2019. On The Stability of Interpretable Models. In *2019 International Joint Conference on Neural Networks (IJCNN)*. 1–8.

[21] Amelia Gulliver, Kathleen M Griffiths, and Helen Christensen. 2010. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. *BMC Psychiatry* 10 (Dec. 2010), 113.

[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 (2017), 43–49. Big data in the behavioural sciences.

[23] Aron Halfin. 2007. Depression: the benefits of early and appropriate treatment. *Am J Manag Care* 13, 4 Suppl (Nov. 2007), S92–7.

[24] Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize?. In *Findings of the association for computational linguistics: EMNLP 2020*. 3774–3788.

[25] Tobias U Hauser, Vasilisa Skvortsova, Munmun De Choudhury, and Nikolaos Koutsouleris. 2022. The promise of a model-based psychiatry: building computational models of mental ill health. *The Lancet Digital Health* 4, 11 (2022), e816–e828.

[26] Gordon Jackson-Koku. 2016. Beck depression inventory. *Occupational Medicine* 66, 2 (2016), 174–175.

[27] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3543–3556.

[28] 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*. European Language Resources Association, Marseille, France, 7184–7190.

[29] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[30] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine* 16, 9 (2001), 606–613.

[31] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700* (2019).

[32] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1675–1684.

[33] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880.

[34] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81.

[35] Silvia Lopez-Larrosa, Vanesa Sánchez-Souto, David E. Losada, Javier Parapar, Álvaro Barreiro, Anh P. Ha, and Edward M. Cummings. 0. Using Machine Learning Techniques to Predict Adolescents' Involvement in Family Conflict. *Social Science Computer Review* 0, 0 (0), 08944393221084064.

[36] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[37] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[38] Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian. 2018. RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Association for Computational Linguistics, New Orleans, LA, 168–173.

[39] Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive

survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics* 113 (2021), 103655.

[40] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[41] Danielle Mowery, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen Coppersmith, Craig Bryan, Mike Conway, et al. 2017. Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study. *Journal of medical Internet research* 19, 2 (2017), e6895.

[42] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1101–1111.

[43] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions.

[44] Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. 2022. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022*. 2563–2572.

[45] Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 8446–8459.

[46] Noratikah Nordin, Zurinahni Zainol, Mohd Halim Mohd Noor, and Lai Fong Chan. 2023. An explainable predictive model for suicide attempt risk using an ensemble learning and Shapley Additive Explanations (SHAP) approach. *Asian Journal of Psychiatry* 79 (2023), 103316.

[47] Cardwell C Nuckols and Cardwell C Nuckols. 2013. The diagnostic and statistical manual of mental disorders,(DSM-5). *Philadelphia: American Psychiatric Association* (2013).

[48] The National Study of Mental Health and Wellbeing. 2022. *Mental health: prevalence and impact*. Technical Report. Australian Institute of Health and Welfare.

[49] OpenAI. 2022. Introducing ChatGPT. [Accessed August 1, 2023].

[50] World Health Organization et al. 2017. *Depression and other common mental disorders: global health estimates*. Technical Report. World Health Organization.

[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318.

[52] Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2023. eRisk 2023: Depression, Pathological Gambling, and Eating Disorder Challenges. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*. Springer, 585–592.

[53] Anxo Perez, Javier Parapar, Alvaro Barreiro, and Silvia Lopez-Larrosa. 2023. BDI-Sen: A Sentence Dataset for Clinical Symptoms of Depression. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[54] A Picardi, I Lega, L Tarsitani, M Caredda, G Matteucci, M P Zerella, R Miglio, A Gigantesco, M Cerbo, A Gaddini, F Spandonaro, M Biondi, and SET-DEP Group. 2016. A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. *J Affect Disord* 198 (March 2016), 96–101.

[55] Anxo Pérez, Javier Parapar, and Álvaro Barreiro. 2022. Automatic depression score estimation with word embedding models. *Artificial Intelligence in Medicine* 132 (2022), 102380.

[56] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21, 1, Article 140 (jan 2020), 67 pages.

[57] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[58] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018).

[59] Esteban A Ríssola, David E Losada, and Fabio Crestani. 2021. A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare* 2, 2 (2021), 1–31.

[60] Farig Sadeque, Dongfang Xu, and Steven Bethard. 2018. Measuring the Latency of Depression Detection in Social Media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (Marina Del Rey, CA, USA) *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 495–503.

[61] Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2931–2951.

[62] Hong-Han Shuai, Chih-Ya Shen, De-Nian Yang, Yi-Feng Lan, Wang-Chien Lee, Philip S. Yu, and Ming-Syan Chen. 2016. Mining Online Social Data for Detecting Social Network Mental Disorders. In *Proceedings of the 25th International Conference on World Wide Web* (Montréal, Québec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 275–285.

[63] Ruba Skaik and Diana Inkpen. 2020. Using Social Media for Mental Health Surveillance: A Review. *ACM Comput. Surv.* 53, 6, Article 129 (dec 2020), 31 pages.

[64] 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*. Association for Computational Linguistics, Hong Kong.

[65] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 3104–3112.

[66] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. *arXiv e-prints*, Article arXiv:1606.05386 (June 2016), arXiv:1606.05386 pages.

[67] Martin Tutek and Jan Šnajder. 2022. Toward Practical Usage of the Attention Mechanism as a Tool for Interpretability. *IEEE Access* 10 (2022), 47011–47030.

[68] Berk Ustun and Cynthia Rudin. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* 102, 3 (01 Mar 2016), 349–391.

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[70] Colin G Walsh, Beenish Chaudhry, Prerna Dua, Kenneth W Goodman, Bonnie Kaplan, Ramakanth Kavuluru, Anthony Solomonides, and Vignesh Subbian. 2020. Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. *JAMIA Open* 3, 1 (01 2020), 9–15.

[71] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems*. Curran Associates Inc., Red Hook, NY, USA.

[72] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2017. Detecting and Characterizing Eating-Disorder Communities on Social Media. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Cambridge, United Kingdom) *(WSDM '17)*. Association for Computing Machinery, New York, NY, USA, 91–100.

[73] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 11–20.

[74] Guangxu Xun, Xiaowei Jia, Vishrawas Gopalakrishnan, and Aidong Zhang. 2017. A Survey on Context Learning. *IEEE Transactions on Knowledge and Data Engineering* 29, 1 (2017), 38–56.

[75] 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*. Association for Computational Linguistics, Copenhagen, Denmark, 2968–2978.

[76] Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Zhu. 2022. Symptom Identification for Interpretable Detection of Multiple Mental Disorders on Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 9970–9985.

[77] Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q. Zhu. 2022. Psychiatric Scale Guided Risky Post Screening for Early Detection of Depression. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 5220–5226. AI for Good.

[78] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*. PMLR, 12697–12706.

[79] Ayah Zirikly, Dana Atzil-Slonim, Maria Liakata, Steven Bedrick, Bart Desmet, Molly Ireland, Andrew Lee, Sean MacAvaney, Matthew Purver, Rebecca Resnik, and Andrew Yates (Eds.). 2022. *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, Seattle, USA.