BDI-Sen: A Sentence Dataset for Clinical Symptoms of Depression

Anxo Pérez
anxo.pvila@udc.es, Information Retrieval Lab, CITIC Universidade da Coruña, A Coruña, Spain
Álvaro Barreiro
alvaro.barreiro@udc.es, Information Retrieval Lab, CITIC Universidade da Coruña, A Coruña, Spain
Javier Parapar
javier.parapar@udc.es Information Retrieval Lab, CITIC Universidade da Coruña, A Coruña, Spain
Silvia López-Larrosa
silvia.lopez.larrosa@udc.es Departamento de Psicología Universidade da Coruña, A Coruña, Spain

ABSTRACT

People tend to consider social platforms as convenient media for expressing their concerns and emotional struggles. With their wide-spread use, researchers could access and analyze user-generated content related to mental states. Computational models that exploit that data show promising results in detecting at-risk users based on engineered features or deep learning models. However, recent works revealed that these approaches have a limited capacity for generalization and interpretation when considering clinical settings. Grounding the models' decisions on clinical and recognized symptoms can help to overcome these limitations. In this paper, we introduce BDI-Sen, a symptom-annotated sentence dataset for depressive disorder. BDI-Sen covers all the symptoms present in the Beck Depression Inventory-II (BDI-II), a reliable questionnaire used for detecting and measuring depression. The annotations in the collection reflect whether a statement about the specific symptom is informative (i.e., exposes traces about the individual's state regarding that symptom). We thoroughly analyze this resource and explore linguistic style, emotional attribution, and other psycholin-guistic markers. Additionally, we conduct a series of experiments investigating the utility of BDI-Sen for various tasks, including the detection and severity classification of symptoms. We also examine their generalization when considering symptoms from other mental diseases. BDI-Sen may aid the development of future models that consider trustworthy and valuable depression markers.

CCS CONCEPTS

Information Systems; Information Retrieval; Retrieval tasks and goals; Clustering and classification.

KEYWORDS

Social media mining; Depression detection; Symptom detection; Depression dataset.

1 INTRODUCTION

People who value independence, privacy and, sometimes, anonymity feel comfortable speaking about mental issues on social media platforms [21]. Researchers have even found that young people are more prone to discuss sensitive issues online than face-to-face [4].

The early identification and diagnosis of mental disorders are critical for effective treatment reducing mortality and morbidity, including the costs associated with misdiagnosis [35]. Numerous studies have highlighted how early interventions reduce the negative impact of mental disorders [15, 44]. In this context, the amount of user-generated data from social media allowed researchers to investigate online indicators of mental health conditions. Social media platforms became a source of information to provide early interventions that are low-cost and non-invasive [11]. Rich bodies of work have shown encouraging results in determining the presence of mental diseases exploiting social media writings from different platforms [6, 8, 50]. In the case of depression, researchers have focused on identifying depressive patterns and linguistic markers to develop predictive models [3, 55], obtaining great accuracy on multiple evaluation benchmarks [41, 62].

Although the pioneer models in the mental health domain showed good predictive performance, there are still significant gaps towards their real integration in clinical settings [12, 57]. One major limi-tation is that these models have a limited generalization capacity (i.e., when extending the models to other social platforms or collec-tions) [16, 30]. Another area for improvement is the interpretability of these models. They often lack transparency in their decision-making processes in a domain where interpretability is essential for clinicians to validate a diagnosis based on automated screening results [57]. To overcome the above limitations, a recent line of work focused on developing models that integrate depressive symptoms as reliable clinical markers. However, most of the existing datasets on depression detection only provide binary labels at the user level (depressive vs control users) [5, 32, 59]. Recently, the eRisk depression severity shared task [28] made pioneer contributions to promote the integration of symptoms detection, as they were the first to release a dataset containing user-produced labels at the symptom level. The eRisk severity estimation collections contain social media user responses to the BDI-II [2], a questionnaire that includes 21 recognized symptoms, such as sadness, irritability or fatigue. These new types of datasets allowed the leverage of depression markers from standard questionnaires to construct detection models. As a result, new symptom-based models demonstrated their potential to improve traditional approaches regarding performance, interpretability and generalization [37, 46, 60].

In this paper, we introduce BDI-Sen to promote further the development of models based on symptom markers to identify depressive signs. BDI-Sen is a dataset comprising 4973 annotated sentences covering depressive symptoms and 41 200 control sentences. Follow-ing a similar approach to PsySym [61], which includes symptoms of different diseases based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-5)

[38], we identify relevant sentences to depressive symptoms. However, in our case, the sentences are associated with users' responses to the 21 BDI-II symptoms. BDI-Sen is also a valid resource for ranking representative sentences of depressive symptoms, following the recent CLEF eRisk task [42]. Adhering to established clinical schemas for diagnosing depression, such as BDI-II, is crucial for facilitating the integration of more effective and consistent diagnostic support tools.

For building our dataset, we first semantically ranked the whole sentences from the annotated users for relevance to a symptom. To do so, we estimated semantic similarities using sentence transformers embeddings [48] and relying on the descriptions of elements provided by the BDI-II as queries. In the second phase, we follow a manual annotation schema as in similar works [20, 32, 36] where experts decided the actual relevance of the filtered candidates.

Our study includes a symptom-by-symptom analysis of the language and emotional characteristics of the annotated sentences. Additionally, we perform experiments to validate the usefulness of BDI-Sen for various tasks, including the detection and severity estimation of symptoms. Using a wide range of classification models, we find that the methods can effectively detect sentences representative of depressive symptoms. However, when considering different severity risk levels, we observe a significant decrease in performance. Further examination via error analysis reveals the challenge of distinguishing between closely related severity levels. Finally, we investigated the generalization of our models to symptoms from other mental diseases, showing that the models trained on our dataset can generalize well. BDI-Sen dataset and the code implemented is available under the eRisk dataset research license[1].

[1] https://erisk.irlab.org/BDISen.html

2 RELATED WORK

Traditional research about social media and mental health activity has thoroughly addressed how the users' self-disclosure of their illnesses and symptoms affects their mental health. Many studies have analyzed the benefit of self-disclosure on social users [17, 18, 31]. Self-disclosure refers to any communication about oneself communicated by an individual to others. Mental health-related self-disclosure on social media may help users to perceive higher levels of social support [17] and reduce psychological distress [51]. In addition, positive disclosure tends to produce more positive feedback for the community, improving the connectedness feelings [33]. Moreover, people being authentic in the disclosure shows improvements in self-esteem [58]. These and other reasons (like anonymity [1] or reduced perception of vulnerability [24]) explain why many social media users are prone to talk about their inner selves in those public forums. The self-disclosure of individuals' information motivated extensive work on detecting mental diseases from social media [14, 53, 54]. Mental disease detection models aim to automatically label users with published traces of mental illnesses based on online information. Early research in this field focused primarily on de-tecting depression [9], which is still the predominant mental disease in this domain. Later, other studies expanded the scope to include multiple conditions [54]. There are two main types of approaches for mental disease detection: those that use traditional machine learning algorithms with engineered features such as bag-of-words, topic modelling, or Linguistic Inquiry and Word Count (LIWC)[3, 43], and those using deep neural networks to encode users' writings [25, 39, 63]. However, models under both paradigms tend to suffer from a lack of generalizability [16, 37]. Additionally, most deep neural methods are considered black-boxes that cannot pro-vide reliable explanations. On the other hand, the use of engineered

features are often too general (e.g., word counts, posting hours) to provide personalized user-level explanations [36, 60].

Recent studies have explored the creation of symptom-based prediction models for signs of depression. These models showed the importance of presenting reliable depression markers to aid health professionals in their diagnosis [7]. Most of them leveraged the use of Large Language Models (LLMs) to design classifiers [37, 46, 60]. For instance, Zhang et al. [60] developed a BERT-based model that aggregates markers from different clinical inventories to calculate the risk of symptoms at the post level. To improve the efficiency of their approach, they designed templates based on standard questionnaires to pre-filter only representative posts.

Nguyen et al. [37] also explored BERT-based methods using symptom classifiers and compared them against a standard depres-sion classifier. They used the nine symptoms from the 9-Question Patient Health Questionnaire (PHQ-9) [22] to design the symptom classifiers. Covering three different datasets, the authors found that these classifiers performed well compared to the standard depression classifier while generalizing better to other datasets. Moreover, the authors found that when leveraging the weights from the attention architecture, these symptom classifiers provide a model that can highlight specific posts based on relevant symptoms, improving their interpretability. In addition to this, the works proposed to solve the *eRisk depression estimation shared task* were pioneering contributions to the development of symptom detection models, where their approaches predicted the BDI-II symptom responses of Reddit users. For a detailed analysis of these works, we refer the reader to the corresponding shared task overviews [28, 29, 41].

To facilitate the development of mental health detection models, the research community created datasets covering various mental illnesses. Due to the complexity and time-consuming nature of data acquisition approaches, researchers in this domain have used different heuristics to automatically label users. These heuristics are commonly known as "proxy diagnostic signals". For depressive disorder, most of these proxies are based on self-diagnostic statements in social media (e.g., "*I have been diagnosed with depression*") [6, 8, 32, 52, 59]. While relying on self-diagnostic statements has its limitations [12], it is an effective way to obtain a sufficient amount of data in a low-effort, and unobtrusive manner. Copper-Smith et al. [6] and Losada and Crestani [27] made pioneer contributions by releasing a public collection based on self-expression from depression diagnoses on Twitter and Reddit, respectively.

Following the works mentioned above, we can find collections that go beyond providing binary labels for depression detection. Some datasets focused on more diverse aspects, such as including temporality in their annotations [32], or the combination of different modalities of data, such as text and images [52]. When considering the time factor, the collections released by the shared tasks of *early* risk detection (eRisk) add further challenges over the classical binary classification problem [28, 29, 41]. Additionally, the eRisk collections on depression severity estimation adopt a human-in-the-loop approach, requiring self-reported information directly from individuals. These collections consist of users' social media posts and the real users' responses to the 21 BDI-II symptoms.

In another exciting contribution to this new trend of symptom-based models for the mental health detection domain, Zhang et al.[61] released the PsySym dataset. This dataset is the first annotated symptom sentence dataset that covers multiple mental disorders. PsySym includes annotations of 38 symptoms from 7 mental disorders. The

authors established the symptom classes according to the DSM-5, accompanied by the descriptions of diverse inventories. While our work shares similar motivations, we differ in our approach by adhering directly to the clinical questionnaire of the BDI-II and providing the actual responses of the writers to the analysed symptoms' questions. The present study represents a step towards considering reliable symptoms as depression markers to design more robust mental health detection models.

3 BDI-SEN DATASET

This section describes the construction and annotation schema of the BDI-Sen dataset. We create a symptom-based dataset with relevant sentences that trace the presence of clinical symptoms. For this reason, we develop an annotation schema based on the BDI-II [2], a highly reliable tool to diagnose depression in clinical settings [23]. The BDI-II covers 21 recognized symptoms, including emotional, cognitive and physical markers. Each item in the questionnaire has four alternative option responses scaled in severity from 0 (least severe) to 3 (most severe). These options have a textual description associated. Table 1 provides an example of the option descriptions for the symptom *Loss of energy*. To create the BDI-Sen dataset, we used as data source the eRisk2019 depression severity collection [28], which contains social media users' publications from Reddit and their responses to the BDI-II symptoms. We used Reddit as the target platform due to its wide acceptance in previous studies [5, 32, 41, 61].

3.1 Dataset Construction

Candidate Sentences Selection. The large volume of publications from eRisk2019 training users requires an exhaustive filter for reasonable annotation efforts. For this

reason, we design an initial retrieval stage based on filtering candidate sentences that may be relevant to each symptom. The retrieval phase uses the options' descriptions (severity descriptions) as queries to select the candidate sentences. We generate four queries (one per severity level) and search the entire set of sentences. For this purpose, we produce semantic rankings using cosine-similarity with sentence transformers [48] leveraging a pre-trained model based on RoBERTa [26]. To obtain a reasonable balance between the amount and quality of candidates, we conducted pilot experiments involving expert annotators[2]. We presented them with candidate sentences from different similarity thresholds. This process resulted in a minimum value of 0.6 to filter out candidates. We further restricted the assessor's work to the first 750 ranked sentences in those symptoms where this threshold produced too many candidates.

Annotation schema. After selecting the candidate sentences, a team of expert annotators consisting of a psychologist, a speech therapist, and a PhD student with knowledge in the field were responsible for annotating BDI-Sen. The three annotators individually examined the whole set of candidates, with all supplementary metadata removed beforehand to avoid potential bias. To ensure the quality of the labels, we conducted training sessions with the annotators. We discussed the labelling rules with all of them, providing examples of positive and negative cases for each symptom. We instructed the annotators on the goal of our study, and explained the concept of relevance: a sentence is deemed relevant if it offers infor-mation about the specific symptom for the individual. Specifically, each annotator answered the following question in a binary setting (Positive/Negative): *Does the sentence offers information about the symptom, and the user talks in first person?* If in doubt, annotators could leave a sentence unlabelled, and

[2] Prior research showed a high variance in symptoms distributions, since for some of them is easier to retrieve relevant sentences [36, 61].

there was no time limit on their annotations. We presented the sentences for each symptom in a different ranking, and the same sentence can appear in the rankings for different symptoms. Each sentence was considered positive following a majority voting approach among the annotators' decisions. Finally, we obtained a total of 4973 annotated sentences. The interannotation agreement among the three annotators was 84.93%, which is a substantial agreement compared to similar works [32, 36, 61].

Overall Annotation Results. Table 2 shows the number of positive, negative and control sentences obtained from the publications of the eRisk2019 users: (1) *Positive sentences* are those identified as relevant to the BDI-II symptoms, with a total of 853 sentences. (2) *Negative sentences* represent the highest percentage of annotations, totalling 4120. Despite being semantically related to the symptom, the negative sentences are not relevant to it. However, they can still be valuable for developing efficient depression detection models, being examples of false positives, one of the main challenges in detecting risks in social media [30]. (3) We include a set of *Control sentences*. For each symptom, we obtain ten sets of control sentences, each set having the same number of sentences as the negative group. The control sentences were randomly obtained from the rest of the sentences not selected for annotation. The experts annotated the 17% sentences from the pool of candidate ones as relevant. Among the BDI-II symptoms, *Loss of pleasure* has the most annotations (739), while *Low libido* has the least (24). Comparing the positive and negative groups, we can see that the number of sentences annotated as negative is always higher than the number of positive ones.

Severity Weak Labels. In addition to the relevance labels provided by our annotators, using the eRisk2019 users as data source allowed us to include severity labels (0-3) for each BDI-II symptom. The severity labels correspond with the responses from users who

authored the sentences to the BDI-II. We leveraged this additional information using a weakly-supervised approach to generate weak labels for each sentence. Specifically, we assigned the severity label corresponding to the user's response to each sentence related to the symptom. Table 3 shows examples of sentences from our dataset, along with their binary relevance labels and weak severity labels for the symptom Sleep issues. For instance, looking at the sentence *"I just have energy to eat and sleep"*, its author responded 3 for that symptom. Therefore, the weak severity label is 3. This information allowed us to study the relationship between language and symptom severity at the sentence level, despite not having severity labels annotated by experts for each sentence.

3.2 Dataset Analysis

Next, we present an analysis of the constructed dataset. This section aims to determine if there are any differences among the three groups (positive, negative, and control) and among the positive group along the different symptoms. Following Rissola et al. approach [49], we analyze psycholinguistic and emotional features that characterize the writing style from the groups [5, 50, 59]. While the previous works studied the overall language of positive individuals vs control ones, we present the analysis at the symptom level in this case. First, Table 2 shows the main statistics and vocabulary comparison of the three groups of sentences for each symptom. The first block (first four rows) corresponds to the average annotation agreement of the symptom, along with the number of sentences per group. We note the high agreement among the symptoms, with only five having an agreement of less than 80%. While we considered including the Cohen's kappa coefficient, we decided against it since our dataset labels were highly unbalanced. In scenarios where labels are very unbalanced (e.g., our positive sentences represent a small

percentage of the candidates), kappa can be a misleading measure of agreement. In particular, for rare classes, very low kappa values may not necessarily reflect low rates of overall agreement [56]. Therefore, using the average annotator agreement may be a more appropriate measure.

Words Usage. The second block of Table 2 corresponds to the Jaccard index between the sentence groups. This index is a statistic used to quantify the diversity of sample sets [13]. Therefore, the higher the Jaccard value, the more similar the use of words from the groups[3]. Visualizing these results, we see that positive vs control are the groups with the least common vocabulary. For example, in some symptoms like *Punishment feelings*, they only share the 4% of the vocabulary. On the other hand, the most similar groups are positive vs negative (average Jaccard index of all symptoms of 15.60%) and control vs control (average of 18.81% over the control sets). The former makes distinguishing between negative and positively labelled sentences hard when only considering bag of words models (e.g. "you might be having trouble sleeping." is a challenging negative sentence)

Words Distribution. We analyzed the differences in word probability distributions among groups. The third block of Table 2 re-ports the difference in word probability distributions among groups. We measured how the probability distributions (i.e., the language models) differ using Kullback-Leibler divergence (KL). If the two distributions are identical, the KL value is 0. Visualizing the numbers, we observe that the word distributions for most symptoms have more KL when comparing positive vs control groups. Again we observe lower similarities between positive and negative groups

[3] Please note that the comparison with the control group is always the averaged value over the ten sampled control sets.

Finally, Figure 1 illustrates the kernel density estimation of the word probabilities of six BDI-II symptoms.[4] The x-axis represents the logarithm of the word probabilities. Thus, the right side of this axis corresponds to the words with higher probabilities (i.e., frequent words). The y-axis corresponds to the kernel density es-timations. We compare the word distributions of the LMs from the three groups considered. We may observe apparent differences between the control vs positive/negative groups. The word probabilities in the control groups result in a high density of words with high frequencies (i.e., the control group uses common words more frequently). However, that is not the case in the positive and negative groups, where many used words correspond to less probable terms (i.e., they use uncommon terms more frequently). Moreover, the distributions of positive vs negative groups show more differences on the right side of the x-axis (associated with high probability words), where the positive group uses more common words than the negative. These differences may correspond with first-person pronoun use (more popular and more used by depressed individuals [40]) versus second-person pronoun usage.

Emotions and Sentiments Association: Similar to prior works that revealed significant differences in emotional expressions between depressive and control groups [8, 49], we investigated to extend this type of analysis at the symptom level. We used the Plutchik set of emotions [45], which considers: 1) eight primary *emotions*: anger, fear, sadness, disgust, surprise, anticipation, trust, and joy and 2) two basic *sentiments*: positive (SP) and negative (SN). To quantify the emotion levels, we relied on the NRC emotion lexicon [34], which includes a set of words associated with the Plutchik emotions. In our analysis, we calculated the percentage of sentences from each group (positive, negative, control) that contain at least one word associated with the primary emotions and sentiments.

[4] Due to page limitations, we did not include the illustration for all the symptoms, but we found similar patterns on them.

When comparing these results among the groups, we identified two different patterns among the symptoms. We illustrate both in Figure 2. The symptoms in the first row show marked differences between the positive and negative/control groups. For example, for the symptom *Social Issues*, the percentage is always the highest in the positive sentences, with terms associated with *fear* or *SN* being in more than 30% of the positive sentences. Interestingly, the percentage of words referring to *SP* is also higher. This aligns with previous studies that demonstrated individuals with depressive conditions tend to be more emotional in social media [50]. However, in the second row, we observe a different pattern. The differences are much lower in this second type of symptom, with a high degree of overlapping.

4 EXPERIMENTS

In this section, we provide an experimental analysis to evaluate the impact of the BDI-Sen dataset preliminarily. As previously discussed, integrating clinical symptoms for developing mental health detection models has significant practical implications. For this reason, we divided our experiments into two tasks: 1) *Symptom Detection* and 2) *Symptom Severity Classification*. In the symptom detection task, we explored models identifying sentences relevant to BDI-II symptoms. On the other hand, the severity classification task leverages the four levels of severity of the BDI-II (corresponding with the four possible responses to each symptom, see Table 1 for an example) to classify the sentences based on them (0-3). In addition, to evaluate the generalization ability of our classification models, we also explored how the models trained on BDI-Sen behave on sentences from symptoms related to other mental diseases from the PsySym dataset [61].

4.1 Models

Similar to recent literature [37, 61, 61], we considered different types of pre-trained large language models (LLMs) formulated as classi-fiers. First, we used BERT-based models [10] for text classification. We finetuned the pre-trained **BERT** base uncased model, which represents a strong baseline. We also finetuned MentalBERT [19](**MBERT**)[5], a masked language model explicitly trained for the mental health domain. MBERT is pretrained with a corpus coming from subreddits associated with various mental diseases. As the last BERT variant, we included **BERT-mini**[6], a cost-effective alternative to BERT with fewer parameters, to explore the performance of a more lightweight model. Finally, we included **T5** (Text-to-Text Transfer Transformer) [47] in our experiments, which we finetuned to generate labels in textual form.

In addition to these deep learning models, we included two traditional classification approaches based on textual features. We used TF-IDF features with a linear classifier based on Logistic Regression to predict the labels (**TF-IDF+LR**). We also explored text features derived from LIWC categories. LIWC [43] provides a set of linguistic categories that can extract psychological features from the text, such as the presence of words related to positive or negative emotions. We extracted the LIWC features for each sentence and employed those with an SVM classifier (**LIWC+SVM**). These two traditional approaches are good baselines for examining the improvements of complex deep learning models.

[5] https://huggingface.co/mental/mental-bert-base-uncased
[6] https://huggingface.co/prajjwal1/bert-mini

4.2 Experimental Settings

In all our experiments, we used three splits of our dataset corresponding with training/validation/testing in a ratio of 7:1:2. For the training and validation sets, we included the sentences annotated as positive, and we randomly selected the same number of control sentences to balance the labels. That resulted in 1194 sentences in the training set and 172 in the validation set. We included more control sentences for the testing split to simulate a more realistic scenario. When processing user data in social networks, most sentences are not about depressive symptomatology. In a real setting, there is a high unbalanced towards the control class. For this reason, the number of control sentences in the test set is always five times greater than the number of positive sentences (resulting in 1026 sentences in the test split).

Regarding model choices and hyperparameters, in case of the TF-IDF+LR model, we removed stopwords and used 5-fold-cross-validation with the regularization strength (C) as hyperparame-ter in the ranges: [0.1, 0.5, 1, 2, 5, 10, 100, 1000]. In LIWC+SVM, we incorporated all the 64 categories from LIWC, used 5-fold-cross-validation with linear and RBF kernels, and the penalty parameter C in the ranges: [0.1, 1, 10, 50, 100]. We follow the same procedure for all the transformer-based models by using existing implementations from the HuggingFace library. Thus, we did not include any additional hyperparameter tuning. Specifically, for MBERT, BERT, and BERT-mini, we used a learning rate of $2e^{-5}$, the maximum sequence length of 128 during 20 epochs and a batch size of 32. For T5, we used a learning rate of $1e^{-3}$, a maximum sequence length of 256 during 10 epochs with a batch size of 16.

4.3 Symptom Detection

Identifying symptoms is crucial for diagnosing and researching mental health diseases [57]. Therefore, detecting depressive symptoms may be highly beneficial for early detection from social media data. In the symptom detection task, our goal is to determine if a sentence is relevant to a depressive symptom or not. We formulate this task as a binary classification problem, where the models detect if the sentence is related to a depressive symptom (1) or not (0). For the T5 model, we finetuned it by training T5 to generate "*true"* or *"false*" tokens. Table 4 shows the results of all our methods considered for symptom detection.

The results in Table 4 show that all the methods have relatively high F1 and AUC, with AUC scores ranging from 0.83 to 0.95, and F1 scores from 0.62 to 0.83. We can see that the transformer-based models, except BERT-mini, performed better than methods based on textual features (TF+IDF and LIWC). The results also show that the standard BERT model and T5 perform similarly despite T5 being pre-trained on a larger corpus of data. In line with prior research [19, 61], the model pre-trained on mental health-related corpora (MBERT) achieved higher scores in almost every metric. The models seem to peform worse in terms of precision, with MBERT obtaining the best value (0.74) and LIWC with the worst (0.49).

To better understand these results, Figure 3 provides a visual representation of the distribution of true and false predictions made by each of our classification models. These numbers show that the transformer-based models have a high ratio of true positives, from 0.94 (BERT-mini) to 0.98 (T5 and BERT). The percentage of false negatives for these models is small, with less than 0.06 in all cases. On the other hand, the number of false positive errors is higher. In the mental health domain, missing an individual at risk of

being reviewed by professionals, is much more worrying than therapists examining a healthy person. For this reason, a good prediction performance for false negatives is crucial. Finally, we can also observe that the prediction errors of the methods using the textual features have the lowest accuracy overall.

4.4 Symptom Detection - Generalization

Recent studies have demonstrated the low generalizability of mental disease detection models [16]. In this experiment, we want to analyze whether models trained on the BDI-Sen data can generalize to detect symptoms from other mental illnesess. We evaluated this premise using the *PsySym* dataset [61]. The mental disorders covered by PsySym are *depression, anxiety, ADHD, bipolar disorder, OCD, PTSD and eating disorder*. We aimed to test the ability of our models to generalize across these conditions, given the potential overlap in symptom expression between different mental disorders. For this purpose, we used the same models in the *symptom detection task* (Section 4.3) trained in BDI-Sen and tested them over the symptoms of the seven mental disorders of PsySym.

Table 5 shows the results of our models on the PsySym data. The number of positive test sentences for each disease is indicated between brackets. We also display their number of common symptoms with the BDI-Sen symptoms (third row). We only considered positive sentences of each illness and reported the precision that the models achieved. Based on the figures, deep learning methods exhibited good generalization capabilities to other mental diseases. However, a significant performance gap exists between the models using textual features (TF+IDF and LIWC+SVM) and transformer-based ones. Specifically, the best-performing model, T5, achieved an average accuracy of 0.81 across

the symptoms for all diseases. Meanwhile, the worst (TF-IDF+LR) had a precision of only 0.44.

These results suggest that the models trained on our dataset can generalize well to symptoms from other mental diseases. However, as shown in the last row of Table 5, the performance varies among illnesses, indicating that some disorders may be more challenging to detect than others. Unsurprisingly, the models have their best accuracy when evaluated in depression, with an average of 0.81. For the other diseases, the results suggest that the more symptoms they share with BDI-Sen, the better the model performs. Specifically, we achieved at least 0.70 accuracy in anxiety, bipolar disorder, OCD, and PTSD. In contrast, the worst results correspond to ADHD and eating disorders, with accuracy numbers of 0.59 and 0.51, respectively. Conducting these types of multi-disease analyses may provide valuable insights into the similarities and differences be-tween different mental health conditions, potentially leading to new avenues of research.

4.5 Symptom Severity Classification

In this experiment, we aim to classify the sentences from BDI-Sen based not only on whether they are relevant to the symptom but according to the declared severity level. This task represents a step forward from our previous experiments enabled by the weak labels we provide in the BDI-Sen from the users' response to the BDI-II. By identifying the severity of each symptom, mental health detection models may provide a more nuanced and accurate diagnosis of an individual's situation. We formulate the task as a multi-classification problem. The models classify each sentence severity according to the BDI-II schema, with the levels ranging from 0 to 3 (see Subsection 3.1). We refer the reader to Tables 1 and 3 to see descriptions and example sentences from our datasets of

the severity levels. In this experiment, we used the same text classification models trained in a multi-class setting and considered two experimental variants:

(1) The first experiment considers all severity levels, which includes a separate category for control sentences that were randomly selected (i.e., unrelated to any symptom). The aim was also to investigate whether the multi-class classification models may distinguish sentences talking about the symptom in a non-negative way (severity level 0) from those unrelated to the symptoms (control). Table 6 presents the results of our classification methods under this setting. We can observe a significant decrease in performance compared to the symptom detection experiments, where only two classes were considered. Although all methods achieved a reasonably good Micro F1 score due to the large number of control sentences in the test set, there was poor performance in sentences in non-control classes. Furthermore, the gap in performance between the transformer-based and textual feature models is reduced, with T5 being the worst-performing method. MBERT remains the top-performing model across all severity levels.

To further analyze these results and examine prediction errors between categories, Figure 4 (a) presents the confusion matrices for the best-performing method (MBERT). The matrix shows very few misclassification errors between severities that are far apart. For instance, for the *True* sentences with severity label 3, none of them were labelled as 0 or control sentences. Similarly, for the sentences with severity level 2, only 6% of them were misclassified with the level 0, and none of them were misclassified as control. Overall, most prediction errors occurred between severity levels 1, 2 and 3, indicating that the models find it challenging to correctly distinguish between categories with subtle differences. These results may point out the need for more severity-labelled sentences to train models accurately with this level of granularity.

(2) After analysing the above results, we performed an additional experiment, combining in one class the control and labelled sentences with severity level 0. The rationale behind this is that, when using severity detection approaches, the main practical interest would be to detect high-risk sentences. Both severity level 0 (i.e., no risk) and control sentences may not provide much value to support the diagnosis (they would sum up zero to the BDI-II final score). The more severe and negative symptoms expressions are more likely to require attention. Therefore, we grouped them to investigate this more practical scenario. Table 7 shows higher Micro F1 values than the previous experiment. The accuracy of this new class is higher for all models than the one of the control class from previous results. However, even on most occasions, their F1 values are improved, the models still struggle to distinguish between severity levels with risk. MBERT is still the top-performing model, and its F1 scores for severity classes were 0.46, 0.25, and 0.41. T5 continues to be the worst model in this multi-class scenario. Finally, we also included the confusion matrices of the MBERT model in Figure 4 (b). As in the previous experiment, the matrices reveal that most misclassifications oc-cur between adjacent severity levels. Specifically, only 3% of the sentences with severity level 3 were mislabeled as severity level 0.

5 CONCLUSIONS

This paper presents BDI-Sen, a symptom-annotated dataset for depression that includes manually labelled sentences addressing the 21 BDI-II symptoms. By leveraging the eRisk2019 collections as data source, our dataset provides binary relevance labels for the BDI-II symptoms and weak labels regarding their severity level. We designed a retrieval phase to filter-out candidate sentences based on the descriptions of the BDI-II elements, and three experts decided the actual relevance of the candidates. We explored this

resource, revealing linguistic and emotional differences among the symptoms. Moreover, we performed two main experiments with state-of-the-art models trained solely on BDI-Sen: symptom detection and symptom severity classification, including an extensive error analysis for both tasks. The good generalization ability of our models further underlines the usefulness of BDI-Sen as a resource for developing robust mental health detection models.

6 ETHICAL IMPLICATIONS

The BDI-Sen sentences were obtained from publicly available sources and collected in such a way that they rely on the exempt status under title 45 CFR §46.104. We adhered to the corresponding data usage policies. We ensured that personal information could not be identified from the data. Despite being experts in the field, we know that the annotation of depressive symptoms may affect annotators. Annotators were not subjected to any time constraints and were free to take all the necessary breaks to mitigate any negative impacts. The annotators did not report any adverse effects after their work. Moreover, we emphasize that the classification models presented in this work aim to supplement their efforts rather than replace health professionals. The development of such technologies must be cautiously approached, ensuring their use is ethical and respects patient privacy and autonomy. We require the BDI-Sen users and researchers to accept a data usage and privacy agreement to avoid possible misuse.

7 ACKNOWLEDGMENTS

This work has received support from projects: PLEC2021-007662 (MCIN/AEI/ 10.13039/501100011033, Ministerio de Ciencia e Innovación, Unión Europea-Next

GenerationEU); Consellería de Ed-ucación, Universidade e Formación Profesional, Spain (accredita-tion 2019–2022 ED431G/01 and GPC ED431B 2022/33, PID2022-137061OB-C21) and the European Regional Development Fund, which acknowledges the CITIC Research Center.

REFERENCES

[1]. Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive Self-Disclosures, Responses, and Social Support on Instagram: The Case of #De-pression. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) (CSCW '17). Association for Computing Machinery, New York, NY, USA, 1485–1500. https://doi.org/10.1145/2998181.2998243

[2]. Aaron T Beck, Robert A Steer, Roberta Ball, and William F Ranieri. 1996. Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. *Journal of personality assessment* 67, 3 (1996), 588–597.

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

[4]. 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.

[5]. Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a Large-Scale Resource for Exploring On-line Language Usage for Multiple Mental Health Conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 1485–1497.

[6]. Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Baltimore, Maryland, USA, 51–60. https://doi.org/10.3115/v1/W14-3207

[7]. Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights* 10 (2018), 1178222618792860.

[8]. Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social Media as a Measurement Tool of Depression in Populations (*WebSci '13*). Association for Computing Machinery, New York, NY, USA, 47–56. https://doi.org/10.1145/2464464.2464480

[9]. 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.

[10]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv*:1810.04805 (2018).

[11]. Adam G Dunn, Kenneth D Mandl, and Enrico Coiera. 2018. Social media interventions for precision public health: promises and risks. *NPJ digital medicine* 1, 1 (2018), 47.

[12]. 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. https://doi.org/10.1145/3290605.3300364

[13]. Sam Fletcher, Md Zahidul Islam, et al. 2018. Comparing sets of patterns with the Jaccard index. *Australasian Journal of Information Systems* 22 (2018).

[14]. 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.

[15]. Aron Halfin. 2007. Depression: the benefits of early and appropriate treatment. *American Journal of Managed Care* 13, 4 (2007), S92.

[16]. 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*. Association for Computational Linguistics, Online, 3774–3788. https://doi.org/10.18653/v1/2020.findings-emnlp.337

[17]. Rebecca A Hayes, Caleb T Carr, and Donghee Yvette Wohn. 2016. One click, many meanings: Interpreting paralinguistic digital affordances in social media. *Journal of Broadcasting & Electronic Media* 60, 1 (2016), 171–187.

[18]. Hsin-Yi Huang. 2016. Examining the beneficial effects of individual's self-disclosure on the social network site. *Computers in human behavior* 57 (2016), 122–132.

[19]. 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. https://aclanthology.org/2022.lrec-1.778

[20]. Payam Karisani and Eugene Agichtein. 2018. Did you really just have a heart attack? Towards robust detection of personal health mentions in social media. In *Proceedings of the 2018 World Wide Web Conference*. 137–146.

[21]. Sylvia Deidre Kauer, Cheryl Mangan, and Lena Sanci. 2014. Do online mental health services improve help-seeking for young people? A systematic review. *Journal of medical Internet research* 16, 3 (2014), e3103.

[22]. 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.

[23]. Lourdes Lasa, Jose L Ayuso-Mateos, Jose L Vázquez-Barquero, FJ Dıez-Manrique, and Christopher F Dowrick. 2000. The use of the Beck Depression Inventory to screen for depression in the general population: a preliminary analysis. *Journal of affective disorders* 57, 1-3 (2000), 261–265.

[24]. Chung-Ying Lin, Peyman Namdar, Mark D Griffiths, and Amir H Pakpour. 2021. Mediated roles of generalized trust and perceived social support in the effects of problematic social media use on mental health: A cross-sectional study. *Health Expectation*s 24, 1 (2021), 165–173.

[25]. Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. 2014. User-level psychological stress detection from social media using deep neural network. In *Proceedings of the 22nd ACM international conference on Multimedia*. 507–516.

[26]. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/ARXIV.1907.11692

[27]. David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, CLEF 2016*, Évora, Portugal, September 5-8, 2016, Proceedings 7. Springer, 28–39.

[28]. 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*. Springer, 340–357.

[29]. David E Losada, Fabio Crestani, and Javier Parapar. 2020. eRisk 2020: Self-harm and depression challenges. In *European Conference on Information Retrieval*. Springer, 557–563.

[30]. Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psy-chology: From Keyboard to Clinic. Association for Computational Linguistics*, New Orleans, LA, 78–87. https://doi.org/10.18653/v1/W18-0608

[31]. Mufan Luo and Jeffrey T Hancock. 2020. Self-disclosure and social media: motivations, mechanisms and psychological well-being. *Current opinion in psychology* 31 (2020), 110–115.

[32]. 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. https: //doi.org/10.18653/v1/W18-0618

[33]. Anna Metzler and Herbert Scheithauer. 2017. The long-term benefits of positive self-presentation via profile pictures, number of friends and the initiation of relationships on Facebook for adolescents' self-esteem and the initiation of offline relationships. *Frontiers in psychology* 8 (2017), 1981.

[34]. Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence* 29, 3 (2013), 436–465.

[35]. C Brendan Montano. 1994. Recognition and treatment of depression in a primary care setting. *The Journal of clinical psychiatry* (1994).

[36]. 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.

[37]. Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the Generalizability of Depression Detection by Lever-aging Clinical Questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 8446–8459. https://doi.org/10.18653/v1/2022.acl-long.578

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

[39]. Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of twitter users. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*. 88–97.

[40]. Rosa María Ortega-Mendoza, Delia Irazú Hernández-Farías, Manuel Montes y Gómez, and Luis Villaseñor-Pineda. 2022. Revealing traces of depression through personal statements analysis in social media. *Artificial Intelligence in Medicine* 123 (2022), 102202. https://doi.org/10.1016/j.artmed.2021.102202

[41]. Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2021. Overview of eRisk 2021: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 324–344.

[42]. 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.

[43]. 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), 2001.

[44]. A Picardi, I Lega, L Tarsitani, M Caredda, G Matteucci, MP Zerella, R Miglio, A Gigantesco, M Cerbo, A Gaddini, et al. 2016. A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. *Journal of affective disorders* 198 (2016), 96–101.

[45]. Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*. Elsevier, 3–33.

[46]. Anxo Pérez, Javier Parapar, and Álvaro Barreiro. 2022. Automatic depression score estimation with word embedding models. *Artificial Intelligence in Medicine* 132 (2022), 102380. https://doi.org/10.1016/j.artmed.2022.102380

[47]. 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. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[48]. Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Em-pirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[49]. Esteban Andrés Ríssola, Mohammad Aliannejadi, and Fabio Crestani. 2020. Be-yond Modelling: Understanding Mental Disorders in Online Social Media. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I (Lecture Notes in Computer Science,* Vol. 12035), Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, 296–310. https://doi.org/10.1007/978-3-030-45439-5_20

[50]. Esteban A. Ríssola, Mohammad Aliannejadi, and Fabio Crestani. 2022. Mental disorders on online social media through the lens of language and behaviour: Analysis and visualisation. *Information Processing & Management* 59, 3 (2022), 102890. https://doi.org/10.1016/j.ipm.2022.102890

[51]. Mihye Seo, Jinhee Kim, and Hyeseung Yang. 2016. Frequent interaction and fast feedback predict perceived social support: Using crawled and self-reported data of Facebook users. *Journal of Computer-Mediated Communication* 21, 4 (2016), 282–297.

[52]. Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, TatSeng Chua, Wenwu Zhu, et al. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*. 3838–3844.

[53]. 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*. 275–285.

[54]. Ruba Skaik and Diana Inkpen. 2020. Using social media for mental health surveillance: a review. *ACM Computing Surveys (CSUR)* 53, 6 (2020), 1–31.

[55]. Marcel Trotzek, Sven Koitka, and Christoph Friedrich. 2018. Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences. *IEEE Transactions on Knowledge and Data Engineering* 32 (04 2018), 588–601. https://doi.org/10.1109/TKDE.2018.2885515

[56]. Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med* 37, 5 (2005), 360–363.

[57]. 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 (2020), 9–15.

[58]. Chia-chen Yang, Sean M Holden, and Mollie DK Carter. 2017. Emerging adults' social media self-presentation and identity development at college transition: Mindfulness as a moderator. *Journal of Applied Developmental Psychology* 52 (2017), 212–221.

[59]. 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. https://doi.org/10.18653/v1/D17-1322

[60]. 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 Intelli-gence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, Luc De Raedt (Ed.). ijcai.org, 5220–5226. https://doi.org/10.24963/ijcai.2022/725

[61]. Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q. 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7- 11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 9970–9985. https://aclanthology.org/2022.emnlp-main.677

[62]. 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 Com-putational Linguistics and Clinical Psychology*. Association for Computational Linguistics, Seattle, USA. https://aclanthology.org/2022.clpsych-1.0

[63]. Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. DepressionNet: Learning Multi-Modalities with User Post Summarization for Depression Detection on Social Media. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (Virtual Event, Canada) (SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 133–142. https://doi.org/10.1145/3404835.3462938

Table 1: BDI-II options for the symptom Loss of energy.

Option	Description
0	I have as much energy as ever
1	I have less energy than I used to have
2	I do not have enough energy to do very much anything
3	I do not have enough energy to do

Table 2: Main statistics and vocabulary comparison of the three sentences groups.

	Sadness	Pessimism	Sense of Failure	Loss of Pleasure	Guiltiness	Punishment	Self-dislike	Self-incrimination	Suicidal Ideas	Crying	Agitation
# Average annotator agreement (%)	81.06	75.43	85.73	79.10	93.70	91.63	93.53	83.62	88.98	87.05	78.89
# Sentences Positive (P)	154	72	62	140	27	18	60	10	44	34	26
# Sentences Negative (N)	490	202	237	599	290	269	445	108	186	244	135
# Sentences Control (C)	4900	2020	2370	5990	2900	2690	4450	1080	1860	2440	1350
Jaccard's Index (P vs N)	0.21	0.21	0.18	0.19	0.15	0.09	0.15	0.13	0.21	0.14	0.13
Jaccard's Index (P vs C)	0.10	0.12	0.10	0.11	0.07	0.04	0.06	0.07	0.11	0.08	0.09
Jaccard's Index (N vs C)	0.15	0.16	0.15	0.17	0.14	0.14	0.15	0.16	0.16	0.15	0.17
Jaccard's Index (C vs C)	0.25	0.21	0.20	0.27	0.21	0.20	0.24	0.18	0.19	0.21	0.18
KL(Positive||Negative)	1.01	1.40	1.54	0.91	1.50	1.69	1.41	1.95	1.20	1.79	1.98
KL(Positive||Control)	1.42	1.67	1.92	1.20	1.94	2.29	1.87	2.36	1.67	2.12	2.37
KL(Negative||Control)	1.24	1.49	1.57	0.96	1.51	1.51	1.28	1.56	1.40	1.37	1.69
KL(Control||Control)	0.74	1.20	1.12	0.63	1.00	1.04	0.78	1.55	1.24	1.09	1.45

	Social issues	Indecision	Worthlesness	Low energy	Sleep issues	Irritability	Appetite issues	Concentration	Fatigue	Low libido
# Average annotator agreement (%)	85.55	76.67	90.48	81.18	87.10	86.45	85.19	86.29	85.53	80.56
# Sentences Positive (P)	25	22	28	16	9	55	8	13	27	3
# Sentences Negative (N)	118	58	119	108	22	220	19	128	102	21
# Sentences Control (C)	1180	580	1190	1080	220	2200	190	1280	1020	210
Jaccard's Index (P vs N)	0.14	0.19	0.15	0.14	0.15	0.16	0.15	0.12	0.16	0.11
Jaccard's Index (P vs C)	0.11	0.11	0.09	0.07	0.09	0.09	0.09	0.07	0.09	0.05
Jaccard's Index (N vs C)	0.17	0.15	0.14	0.14	0.13	0.16	0.13	0.15	0.16	0.14
Jaccard's Index (C vs C)	0.18	0.15	0.18	0.17	0.12	0.20	0.12	0.18	0.18	0.13

	Social issues	Indecision	Worthlesness	Low energy	Sleep issues	Irritability	Appetite issues	Concentration	Fatigue	Low libido
KL(Positive||Negative)	1.62	1.64	1.92	1.96	2.54	1.24	2.45	1.97	1.61	2.61
KL(Positive||Control)	1.94	2.24	2.30	2.53	2.95	1.95	2.92	2.50	2.21	3.50
KL(Negative||Control)	1.45	1.89	2.00	1.85	2.60	1.41	2.61	1.45	1.68	2.62
KL(Control||Control)	1.50	1.99	1.43	1.60	2.65	1.07	2.71	1.44	1.58	2.59

Table 3: Examples of paraphrased sentences for the symptom *Sleep issues*.

(Relevance, Severity)	Sentence
(0,1)	I'm lying in my bed, and I'm still feeling it.
(0,2)	You might be having trouble sleeping or anything
(0,3)	If it persists, consult a sleep expert
(1,1)	I have sleeping issues, that's why I miss school
(1,2)	Even when I'm exhausted, I can't sleep
(1,3)	I just have energy to eat and sleep

Table 4: Symptom detection results of our classification models on BDI-Sen.

Method	AUC	P	R	F1
TF-IDF+LR	0.87	0.61	0.85	0.71
LIWC+SVM	0.83	0.49	0.83	0.62
MBERT	**0.95**	**0.74**	0.96	**0.83**
BERT	0.93	0.63	0.98	0.77
BERT-mini	0.90	0.57	0.94	0.70
T5	0.94	0.65	**0.98**	0.78

Table 5: Generalization ability results of our models with other mental diseases. Precision of the proposed sentence classification models when confronted with the positive sentences for the different disorders from the PsySym dataset [61]. Second row displays the number of test sentences from that disease. In the third row, c.s. (common symptoms) refers to the number of symptoms that are in common between the disease and BDI-Sen.

Method	Depression (1433 sent.) (14 c.s.)	Anxiety (2822 sent.) (19 c.s.)	ADHD (528 sent.) (4 c.s)	Bipolar Disorder (1131 sent.) (14 c.s)	OCD (449 sent.) (2 c.s)	PTSD (1284 sent.) (5 c.s)	Eating Disorder (907 sent.) (4 c.s)	Method Average
TF-IDF+LR	0.60	0.46	0.30	0.55	0.43	0.43	0.33	0.44
LIWC+SVM	0.69	0.73	0.46	0.53	0.64	0.62	0.49	0.59
MBERT	0.88	0.80	0.61	0.82	0.83	0.82	0.48	0.75
BERT	0.89	0.80	0.71	0.84	0.79	0.82	0.58	0.78
BERT-mini	0.85	0.84	0.74	0.79	0.62	0.69	0.58	0.73
T5	0.93	0.87	0.72	0.87	0.87	0.82	0.58	**0.81**
Disease Average	**0.81**	0.75	0.59	0.73	0.70	0.70	0.51	0.68

Table 6: Symptom severity classification results of our methods considering all the BDI-II severity levels and control sentences.

Method		Severity 0			Severity 1			Severity 2			Severity 3			Control		
	Micro F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TF-IDF+LR	0.79	0.15	0.30	0.20	0.55	0.27	0.36	0.21	0.23	0.22	0.41	0.37	0.39	0.87	0.96	0.92
LIWC+SVM	0.76	0.00	0.00	0.00	0.80	0.24	0.37	0.00	0.00	0.00	0.03	1.00	0.06	0.83	0.95	0.89
MBERT	0.80	0.35	0.08	0.13	0.52	0.37	0.43	0.32	0.22	0.26	0.44	0.39	0.42	0.86	0.99	0.92
BERT	0.77	0.24	0.01	0.10	0.56	0.32	0.40	0.34	0.19	0.24	0.25	0.20	0.22	0.82	0.99	0.90
BERT-mini	0.70	0.05	0.50	0.09	0.89	0.25	0.39	0.03	0.33	0.05	0.00	0.00	0.00	0.83	0.98	0.90
T5	0.66	0.12	0.05	0.07	0.09	0.09	0.09	0.06	0.06	0.06	0.09	0.07	0.08	0.76	0.83	0.80

Table 7: Symptom severity classification results considering grouping the BDI-II severity level 0 and control sentences.

Method		Severity 0 + Control			Severity 1			Severity 2			Severity 3		
	Micro F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TF-IDF+LR	0.83	0.91	0.96	0.93	0.45	0.32	0.37	0.26	0.22	0.24	0.38	0.35	0.37
LIWC+SVM	0.84	0.94	0.93	0.93	0.54	0.31	0.40	0.00	0.00	0.00	0.00	0.00	0.00
MBERT	0.86	0.93	0.98	0.95	0.54	0.40	0.46	0.32	0.20	0.25	0.41	0.40	0.41
BERT	0.85	0.92	0.99	0.95	0.58	0.34	0.43	0.18	0.17	0.17	0.35	0.41	0.38
T5	0.75	0.85	0.88	0.86	0.15	0.16	0.15	0.00	0.00	0.00	0.20	0.14	0.16

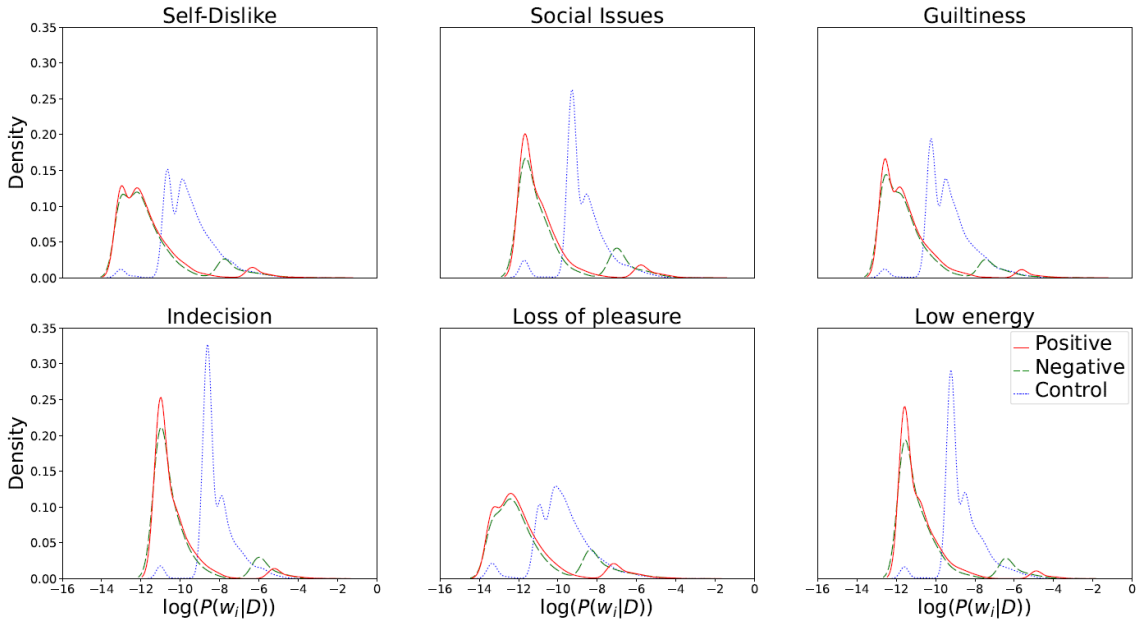

Figure 1: Density plot comparing word distributions of the three sentence groups for different symptoms.

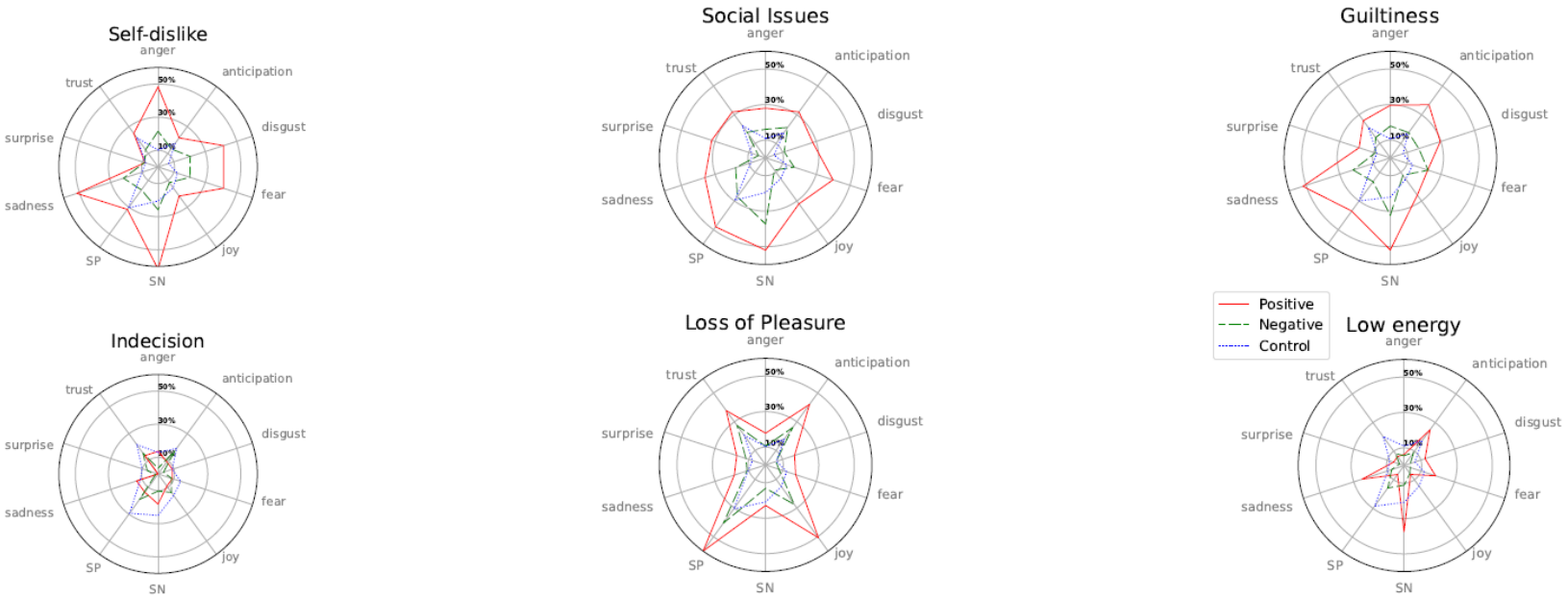

Figure 2: Radar plots illustrating the percentage of sentences that contain a word associated to the Plutchik emotions for each group (positive, negative, control).

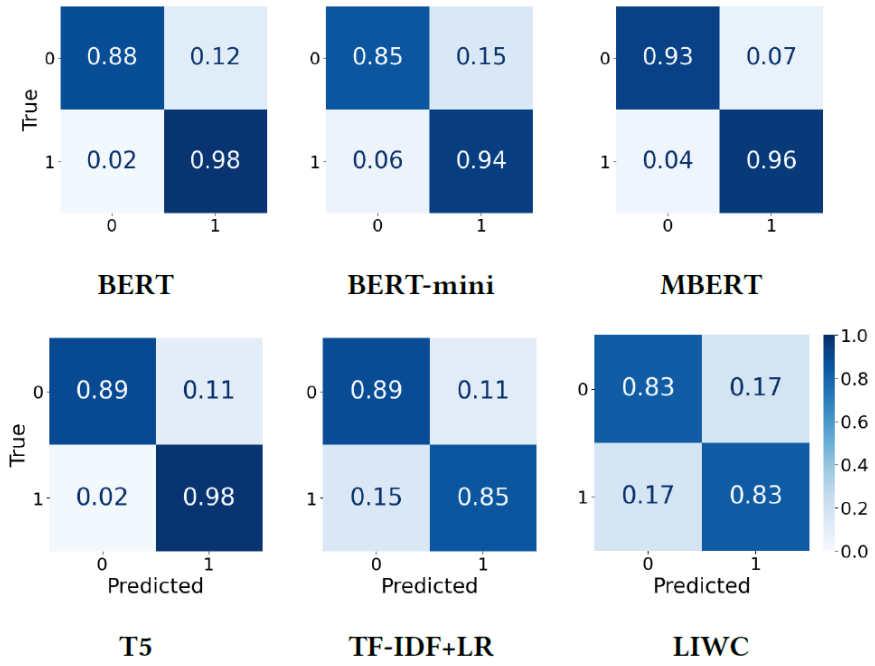

Figure 3: Confusion matrices showing the predictions accu-racy of our symptom detection methods.

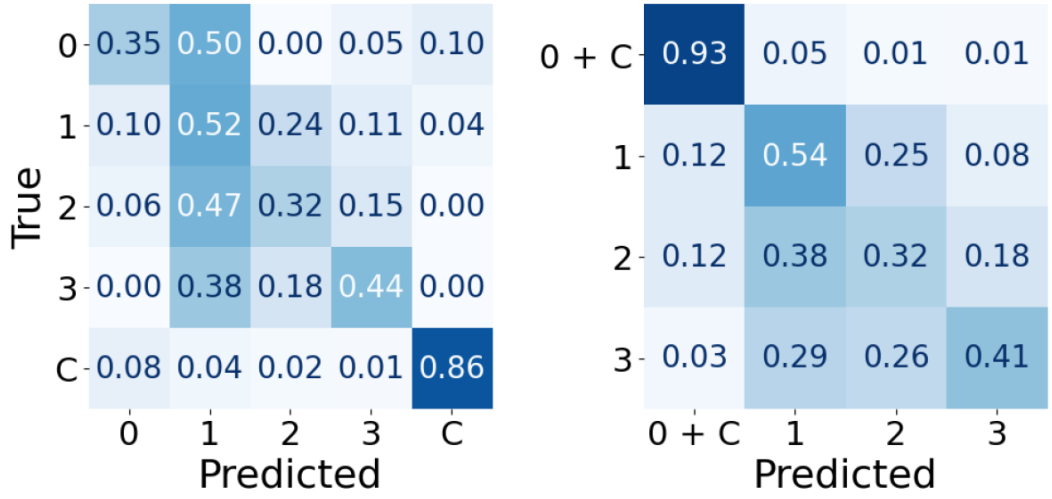

Figure 4: Confusion matrices of our best method (MBERT) classifying different symptom severity levels.