Artificial Intelligence In Medicine 132 (2022) 102380

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Automatic depression score estimation with word embedding models

Anxo Pérez *, Javier Parapar, Álvaro Barreiro

Information Retrieval Lab, CITIC, Universidade da Coruña, Campus de Elviña s/n, 15071 A Coruña, Spain

ARTICLE INFO

Keywords:
Depression prediction
Neural language models
Social media
Word embeddings

ABSTRACT

Depression is one of the most common mental health illnesses. The biggest obstacle lies in an efficient and early detection of the disorder. Self-report questionnaires are the instruments used by medical experts to elaborate a diagnosis. These questionnaires were designed by analyzing different depressive symptoms. However, factors such as social stigmas negatively affect the success of traditional methods. This paper presents a novel approach for automatically estimating the degree of depression in social media users. In this regard, we addressed the task *Measuring the Severity of the Signs of Depression* of eRisk 2020, an initiative in the CLEF Conference. We aimed to explore neural language models to exploit different aspects of the subject's writings depending on the symptom to capture. We devised two distinct methods based on the symptoms' sensitivity in terms of willingness on commenting about them publicly. The first exploits users' general language based on their publications. The second seeks more direct evidence from publications that specifically mention the symptoms concerns. Both methods automatically estimate the Beck Depression Inventory (BDI-II) total score. For evaluating our proposals, we used benchmark Reddit data for depression severity estimation. Our findings showed that approaches based on neural language models are a feasible alternative for estimating depression rating scales, even when small amounts of training data are available.

1. Introduction

Mental well-being is a fundamental component of World Health Organization's (WHO) definition of health [1]. Mental disorders are complex and can take many forms. These disorders directly affect how we think, feel, and act [2]. Good mental health enables people to develop their potential, cope with the everyday stresses of life, work productively, and lead a fulfilling life [3].

Depression is one of the most frequent and debilitating psychiatric disorders [4,5]. Depression alone affects more than 270 million people [6], being a primary reason for suicide (about 50% of all) [7] and one of the most significant causes of disability worldwide. It can affect anyone, regardless of age, gender, financial or social status [2]. The number of existing cases of these disorders maintains an upward trend over the years, and they already constitute a notable public health problem with vital consequences for society [8]. The significant impact on the prevalence of psychological disorders shown by the first massive studies of the effects of the Covid-19 pandemic is also noteworthy [9]. Despite its many harmful effects, there are validated and effective treatments, which can be boosted along with therapies and prevention programs [10]. Still, many individuals suffering from forms of depression remain untreated [11]. Early recognition is one of the keys to success in these treatments, minimizing their impact on public health and reducing cases' escalation. Several studies have addressed how early detection drastically reduces the disorder's negative impacts [12–14].

The standard way to measure the severity of depression by the clinicians relies on validated psychometric tests. These tests are known to have a satisfactory performance in the diagnosis at individual-level [15]; some relevant examples are the Patient Health Questionnaire 9 [16], the Hospital Anxiety and Depression Scale [17], the Center for Epidemiologic Studies Depression Scale [18] or the Hamilton Rating Scale for Depression [19]. One of the most reliable and trusted instruments is the Beck Depression Inventory-II (BDI-II) [20]. The BDI-II has a great deal of evidence of its performance [21]. It comprises 21 items related to major symptoms, covering aspects such as hopelessness or sadness, cognition symptoms such as punishment or guilty feelings, and physical indicators such as fatigue [22].

However, self and family reports are the most frequent ways to detect cases of depressive illness [23]. Population-level analysis via traditional methods is expensive. Phone surveys [24] are a common approach that often implies a crucial delay in obtaining practical results. Many health organizations also make these questionnaires available to users to fill them in by themselves. These online tests can even suggest visiting a medical professional. Per contra, when aiming for

* Corresponding author.
E-mail addresses: anxo.pvila@udc.es (A. Pérez), javierparapar@udc.es (J. Parapar), barreiro@udc.es (Á. Barreiro).

https://doi.org/10.1016/j.artmed.2022.102380
Received 11 June 2021; Received in revised form 8 June 2022; Accepted 18 August 2022
Available online 24 August 2022

a precise global diagnosis, conventional procedures may have distinct limitations. Revealing your mental conditions continues to be a highly complex process surrounded by stigmas, with a predominant lack of awareness in society. Moreover, the total score of the questionnaires is easily manipulable by the respondents. Existing studies analyzed how these responses can drastically vary depending on changing factors [25,26], and the final scores can be easily minimized or exaggerated. Bowling [26] studied the quality results' variations based on the administration of these tests. Social expectations, such as doing a test in front of a doctor, would change the results drastically compared to doing it in a friendly environment like your room.

As a new communication paradigm, people use social networks as a comfortable mean to share emotions, feelings, and thoughts. A significant part of mental health research requires the study of thoughts and behavior. Hence, social media is an excellent source to capture feelings that would often characterize disorders like clinical depression [27]. In these environments, users can preserve their anonymity while receiving the support and experience from others [28]. The emotional distance provided by these platforms is an incentive to interact and share their true feelings; individuals can express themselves in a leisurely manner, where depressed people gather information about their disease and discuss their symptoms. Based on this idea, researchers have analyzed user-generated content from sources like Reddit [29,30], Twitter [31,32] or Facebook [33] to develop predictive solutions for detecting mental illnesses.

In this work, we study the potential of neural language models for detecting depressive states based on social media content. Such a challenging task has been traditionally tackled as a binary classification problem, i.e., classifying depressed and non-depressed individuals. However, little effort has gone into finer-grained analysis, distinguishing between the different symptoms that characterize depression. The present study is a step towards that direction: we automatically fill in the BDI-II questionnaire by estimating all its 21 symptoms. In this regard, our primary research objective is to investigate whether or not word embeddings can capture signals of specific symptoms. For that, our proposed solutions build symptom-classifiers that exploit word embeddings as input features. We evaluate these proposals in benchmark collections using Reddit as data source.

As we work with the symptoms collected in the BDI-II, we observed that their sensitivity varies greatly. We define sensitivity as a property of the symptoms. The more sensitive a symptom is, it reduces the willingness of users to comment on it publicly. For instance, some symptoms (e.g. fatigue, changes in appetite) may be less intimate, making the users comment on them more easily. For these symptoms, it would be easier to search for direct mentions of the user concerns. On the contrary, others (e.g. loss of interest in sex, crying) are more sensitive, and users avoid explicitly talking about them. In this case, we would need to search for other signals in their language. This follows our intuition that there is great diversity in the degree to which users are more or less prone to talk about specific aspects of their lives explicitly. Our hypothesis indicates that signals of depressive symptoms differ depending on the symptom sensitivity. Our second objective exploits and analyzes this observation.

For the previous reasons, we designed two exploratory methods: The first one captures users' general language, searching for communication patterns by analyzing their publications at sentence-level. Contrarily, the second seeks only direct mentions of symptom-related concerns, focusing on specific responses instead of the rest of the publication. Attempting to validate this idea, we performed a symptom by symptom analysis, resulting in the development of a third hybrid solution that, depending on the symptom, will use one or another method.

The remainder of the paper is structured as follows: in Section 2 we present the actual approaches that exploit the use of clinical questionnaires and a summary of the work related to our line of research. Section 3 describes the collections that will be used along with the characteristics of the BDI-II, and we introduce our framework to estimate the presence of symptoms. The results of our methods are presented in Section 4 based on the experimental settings introduced at the beginning of that same section, and we also report a comparative symptom-by-symptom analysis. We discuss our conclusions and future studies in Section 5. Finally, the ethical considerations and practical implications of our approaches are discussed in Section 6.

2. Related work

Traditional studies on depression detection are based on different self-diagnosis questionnaires [16,18–20]. These questionnaires are aimed at an individual level, by establishing rating scales that associate the total score to the depression severity of the patient. There are also existing efforts towards studying and detecting depressed people at a population level [34]. Most of them have relied on sharing those inventories as surveys. One example is the Behavioral Risk Factor Surveillance System (BRFSS), a United States health survey administered via telephone. Its main objective is to estimate the rate of depression among adults in the US [24]. Nevertheless, these techniques face methodological challenges:

- Depression sufferers often manifest a lack of willpower and strength, discouraging factors from filling questionnaires honestly [35,36].
- Stigmas surrounding mental health diseases make it even more difficult to reveal your feelings and decide to seek help [37].
- These procedures imply notable financial requirements and active participation of subjects. Moreover, it is required considerable time to have enough people to make complete conclusions, involving long delays between data collection and actual findings [36].

There is a great deal of prior work on understanding the underlying causes of depression. Research on topics related to depressive conditions has been performed within the medicine or psycholinguistics fields [38–40]. All of them have tried to identify the presence of symptoms, causes and how to perform a precise diagnosis. In recent years, social media have become a pivotal source to provide data related to mental health disorders [29,41,42]. This amount of user-generated information allows researchers to analyze user language and behavior. Several studies showed how social network content is valuable for identifying depression and other mental health problems. In such a way, computational linguistic solutions have been applied to analyze mental disorders like suicidal ideation [43,44], schizophrenia [45], or eating and anxiety disorders [46,47].

Concerning to depression, De Choudhury et al. [32,48,49] made pioneer contributions to the field. Their studies extracted relevant features in the language of depressive individuals (i.e., higher self-attentional focus and negative emotions). In this regard, several researchers started to investigate a wide range of different features: linguistic, emotional expression, semantic, lexicon-based or social network properties, among others. For instance, the role of emotions in depressive-content has been significant in identifying the disorder on Twitter [50] and Reddit [30]. There were several efforts focusing on contact's network structure [32,51] and the relevance of personal statements [52] (i.e., information present in phrases with singular first pronoun) and the frequency of personal pronouns [53]. In this context, we are also aware of two recent proposals that incorporate questionnaire information to identify depression as a binary classification task [54,55].

For the last few years, most recent work involving the capture of semantic features is built on the use of neural language models. These models encode the word meanings as vectors, known as word embeddings. Words that appear in similar contexts will be closer in a multi-dimensional space. Traditional word embedding architectures,

including Word2Vec[56] and Glove[57], are designed as static models. Static models use a single global embedding to represent each word of the vocabulary. Recent advancements in the Natural Language Processing (NLP) field introduced context-aware approaches thanks to the use of pretrained language models (PLMs). PLMs come from modern transformers architectures, such as BERT-based models [58], GPT [59] or T5 [60]. One of the main advantages of these models is its ability to be fine-tuned into smaller datasets and obtain outstanding performance in downstream tasks [61,62].

Recent works in the literature adapted different PLMs for the clinical [63] and biomedical[64] domain. Researchers also leveraged the potential of transformer models in the mental healthcare field. Related to depression, most of the prior works based their methods by fine-tuning the original BERT with an additional classification head. For this purpose, they train these models as sequence classification task, giving them as input publications of depressed vs. non-depressed users. After the training process, the models are able to produce binary decisions at publication level [55,65,66].

Due to the recent popularity of this field, shared tasks emerged to promote the research on mental health detection from the perspective of NLP. The Computational Linguistics and Clinical Psychology (CLPsych[1]) [67] and the Early Risk Prediction on the Internet (eRisk[2]) are the evaluation forums that have become standard benchmarks for predictive methods. The CLPsych-2015 aimed to detect evidence from users having depression using Twitter data. Participant systems experimented with bag-of-words together with supervised standard classifiers, or developing rule-based approaches, topic models and traditional machine learning algorithms.

Additional challenges to help improve research on depression detection go beyond single binary classification (depressed vs not depressed) scenarios. The eRisk initiative proposes novel tasks in this new context.[3] For instance, detecting depression in early stages has great importance to avoid further consequences. For that reason, eRisk organized the first shared task on *early risk detection of depression* (ERD) in 2017. *ERD* task consisted of early detecting signs of depression by sequentially processing social media posts from Reddit users. In this sense, organizers include time-aware metrics to reward early alerts and penalize later ones. The second edition was organized in 2018. In 2018, participants included bag-of-words representation and the exploit of domain-specific vocabulary, and started to leverage word embeddings representations. Still, the best performance obtained so far did not make use of these kind of semantic representations. Burdisso et al. [72] improved all other participants using a text classification method, called the SS3 classifier. The main idea of SS3 is to value words in relation to categories. The authors built a dictionary of words associated to depression and non-depression categories. In the classification stage, the method uses this information to calculate if the score of the summarized words of a user is related to one of the categories.

In the last two editions, a new task came up aimed at estimating the level of depression, called *Measuring the Severity of the Signs of Depression*. Our work is directly related with this task. In 2019 edition, few approaches built solutions based on word embeddings. van Rijen et al. [73] proposed a ensemble of models based on word polarities, mutual information and semantic similarity. For the semantic features, they used word embeddings from GloVe to extract features for each symptom. Then, they compared these symptom embeddings with the user posts. Abed-Esfahani et al. [74] proposed a similar approach, but used a finetuned GPT-1 model to generate the vector representations. Again, the best results in this collection were recently published by Burdisso et al. [75]. The authors extended the SS3 classifier values mapping them to the BDI-II possible categories, obtaining a promising performance.

The next edition, more researchers adapted different PLMs to build their solutions for the 2020 collection. Most of them obtained their own datasets from Reddit. Then, they fine-tuned these models as sequence classification to produce the inference decisions [65,66]. However, three of the top four participants did not rely on the use of embeddings. These solutions were focused on the use of hand-crafted, emotional and psycholinguistic features [76–78]. We give a detailed description of each top performing method in Section 4.2. In this regard, our method differs from prior works in two main ways. (1) We do not rely on fine-tuning on specific crawled datasets, or base our approaches in complex classification processes. This point contributes to the explainability of our decisions. (2) Our methods are model-agnostic and can be easily transferred to other diseases, questionnaires and symptoms without additional effort. (3) Finally, our classifiers are not based in a specific set of hand-crafted features and resources, such as the exploit of depression or emotion lexicons.

Table 1
Options available for the item 10 'Pessimism in the future'.

Option value	Option sentence
0	I am not discouraged about my future
1	I feel more discouraged about my future than I used to
2	I do not expect things to work out for me
3	I feel the future is hopeless and will only get worse

3. Material and methods

3.1. Task: Measuring the Severity of the Signs of Depression

The aim of the task is to estimate the level of depression based on a thread of user submissions. The levels go accordingly to the BDI-II questionnaire scores. For each user, the participants were given its entire history of postings, and they had to predict the BDI-II responses for that same user. The predicted answers are based on the evidence found in the history of postings. Thus, for each user, the collection has two elements: (1) their real responses to the symptoms of the BDI-II (ground truth) and (2) their full writing history (WH).

BDI-II[4] is composed of 21 items that measure characteristic attitudes and symptoms of clinical depression [20,79]. Each item is related to a different symptom, containing four answer options accompanied by a sentence explaining its meaning. The options are rated from 0 to 3 according to the Likert scale [80]. Options scale in terms of severity, from the total absence of the symptom to a total identification. It takes approximately 10 to 15 min to complete it.

Table 1 shows one example of an item, in this case, the item is related to pessimism about the future. The accumulating result of all the 21 items is associated with a scale of depression manifestation: minimal depression (0–9), mild depression (10–18), moderate depression (19–29), and severe depression (30–63).

3.2. Datasets

For experimental purposes, this study uses the datasets provided by eRisk 2019 and 2020 editions [70,71]. eRisk organizers contacted with users from the Reddit[5] platform to fill in the BDI-II. With their agreement, they extracted their complete WH. Both collections contain posts from English-speaking users, with 20 users in 2019 and 70 in

[1] https://clpsych.org/
[2] https://erisk.irlab.org/
[3] The tasks related to each edition are described in detail in the CLEF's eRisk overviews [68–71].
[4] The BDI-II can be consulted at https://early.irlab.org/2019/ (Task 3).
[5] Reddit is an open-source platform where members can submit content such as links, text or images (https://www.reddit.com/).

Table 2
Statistics of eRisk 2019 and 2020 collections for the depression estimation task.

Edition		2019	2020
Users		20	70
Number of posts			
	Total	10397	33638
	Min per user	25	16
	Max per user	1257	1258
	Avg per user	519.8	480.5
	Words/Post avg	46.9	42.87
Number of titles			
	Total	1123	3865
	Min per user	0	0
	Max per user	512	526
	Avg per user	56.1	55.2
	Words/Title avg	10.8	9.28

2020. For each user, the dataset provides its real responses to the questionnaire along with its complete history of postings. The collections can be obtained on request from the eRisk organizers.

In our proposals, we used the 2019 corpora as training data and the 2020 for testing. The collection provides an XML file for each subject, which contains all user posts ordered by chronological order. Each post consists of the following elements: id, a unique identifier for each Reddit user; and writing, which constitutes a publication made on the platform. Simultaneously, each writing contains the fields (title, date, info and text). The field title represents the Reddit thread title; date indicates the exact time of the publication, info designates the platform used (just Reddit), and text represents the user's publication text. Table 2 provides general descriptive statistics of the two collections.

3.3. The symptom-classifiers framework

Since we are introducing a new classification framework, instead of going straight to the specific methods, we have decided to include the general idea first and then, the three different solutions that came up based on it. Our proposed approach was intended to be used as a general framework for depression estimation, considering it is (1) model-agnostic and (2) flexible enough to be implemented in several manners depending on the symptom or even questionnaire. Thus, the next subsections shows our framework's main aspects along with intuitive examples highlighting the process.

We addressed our proposal as a classification task. Instead of relying on a unique classifier, we build 21 symptom-classifiers, each one corresponding to a different symptom present in the BDI-II. The symptom-classifiers are designed as a four-class problem, and each class is associated with one of the possible answer options (0–3). For instance, the options shown in Table 1 *Pessimism in the future* correspond to the four possible classes of that symptom. With the use of these classifiers, we can infer the answered option for each one of BDI-II symptoms. Finally, to predict the user's total BDI-II score, we simply aggregate the decisions of the 21 classifiers. Keeping this in mind, our proposal relies on three critical choices: (*i*) data filtering (*ii*) users and options feature extraction (*iii*) the use of symptom-classifiers:

3.3.1. Data selection strategy

(*i*) One of the major challenges in research on depression from social media is the creation of collections with enough reliable cases. Labels are typically assigned at user level (in our case, the labels are the responses of the BDI-II of each user). However, our approach extracts the features at a sentence-level instead of considering the whole user WH. Providing labels at the sentence level would be a high-cost process. For this reason, we extend the labels of the users' symptoms (0−3) to all their WH. This step generates noise for many samples, considering that the users write many publications out of context. As a consequence, if we had an user $U1$ who answered the option 3 for the symptom *'Self-dislike'* and another user $U2$ who answered the option 0, and they both have a publication with the content *'Went to the cinema today'*, the publications would have opposite labels despite being the same text. To reduce this issue, we applied a simple but effective filtering method in both the training and inference phases. The data selection strategy allows us to gather candidate posts from the entire set of publications that have higher relevance to the symptom.

As publications on Reddit have variable content length, we first split the whole WH of the users into smaller units (sentence-level). As our approach is based on semantic search, we find it very convenient to work with sentences rather than the whole publication. Working at smaller units improves the quality of the semantic representations.

To filter only relevant sentences from the whole set of the sentence labels, we use a measure of textual similarity: Okapi Best Matching 25 (BM25) [81]. In the training phase, we use the 2019 corpus to gather all the sentences from the users that answered each option. Using BM25, we can retrieve the top k candidate sentences that better characterize the option. For that, we use as query the statement containing the description of that same option. For instance, if we want to get the features of the option 0 in symptom 14 *Loss of energy*, the query would be: *I have much energy as ever*, and we would only select the top k sentences from the users that answered that option. This allows us to discard many writings that would not add any value for feature extraction.

In the case of a test user, we apply a similar process. However, in the inference phase we do not know the test user's label. Therefore, in this case, we use the queries to obtain relevant sentences of all four options. For example, for the symptom 14, the query would be: (I have much energy as ever I have less energy than I used to have I do not have enough energy to do very much I do not have enough energy to do anything). The filter will select only the top k sentences from those users that match the query. These candidate sentences will be the ones used to extract the features.

3.3.2. Extraction of Users and Options features

(*ii*) Our main intention is to investigate the capture of the semantics of the BDI-II symptom options, $o \in \{0, 1, 2, 3\}$. As sentence embeddings have been very successful in semantic similarity tasks, we decided to create an vector embedding to capture these semantics. This resulted in a total of 84 option vectors (21 symptoms with four options each). In all our experiments, we extracted the features using word2vec models. This process is carried out in the training phase, where we retrieve all the filtered sentences from the users that replied to each option. To obtain the features, we assume $e(s)$ as a function that takes a sentence s and maps it into its vector representation. Following this, the computation of the final features vector of a symptom i and an option j, defined by $\vec{o}_j^i$, follows the equation:

$$\vec{o}_j^i = \frac{1}{|S_j|} \sum_{s_j \in S_j} e(s) \tag{1}$$

where $|S_j|$ denotes the total number of filtered sentences for the option. $\vec{o}_j^i$ is then calculated by averaging the sentences embeddings of the users that answered the specific option j. In inference, we also summarized the semantic of a test user into a single vector embedding. For that, we average all the features from its filtered sentences.

3.3.3. Use of symptom-classifiers

(ii) Every symptom-classifier has two phases: training and inference. The training process consists in generating a feature vector representing each possible option. As a result, we train 21 symptom-classifiers by calculating the option vectors, where $C_n = \{\vec{o}_0^n, \vec{o}_1^n, \vec{o}_2^n, \vec{o}_3^n\}$, with $1 \leq n \leq 21$. We illustrate the options vector generations process in the Fig. 1. It is exemplified for one of the symptom, 14 : (*Loss of energy*). The process for the rest of symptom's options remains the same.

First, we retrieve all the training users that answered each option: (0. *I have much energy as ever,* 1. *I have less energy than I used to have,*

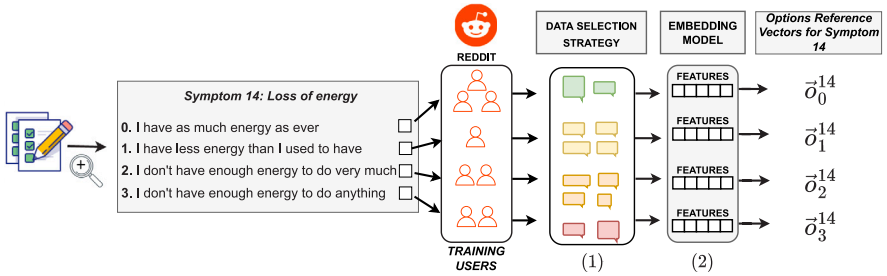

Fig. 1. Training overview showing the extraction of option feature vectors of symptom 14: *Loss of energy*.

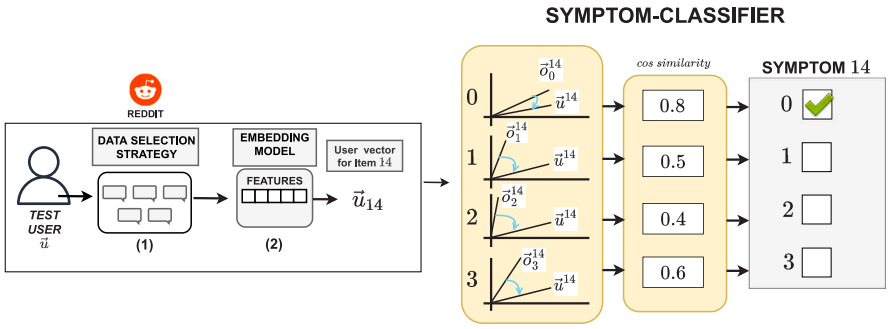

Fig. 2. User feature vector extraction u_{14} for an symptom 14 and the posterior classification decision.

2. *I do not have enough energy to do very much*, 3. *I do not have enough energy to do anything*). From their WH, we apply the data selection strategy to filter the relevant sentences most related to the symptom. At this moment, we already have k candidate sentences $\{s_{j1}, s_{j2}, \ldots, s_{jk}\}$ for each option o_j. Finally, we extract the features from these filtered sentences to compute the vectors o_j^{14} for the symptom 14.

In inference, we classify the BDI-II answers of a test user after measuring the similarity with the previously calculated option vectors. We illustrate the classification process in Fig. 2. It is exemplified for only one test user and the symptom 14. First, we filter its corresponding WH following the data selection strategy described in Section 3.3.1.[6] Then, we apply the $e(s)$ function to extract features at sentence-level, and we compute the user vector $\vec{u^{14}}$ for the symptom 14 by averaging all the sentences features. We produce a different user vector for each symptom, containing only the sentences more related to it. Finally, to obtain the predictions, we simply compare the similarity of $\vec{u^{14}}$ with the option vectors $\vec{o}^{14}_{[0...3]}$. We use cosine similarity to obtain a result for each option. In the last step, we classify the given test user representation u with the option that has the highest cosine similarity with her/him.

3.4. Symptom-classifiers variants

This subsection describes the three different variants we used to build the symptom-classifiers. All of them are based on the classification framework described above. As a result, these classifiers are able to generate decisions for each symptom.

3.4.1. General symptom-classifiers

We call the first approach General symptoms-classifiers, which objective is to capture the general language of the users. In this method, we look for communication patterns that may indicate the presence of the symptoms. The procedure for obtaining the representations of options and users is analogous to the pipeline described above. We simply obtain all the k sentences that satisfy the filtering process and extract all its features in training. In the inference phase, the inputs for the symptom-classifiers are the test users' vectors. To compute the final BDI score of the users, we aggregate the decisions of all these classifiers.

3.4.2. Direct answers symptom-classifiers

We called this second approach Direct answers classifiers. Contrarily to the general symptom-classifiers, here we seek only direct mentions or concerns associated with the symptoms. As a result, this seeking process will only extract the specific span of the sentences containing explicit answers about how the user feels about the symptom. For this step, we have used a Question Answering (QA) model.

QA is one of the NLP tasks that has significantly disrupted. QA systems are based on triplets (P, Q, A), which can generate an answer A from a passage P and a question Q. The idea is to get potential answers for each symptom directly from the user's writings. Our model obtains the answer A by using the user's publications as a passage P and the BDI item as a question Q. Behind the premise that particular items are more likely to be commented on in a more direct way than others, we can experiment if capturing only direct answers to symptoms can improve prediction performance. We give the details of the training process of the QA model in the experimental settings (Section 4.1). We constructed our QA model as an application of BERT [58], a neural language model based on the transformers architecture.

At this point, it is necessary to mention that the BDI-II does not provide a question per symptom. Instead, it simply presents the possible options, and the respondent have to choose the option with which he/she most strongly identifies. We, therefore, had to manually construct the questions for each item as questions to our QA model. For this, we formulated a simple question containing keywords related to the symptom. Table 3 shows some of the questions we constructed for the items in the left column, as well as some extracted answers from those questions in the training collection.[7]

This proposal uses the same configuration as the general symptom-classifiers for the feature extraction. However, the difference is that we now apply a different technique to select the sentences for extracting

[6] If no posts are found after filtering, we assume the answer estimated is 0 as there are no information traces from the user.

[7] The answers are paraphrased in accordance to test collections' license.

Table 3
Example of four questions that we used for BDI symptom along with an extracted answer from these same questions.

BDI Item	Model question	Extracted answer
9. Crying	Do you usually cry?	I have grown used to crying over anything that occurs
11. Social withdrawal	Have you lost interest in people or social life?	I have been increasingly estranged from most of my peers
13. Worthlessness	Do you feel worthless or insignificant?	Always feeling guilty and unworthy dude
14. Tiredness	Are you usually tired or fatigued?	It is difficult to be creative or even get out of bed these days

the features. For this reason, we use the QA model to search for only explicit answers to the symptoms. If the QA model does not output any answer for a selected sentence, we discard it. Consequently, the embedding model will only extract features from short answers instead of the whole candidate sentences. With this step, we drastically reduce the total number of sentences represented as vectors. The rest of the classification process, both in training and inference, is analogous to the general architecture. We construct the option vectors for each symptom-classifier in training and the test user vectors for inference. Finally, the computation of the final BDI score is the aggregation the decisions of the direct symptom-classifiers.

3.4.3. Mixed symptom-classifiers

This final approach, called Mixed classifier, is a hybrid solution leveraging the two previous methods. The mixed classifier does not correspond to any new method. Instead, it uses the general or direct symptom-classifiers depending on the symptom. As these two have different objectives to capture (general vs direct language), we conducted a symptom-by-symptom analysis to determine which model performed better for each symptom. With that goal, we performed leave-one-out cross-validation in our 20 training users, using Average Hit Rate (AHR) as objective metric. AHR computes the ratio of times a method correctly estimated an option averaged for all the users (evaluation metrics are explained in Section 4). The complete analysis can be seen at Section 4.3. The best performing classifiers per symptoms for the training users are presented below:

- The direct classifier obtained better results in the following symptoms: *'Pessimism for the future'*, *'Sense of being a failure'*, *'Social withdrawal'*, *'Guilty feelings'*, *'Sense of punishment'*, *'Irritability'*, *'Changes in appetite'* and *'Tiredness or fatigue'*.
- The general classifier obtained better results in: *'General sadness'*, *'Lack of satisfaction'*, *'Self-dislike'*, *'Crying'*, *'General agitation'*, *'Indecisiveness'*, *'Loss of energy'*, *'Concentration difficulty'* and *'Loss of interest in sex'*.
- Both classifiers had the same AHR values in: *'Sleep disturbance'*, *'Self-incrimination'*, *'Worthlessness'*, *'Suicidal ideation'*.

Based on these results, in the test set, the mixed classifier decides to use the best of the two previous classifiers that best adjusts to each symptom. In the case of the symptoms which both classifiers had the same AHR result, we considered to use the general classifier as it showed more consistency in the rest of the metrics.

4. Analysis and results

This section covers the experimental analysis of our symptom-classifiers on the depression estimation task. The experiments were conducted on the eRisk 2020 Task 2: *Measuring the Severity of the Signs of Depression*. To evaluate our results, we used the 2019 collection as training data and the 2020 collection as test. It is organized into three main subsections. The Section 4.1 describes the experimental configuration of the proposed approaches and we introduce the four metrics used to evaluate the performance of the classifiers. Section 4.2 shows the obtained results and compare them to the state-of-the-art methods in the shared task. Finally, Section 4.3 presents an symptom-analysis of the obtained results.

4.1. Experimental configuration

We experimented with different embedding models to extract the features from the sentences. Two of them are based on word2vec. The first one is FastText, an incremental word2vec technique that also encodes the morphology of words [82]. The second is sense2vec, a word2vec variant that uses supervised disambiguation to generate unique embeddings for each word sense [83]. Furthermore, sense2vec was trained on Reddit comments from 2015, making it even more suitable for our collections. Additionally, we also experimented using the pretrained BERT model to get these embeddings. In this case, we feed the sentences to $\text{BERT}_{\text{BASE}}$ and take the first vector of the hidden state (CLS token). The CLS was used as embedding of the sentence. The performance of BERT was slightly worse than word2vec models. For this reason, we leave for future work a thorough comparison between the impact of the word2vec and more sophisticated BERT embeddings on the performance of our framework.

Different settings to develop the classification framework were also investigated. Table 4 shows the hyperparameters that produced better results. We applied leave-one-out cross-validation using only the training set. The following is a summary of them: (1) stopwords, we considered removing stopwords when gathering the set of posts. (2) Apply the data selection strategy or instead consider all the publications to generate the options representations. In the training process, we found that not using any filter improves the direct classifier. In contrast, the filtering strategy improves the performance of the general classifier. (3) Apply the selection strategy in the user representations. In this case, we are processing a lower amount of sentences (only one user). Thus, not filtering the sentences of the test user obtained better results.

Furthermore, (4) and (5) show the number of the best top k sentences filtered with BM25. We tuned the k value from 100 to all the possible sentences matched in increments of 100. The training user with more k sentences is 2608. Hence, using 2608 corresponds to retrieving all the sentences with BM25 (BM25_ALL), which was the best value obtained in our experiments. As the filtering method only improved for the general classifier, the rest of the values were not used. Finally, (6) we tuned the text units considered from the set of publications selected. In the case of the general classifier, using sentences for embedding yielded better results. In the case of the direct classifier, using full publications as input to the QA system produced better results than dividing it first into individual sentences. We assume that the QA model works better with the whole publication as it has more context available to look for direct answers. Furthermore, the direct classifier discards sentences when the QA system does not find an answer in them. We also believe that this is the reason of why a filtering strategy is not needed in the direct classifier.

As mentioned in Section 3, the direct-answers classifier employs a QA model to detect potential answers and concerns about the symptoms in the set of publications. For that, we trained a retrospective reader designed by Zhang et al. [84]. The collection used to train this model was the Stanford Question Answering Dataset (SQuAD). SQuAD is a reading comprehension dataset consisting of crowdsourced question/answer pairs on a set of Wikipedia articles [85]. The model ranks the second position in the SQuAD competition in 2020. Table 5 shows the parameters we applied to train the model. We maintained the same parameters reported by the original paper, except for batch size, due to efficiency issues. Moreover, we limited the maximum answer length to 30 as we wanted to capture concrete and short responses.

Table 4
Tuned hyperparameters experimented in the training process of our classifiers.

Hyperparameter	General classifier	Direct classifier
1) Stopwords	Remove	Remove
2) Data selection on options representations	BM25	Not filtered
3) Data selection on users representations	Not filtered	Not filtered
4) Best k on options representations	BM25_ALL (2608)	–
5) Best k on users representations	–	–
6) Text unit	Sentence level	Publication level

Table 5
Parameter values of the question answering model.

Parameter	
Learning rate	$2^{e^{-5}}$
Training steps	260000
Maximum context length	512
Batch size	4
Warm-up steps	814
Maximum answer length	30

For evaluating the participant methods, organizers considered four different metrics. We used the same metrics for all our experiments. These metrics evaluate the quality of the estimated answers to the BDI-II [71]. Next, we briefly describe them:

- Average Hit Rate (AHR): Hit rate (HR) is a measure that computes the ratio of items the system has estimated the same answer option as the user. If the HR for a user is 10/21, then for 10 of the 21 items, the system estimated for the test user the same option as the golden user response.
- Average Closeness Rate (ACR): Closeness Rate (CR) calculates the absolute difference between the actual answer and the estimated answer, and subsequently, an effectiveness score is applied as follows: CR = (*mad* - *ad*), where *mad* represents the maximum absolute difference, and *ad* the actual difference. If a real user has answered '0' for an item, this metric will penalize a system that has estimated '3' more than one that has estimated '1', as the latter is closer to getting it right.
- Average DODL (ADODL): Corresponds to Difference Between Overall Depression Levels (DODL). It calculates the absolute difference (ad_overall) between the user's actual BDI-II score and the system estimated BDI-II score. DODL is normalized to [0,1]. The formula is DODL = ((63 - ad_overall)/ 63), where 63 is the maximum absolute difference that can be obtained when estimating the total score.
- Depression Category Hit Rate (DCHR): The BDI-II associates 4 specific depressive disorders categories depending on the score obtained in it:
  - minimal depression (0–9)
  - mild depression (10–18)
  - moderate depression (19–29)
  - severe depression (30–63)

  DCHR computes the total number of cases where a system has estimated the user's real category based on the BDI-II score.

4.2. Baselines and results comparison

Table 6 presents the performance of the selected baselines, the state-of-the-art approaches for each metric, and the methods we presented. The first three rows (upper block) include baselines proposed by the organizers in order to gain some perspective [71]. The first and second rows are all 0s and all 1s, consisting of filling in the same option (0 or 1) for all the items. These methods represent good baselines for the metrics that consider the closeness of the predictions (ACR and ADODL). Visualizing their results, we can see how the options are distributed between 0's and 1's. Finally, the third row shows the performance of an algorithm that fills in the options randomly. As we included the participants' approaches with better results for each metric in 2020, we will summarily explain their approaches:

The BioInfo@UAVR [77] method used an external dataset to train a rule-based approach. The authors captured different psycholinguistic patterns and behavioral features to model each rules. More specifically, they represent the user's writing history as a vector composed of several depression-specific features. Some examples are guilt and cry, sleep, anxiety, irritation or depression. For example, authors measure the depression category using the average polarity of the writings, the use of self-related words (I, myself, mine) and the mention of specific words related to mental disorders and anti-depressants. For each category, they calculated a user score using the frequency of the categories for that user with respect to the total number of occurrences over the training dataset. These scores were then normalized to the interval [0–3] of the BDI-II.

ILab [65] obtained the best results in ACR. Their method used BERT-based classifiers trained explicitly for the task. Similar to our approaches, they addressed the problem as a multi-class labeling task. Authors fine-tuned the base language model with a head for multi-classification for every question. They also balanced the weights of the classes due to the sparsity of training data. In inference, for a given user, they predict the answer obtaining the softmax prediction for every publication. The class with the highest accumulated value is the estimated answer by the system. They experimented with different neural language models, and XLM-RoBERTa achieved the greatest performance.

PRHLT-UPV obtained the best results in DCHR. Sabina et al. [78] submitted three different runs. Their first two methods used linguistic and emotion features like LIWC, representing the users as continuous vectors. To obtain user levels representations, authors averaged the values of these vectors for each user post. They also included the standard variation of the features trying to capture the evolution of the language. Finally, the third run leveraged pre-trained language models to obtain semantic representations of the users social media posts. In this regard, they extracted the features at sentence-level using Universal Sentence Encoder (USE) [86]. They experimented these three solutions with traditional machine learning algorithms. Their best results were obtained by using the linguistic and emotional features with a SVM classifier.

RELAI obtained the best results in ADODL. Maupome et al. [76] addressed the problem as authorship attribution, which relies on decision models to predict the probability of a pair of documents written by the same user. Authors exploit the decision models in two variants: the first attempt to relate users to each other (user-based), and the second to relate a user with the text contained in the BDI-II itself (answer-based). These models included three approaches: Topic Models, Contextualizer and a Stylometry-based solution. Their best model was topic modeling using Linear Discrimination Analysis (LDA). The user-based variant creates topic vectors for users and then computes the distance between them, and the answer-based between the topic representation of answers. Their LDA model was trained on an external dataset, finding better results when requiring the model to calculate 30 topics.

Our solutions include the general, direct-answers and mixed classifiers with the sense2vec and FastText embedding model variants. Next, we will briefly explain and compare the results of our methods with the baselines considered.

Table 6
Results of our classifiers along with the baselines and the best runs of eRisk 2020. S2V and FT stand for our two embedding models, sense2vec and FastText, respectively. The numbers in parenthesis after the score corresponds to the position it would have obtained if our methods had participated in the task. Bold values highlight the best value obtained in the metric.

Run	AHR (%)	ACR (%)	ADODL (%)	DCHR (%)
all 0s	36.26	64.22	64.22	14.29
all 1s	29.18	**73.38**	81.95	25.71
random	23.94	58.44	75.22	26.53
BioInfo@UAVR [77]	38.30	69.21	76.01	30.00
ILab [65]	37.07	69.41	81.70	27.14
Relai [76]	36.39	68.32	**83.15**	34.29
Prhlt-Upv_svm [78]	34.56	67.44	80.63	35.71
General_Classifier_S2V	$38.23_{(2)}$	$69.23_{(2)}$	$81.56_{(5)}$	$\mathbf{44.29}_{(1)}$
General_Classifier_FT	$38.57_{(1)}$	$69.16_{(3)}$	$80.54_{(7)}$	$38.57_{(1)}$
Direct_Classifier_S2V	$36.94_{(4)}$	$69.39_{(2)}$	$81.41_{(5)}$	$28.57_{(9)}$
Direct_Classifier_FT	$35.64_{(9)}$	$67.89_{(9)}$	$80.91_{(6)}$	$28.57_{(9)}$
Mixed_S2V	$\mathbf{38.97}_{(1)}$	$70.10_{(1)}$	$82.61_{(3)}$	$37.14_{(1)}$
Mixed_FT	$38.51_{(1)}$	$70.00_{(1)}$	$81.80_{(3)}$	$30.00_{(7)}$

The *all 0's* method obtains noticeably good results in AHR compared with its performance in the other metrics. This evidences that option 0 was the most popular option for the test collection. Moreover, as expected, *all 1's* method achieves favorable result for metrics that consider absolute differences (ACR and ADODL). As the collection is balanced between depressive and non-depressive users, estimating 1's will always be in the middle range, representing a good baseline for these metrics.

From this table, it is possible to observe competitive results with respect to the top participant methods. The sense2vec variant was always slightly superior than the FastText variant. In the vast majority of metrics, both our general and mixed classifiers have improved all systems. Furthermore, while most participants struggled to perform well on all the four metrics, our methods showed a high level of consistency in performance.

Our mixed solution using sense2vec (*MIXED_S2V*) outperformed all the participants in three of the four metrics evaluated (AHR, ACR, DCHR), and ranked third in ADODL. We note that, for the DCHR metric, we increased performance by about 29% compared to the best participant result. With the FastText model, (*MIXED_FT*), we rank first in AHR and ACR, while being third in ACR. Our mixed solutions always improved the performance of our individual methods.

In addition, the general classifiers also obtained better results than any participant overall. Using the sense2vec model, *General_Classifier_S2V*, we are close to rank first in AHR and ACR. In DCHR, we outperform the best participant by a wide margin (nearly 30% improvement to the best solution). With the FastText model, *General_Classifier_FT*, we are first in AHR and DCHR and third in ACR. However, this classifier has been found to perform worse in ADODL.

Finally, our direct-answers classifiers are the ones that performed more modest. Using the sense2vec model, we are in the top 5 of three metrics (AHR, ACR, and ADODL). In DCHR is where we obtained the worst position. This drop in the performance may illustrate that, for most symptoms, capturing general language use rather than searching for direct answers to the item of the questionnaire is more appropriate. From the results, we can conclude that the proposed classification framework performs considerably well in the depression estimation scenario. Despite the simplicity of our approaches, they still show better performance than presented baselines. Moreover, we have to stress that, in contrast to the participant systems, our framework does not (*i*) use external datasets, (*ii*) apply an elaborated set of textual and hand-crafted features, or (*iii*) rely on complex decision models based on ensembles of different machine learning classifiers and features. Furthermore, the decisions are easily interpretable as we can compute the similarity value of all the options contributing to the explainability of the model.

In previous results reported in Table 6, our methods used the best value of top k obtained in the training process. In addition to this, we ran experiments to evaluate the impact on the number of sentences filtered in the data selection strategy. For this reason, we experimented with different values of k and evaluated directly on the test set. This allowed us to carry out an analysis covering a larger number of users. We tuned different k values to generate the options representations. More specifically, from 100 to the max possible value (2600) with increments of 100. We also experimented with a few low k values, such as 25 and 50. We note that the users in the training collection have different number of sentences. The maximum top k retrieved depends on each user. For example, the user with the most retrieved sentences with BM25 is 2608. However, this value does not exceed 600 sentences for most users. From 600 to 2600, the number of additional retrieved sentences is very low and has minimal impact on the results. Thus, we decided to cut the y-axis at 600 and report the results after considering almost all available sentences $(2400, 2500, 2600)$.

We used the general classifier in these experiments as it was the most sensitive to the filter process.[8] Visualizing Fig. 3, we can compare the influence of using more or less restrictive data selection strategies. The x-axis is the top k sentences selected for each training user. The y-axis is the performance for the AHR and DCHR metrics, showing the impact at symptom level and the total BDI score, respectively. Looking at these results, the performance for both metrics increases as you consider more sentences filtered by BM25. This suggests that, using the general classifier, the filtering technique improves by using more sentences retrieved by BM25. The improvement in the performance is even more significant in DCHR (predicting depressive categories). As the goal of the general classifier is to capture patterns in the language, this indicates that using more candidate sentences is useful for this method.

4.3. Symptom-by-symptom comparative study

Our main hypothesis is that the way depressive symptoms manifest in social media may differ depending on the privacy of the symptoms. For checking that, we designed a comparative symptom-by-symptom analysis of the general and direct classifiers. We applied leave-one-out cross-validation on the training data to determine which classifier performed better for each symptom. The metric maximized is the Averaged Hit Rate (AHR). Fig. 4 illustrates the results obtained in this process. The (x, y) points represent the symptoms. The x-axis is the AHR value obtained for the symptom when considering the best classifier. The y-axis for that symptom is the AHR difference between the direct and general classifier, and there we can see the differences in the performance of the two classifiers considered. On the positive y-axis, we can see the symptoms for which the direct classifier obtained a better value than the general one in terms of AHR (this difference is positive). The higher the value of the y, the higher is the performance of the direct respect to the general classifier. On the negative axis we see the symptoms better captured by the general classifier (the difference is negative), so lower values means that the difference of the general respect to the direct is more significant. Finally, we can also visualize the percentage of improvement of the best classifier over the results of the lowest performing one. The circle size depicts this percentage, which helps to reflect the variability of results depending on the symptom. Bigger circles mean that the improvement is greater.

Analyzing this figure, we can extract relevant conclusions: (*i*) the results suggest a great variability of the performance depending on the symptom. For instance, while for *Sense of Punishment*, we got around 80% of the options correct with the best classifier, in other like *Sleep changes* the best value drops to less than 20%. This fact

[8] As reported in Section 4.1 the direct classifier always obtained better results without any previous filter.

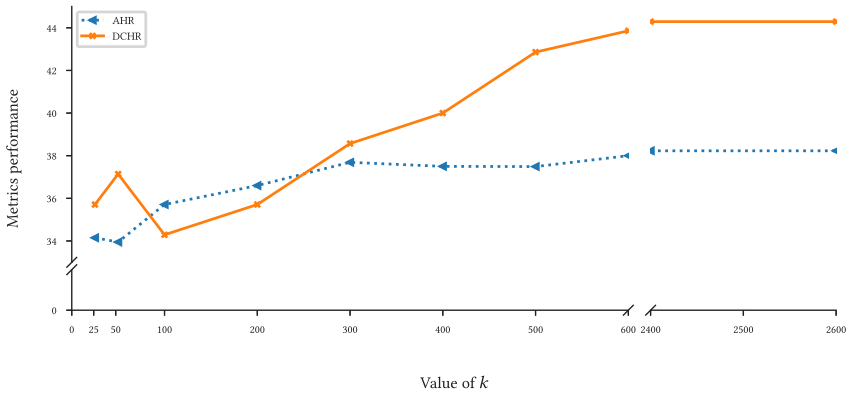

Fig. 3. Effect on the test collection depending on the number of filtered sentences (k) using the general symptom-classifiers.

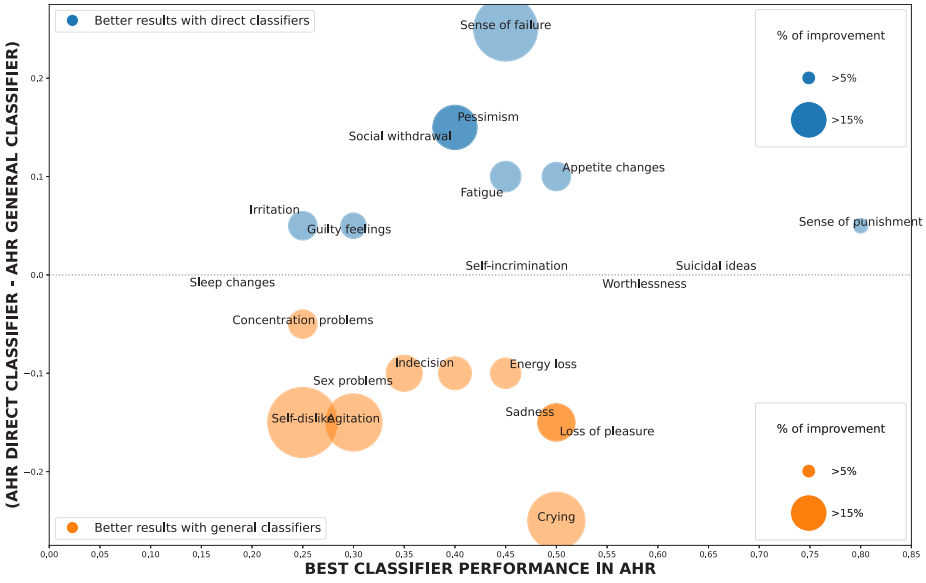

Fig. 4. Results of the comparative study of the symptoms considering the general and direct classifiers.

shows that there are symptoms very challenging to capture from the writings regardless of the method. For most symptoms, the best AHR is just under 50%. (*ii*) The two classifiers show remarkable differences in their results, indicating that there is also a significant variability in the performance depending on the method considered. Only in four symptoms we found no difference in AHR (*Sleep changes, Self-incrimination, Worthlessness and Suicidal ideas*), which means that both proposals achieved the same results. The percentage of improvement also supports this assumption. For instance, in the symptom *Sense of failure*, we obtained an increment of 125% using the direct classifier, while in *Crying* the general exceeded the 100% improvement. This symptom-analysis is consistent with the results obtained in the test collection, where we achieved an improvement using the mixed classifier. All this evidence validates our initial hypothesis about the different nature of depressive symptoms, exemplifying how performance can drastically vary depending on the symptom particularities. Lastly, we want to clarify that as the collections correspond to English-speakers, the embedding models were previously trained on texts in the same language. However, the architectures of these models, as well as our whole framework, are completely transferable to any language.

To further analyze the results obtained for each symptom, we also examined the impact that the distribution of training and test users may have on our symptom-classifiers. For this purpose, we show the balance of the answers ($0-3$) for all the symptoms in Fig. 5. The first two rows correspond to the distribution of the options answered by the users in the training and test sets. The last row corresponds to the distribution of our general symptom-classifier estimations.[9]

Visualizing Fig. 5, we can first see that for certain symptoms in the training set there are no users for all the options. For instance, there are no users answering option 3 for the symptoms *Sadness*, *Suicidal ideas* and *Irritability*. Due to this lack of data, our approach was unable to generate the representations for the option 3 in these symptoms. As a consequence, our classifiers cannot classify the option 3 for these symptoms in inference. That is a limitation of our approach that relies on the existence of training data. Secondly, regarding our results, the distributions of our general classifier answers shows that our classifier often underestimates the severity of symptoms. The median of our decisions is always in options 0 and 1. However, we see a higher presence of more severe options in both the training and test set. Finally, we do not see evidence that our trained models may be overfitted towards the majority class on the trained data. This is a good behavior given the limited amount of training data available.

5. Conclusions and future work

In this paper, we explored the potential of using neural language models to serve as a tool to estimate depressive states. For that, we

[9] We display the results from the general classifier as it was the one with the most consistent results across metrics.

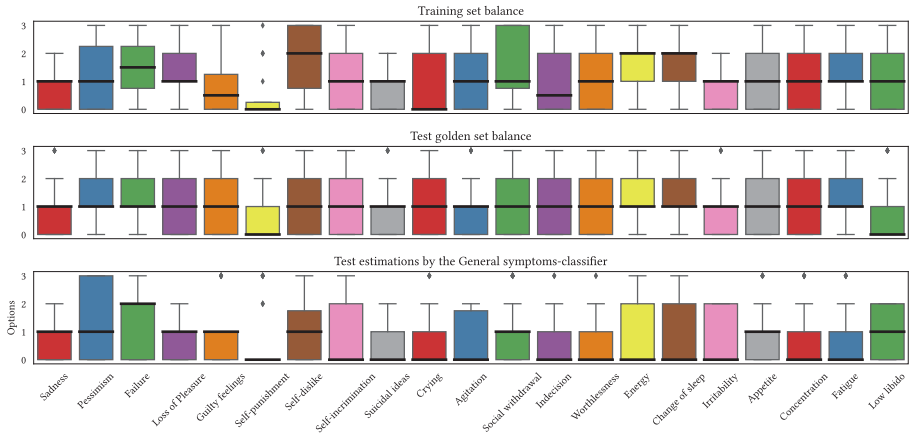

Fig. 5. Distribution of the options in the training (first row), test set (second row) and general symptom-classifier decisions (third row) for all the BDI symptoms.

designed a classification framework to estimate the severity level of the symptoms that characterize depression. Traditional depression detection approaches addressed this task as a depressed vs not depressed classification problem. In contrast, the work presented in this article produces a prediction for each depressive symptom. We used the 21 symptoms collected in the BDI-II questionnaire as the base of our research. Our work is inspired by one main observation: the symptoms can differ a lot in terms of sensitivity and openness to talk about them. Consequently, patients will be more likely to be mention and comment on some symptoms publicly on social media. To assess this idea, we proposed and evaluated two variants: (1) general symptom-classifiers, which captures individuals' general use of the language and (2) direct symptom-classifiers, which only captures direct answers related to the symptoms. Our results showed that the proposed approaches are simple but effective for estimating depressive levels, and at the same time, are flexible and easier to interpret than other presented baselines. The training and test set was composed by 20 and 70 users, respectively. Considering the small number of training users in the collection, we believe that the results of the presented methods may be benefited from larger training corpus.

Our study included a comparative symptom-by-symptom analysis of these two methods to determine which one performs best for each symptom. The analysis showed that: (1) there is great variability in performance between the two methods, and (2) certain symptoms are much more complex to capture than others, regardless of the method. Following our comparative study, we proposed a mixed solution that uses the best approach for each the symptom. This new mixed classifier achieved state-of-the-art results, outperforming all previous methods. Our findings suggest that there may be an relevant connection between the sensitivity of the symptoms and the performance of predictive approaches.

The methods presented in this study have a variety of potential applications. As future work, we are interested in using other datasets to confirm our results, considering other social platforms and languages. Besides, we will examine different psychological diseases that also incorporate the use of questionnaires for their detection. More specifically, we will attempt to translate our models to similar diseases: gambling by using the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)[87], or eating disorders with the Eating Disorder Inventory-III (EDI-III)[88]. Finally, we intend to investigate different directions for further improvement. Due to the adaptability of our framework, we will experiment with novel sentence representation architectures instead of word-level representations.

6. Ethical and practical considerations

Due to the sensitivity of any topic related to mental health content, we find necessary to discuss the ethical considerations and implications of our approaches. The collections used in this work are publicly available following the corresponding data usage policies. In this context, all users have an anonymous state, and the publications shown here are paraphrased to preserve their privacy. Similarly, we have not attempted to identify any traits that would attempt to reveal any personal information such as the users location, age or gender.

Considering the impact in real life settings, the performance of our solutions is far from ideal, and there is still much work to be done to advance toward more effective depression screening tools. Our models are exclusively trained on data from Reddit platform. For this reason, they are likely to have to be re-trained when considering other data sources, such as different social platforms. Moreover, the models are likely to have a poor generalization when processing data from different contexts (e.g., clinical records). For all these reasons, it is essential to recall that these systems do not intend to replace health professionals' tasks but rather serve as support tools. We envision automated technologies that could complement current online screening approaches to improve the actual detection rates.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has received support from projects: project RTI-2018-093336-B-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación & ERDF, Spain); project PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next GenerationEU, Spain); Consellería de Educación, Universidade e Formación Profesional, Spain (accreditation 2019–2022 ED431G/01 and GPC ED431B 2022/33) and the European Regional Development Fund, which acknowledges the CITIC Research Center, an ICT of the University of A Coruña as a Research Center of the Galician University System. The first author also acknowledges the predoctoral grant contract ref. ED481 A 2021/034 funded by Xunta de Galicia, Spain and the European Social Fund (ESF). The authors also want to thank the funding for open access charge: Universidade da Coruña/CISUG.

References

[1] Preamble to the constitution of the World Health Organization as adopted by the International Health Conference. 1948.

[2] Prince M, Patel V, Saxena S, Maj M, Maselko J, Phillips M, et al. No health without mental health. Lancet 2007;370:859–77. http://dx.doi.org/10.1016/S0140-6736(07)61238-0.

[3] World Health Organization, et al. Promoting mental health: concepts, emerging evidence, practice: A report of the world health organization, department of mental health and substance abuse in collaboration with the victorian health promotion foundation and the University of Melbourne. World Health Organization; 2005.

[4] Hollon SD, Thase ME, Markowitz JC. Treatment and prevention of depression. Psychol Sci Public Interest 2002;3(2):39–77.

[5] Kessler RC, Aguilar-Gaxiola S, Alonso J, Chatterji S, Lee S, Ormel J, et al. The global burden of mental disorders: an update from the WHO World Mental Health (WMH) surveys. Epidemiologia E Psichiatria Sociale 2009;18(1):23.

[6] Saxena S, Funk M, Chisholm D. Comprehensive mental health action plan 2013–2020. EMHJ-Eastern Mediterr Health J 2015;21(7):461–3.

[7] Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, et al. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the National Comorbidity Survey. Arch Gen Psychiatry 1994;51(1):8–19.

[8] Patel V, Chisholm D, Parikh R, Charlson FJ, Degenhardt L, Dua T, et al. Global priorities for addressing the burden of mental, neurological, and substance use disorders. 2016.

[9] Pierce M, Hope H, Ford T, Hatch S, Hotopf M, John A, et al. Mental health before and during the COVID-19 pandemic: a longitudinal probability sample survey of the UK population. Lancet Psychiatry 2020;7(10):883–92.

[10] Duarte PS, Miyazaki MC, Blay SL, Sesso R. Cognitive–behavioral group therapy is an effective treatment for major depression in hemodialysis patients. Kidney Int 2009;76(4):414–21.

[11] Alonso J, Codony M, Kovess V, Angermeyer MC, Katz SJ, Haro JM, et al. Population level of unmet need for mental healthcare in Europe. Br J Psychiatry 2007;190(4):299–306.

[12] Picardi A, Lega I, Tarsitani L, Caredda M, Matteucci G, Zerella M, et al. A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. J Affect Disord 2016;198:96–101.

[13] Halfin A. Depression: the benefits of early and appropriate treatment. Am J Managed Care 2007;13(4):S92.

[14] Rost K, Smith JL, Dickinson M. The effect of improving primary care depression management on employee absenteeism and productivity, a randomized trial. Med Care 2004;42(12):1202.

[15] Smarr KL, Keefer AL. Measures of depression and depressive symptoms: beck depression inventory-II (BDI-II), center for epidemiologic studies depression scale (CES-d), geriatric depression scale (GDS), hospital anxiety and depression scale (HADS), and patient health questionnaire-9 (PHQ-9). Arthritis Care Res 2011;63(S11):S454–66.

[16] Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606–13.

[17] Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983;67(6):361–70.

[18] Eaton WW, Smith C, Ybarra M, Muntaner C, Tien A. Center for epidemiologic studies depression scale: review and revision (CESD and CESD-R). 2004.

[19] Hamilton M. Rating depressive patients. J Clin Psychiatry 1980.

[20] Beck AT, Steer RA, Brown G. Beck depression inventory–II. Psychol Assess 1996.

[21] Dozois DJ, Dobson KS, Ahnberg JL. A psychometric evaluation of the Beck Depression Inventory–II. Psychol Assess 1998;10(2):83.

[22] Beck AT, Steer RA, Carbin MG. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. Clin Psychol Rev 1988;8(1):77–100.

[23] Sanchez-Villegas A, Schlatter J, Ortuno F, Lahortiga F, Pla J, Benito S, et al. Validity of a self-reported diagnosis of depression among participants in a cohort study using the Structured Clinical Interview for DSM-IV (SCID-I). BMC Psychiatry 2008;8(1):1–8.

[24] US Department of Health and Human Services, et al. Centers for disease control and prevention (CDC) behavioral risk factor surveillance system survey data. 1993–2012 Atlanta, Georgia. 10, 2012, Retrieved on January.

[25] Cameron IM, Cardy A, Crawford JR, du Toit SW, Hay S, Lawton K, et al. Measuring depression severity in general practice: discriminatory performance of the PHQ-9, HADS-D, and BDI-II. Br J Gen Pract 2011;61(588):e419–26.

[26] Bowling A. Mode of questionnaire administration can have serious effects on data quality. J Public Health 2005;27(3):281–91.

[27] De Choudhury M, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: Proceedings of the 5th annual ACM web science conference, 2013. p. 47–56.

[28] Colineau N, Paris C. Talking about your health to strangers: understanding the use of online social networks by patients. New Rev Hypermedia Multimedia 2010;16(1–2):141–60.

[29] Losada DE, Crestani F. A test collection for research on depression and language use. In: International conference of the cross-language evaluation forum for european languages. Springer; 2016, p. 28–39.

[30] Aragón ME, Monroy APL, González-Gurrola LC, Montes M. Detecting depression in social media using fine-grained emotions. 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), 2019. p. 1481–6.

[31] Prieto VM, Matos S, Alvarez M, Cacheda F, Oliveira JL. Twitter: a good place to detect health conditions. PLoS One 2014;9(1):e86191.

[32] De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. In: Proceedings of the international AAAI conference on web and social media, Vol. 7, 2013.

[33] De Choudhury M, Counts S, Horvitz EJ, Hoff A. Characterizing and predicting postpartum depression from shared facebook data. In: Proceedings of the 17th ACM conference on computer supported cooperative work & social computing, 2014. p. 626–38.

[34] Stordal E, Mykletun A, Dahl A. The association between age and depression in the general population: a multivariate examination. Acta Psychiatr Scand 2003;107(2):132–41.

[35] Lienemann BA, Siegel JT, Crano WD. Persuading people with depression to seek help: Respect the boomerang. Health Commun 2013;28(7):718–28.

[36] Barney LJ, Griffiths KM, Jorm AF, Christensen H. Stigma about depression and its impact on help-seeking intentions. Aust N Z J Psychiatry 2006;40(1):51–4.

[37] Pennebaker JW. Opening up: the healing power of expressing emotions. Guilford Press; 1997.

[38] Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language use: Our words, our selves. Ann Rev Psychol 2003;54(1):547–77.

[39] Campbell RS, Pennebaker JW. The secret life of pronouns: Flexibility in writing style and physical health. Psychol Sci 2003;14(1):60–5.

[40] Rude S, Gortner E-M, Pennebaker J. Language use of depressed and depression-vulnerable college students. Cogn Emot 2004;18(8):1121–33.

[41] MacAvaney S, Desmet B, Cohan A, Soldaini L, Yates A, Zirikly A, et al. Rsdd-time: Temporal annotation of self-reported mental health diagnoses. 2018, arXiv preprint arXiv:1806.07916.

[42] Cohan A, Desmet B, Yates A, Soldaini L, MacAvaney S, Goharian N. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. 2018, arXiv preprint arXiv:1806.05258.

[43] De Choudhury M, Kiciman E. The language of social support in social media and its effect on suicidal ideation risk. In: Proceedings of the international AAAI conference on web and social media, Vol. 11, 2017.

[44] Desmet B, Hoste V. Online suicide prevention through optimised text classification. Inform Sci 2018;439:61–78.

[45] Mitchell M, Hollingshead K, Coppersmith G. Quantifying the language of schizophrenia in social media. In: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, 2015. p. 11–20.

[46] Wang T, Brede M, Ianni A, Mentzakis E. Detecting and characterizing eating-disorder communities on social media. In: Proceedings of the tenth ACM international conference on web search and data mining, 2017. p. 91–100.

[47] Coppersmith G, Dredze M, Harman C, Hollingshead K. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, 2015. p. 1–10.

[48] De Choudhury M, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: Proceedings of the 5th annual ACM web science conference, 2013. p. 47–56.

[49] De Choudhury M, De S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In: Eighth international AAAI conference on weblogs and social media.

[50] Chen X, Sykora MD, Jackson TW, Elayan S. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In: Companion proceedings of the the web conference 2018, 2018. p. 1653–60.

[51] Kawachi I, Berkman LF. Social ties and mental health. J Urban Health 2001;78(3):458–67.

[52] Ortega-Mendoza RM, Hernández-Farías DI, Montes-y Gómez M, Villaseñor-Pineda L. Revealing traces of depression through personal statements analysis in social media. Artif Intell Med 2022;123:102202.

[53] Jamil Z, Inkpen D, Buddhitha P, White K. Monitoring tweets for depression to detect at-risk users. In: Proceedings of the fourth workshop on computational linguistics and clinical psychology — from linguistic signal to clinical reality. Vancouver, BC: Association for Computational Linguistics; 2017, p. 32–40. http://dx.doi.org/10.18653/v1/W17-3104, URL https://aclanthology.org/W17-3104.

[54] Tlelo-Coyotecatl I, Escalante HJ, Montes M, et al. Depression recognition in social media based on symptoms' detection. Procesamiento Del Lenguaje Natural 2022;68:25–37.

[55] Nguyen T, Yates A, Zirikly A, Desmet B, Cohan A. Improving the generalizability of depression detection by leveraging clinical questionnaires. 2022, arXiv preprint arXiv:2204.10432.

[56] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. Proceedings of Workshop at ICLR 2013;2013.

[57] Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014. p. 1532–43.

[58] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019, p. 4171–86. http://dx.doi.org/10.18653/v1/N19-1423.

[59] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020;33:1877–901.

[60] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 2020;21:140:1–67, URL http://jmlr.org/papers/v21/20-074.html.

[61] Kenter T, De Rijke M. Short text similarity with word embeddings. In: Proceedings of the 24th ACM international on conference on information and knowledge management, 2015. p. 1411–20.

[62] Giatsoglou M, Vozalis MG, Diamantaras K, Vakali A, Sarigiannidis G, Chatzisavvas KC. Sentiment analysis leveraging emotions and word embeddings. Expert Syst Appl 2017;69:214–24.

[63] Alsentzer E, Murphy J, Boag W, Weng W-H, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd clinical natural language processing workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019, p. 72–8. http://dx.doi.org/10.18653/v1/W19-1909, URL https://aclanthology.org/W19-1909.

[64] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–40.

[65] Martínez-Castaño R, Htait A, Azzopardi L, Moshfeghi Y. Early risk detection of self-harm and depression severity using BERT-based transformers: iLab at CLEF eRisk 2020. Early Risk Predict Internet 2020.

[66] Bucur A-M, Cosma A, Dinu LP. Early risk detection of pathological gambling, self-harm and depression using BERT. In: CLEF. 2021.

[67] Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M. CLPsych 2015 shared task: Depression and PTSD on Twitter. In: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, 2015. p. 31–39.

[68] Losada DE, Crestani F, Parapar J. eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In: International conference of the cross-language evaluation forum for european languages. Springer; 2017, p. 346–60.

[69] Losada DE, Crestani F, Parapar J. Overview of eRisk 2018: Early Risk Prediction on the Internet (extended lab overview). In: Proceedings of the 9th international conference of the CLEF association, CLEF, 2018. p. 1–20.

[70] Losada DE, Crestani F, Parapar J. Overview of erisk 2019 early risk prediction on the internet. In: International conference of the cross-language evaluation forum for european languages. Springer; 2019, p. 340–57.

[71] Losada DE, Crestani F, Parapar J. Erisk 2020: Self-harm and depression challenges. In: European conference on information retrieval. Springer; 2020, p. 557–63.

[72] Burdisso SG, Errecalde M, Montes-y Gómez M. A text classification framework for simple and effective early depression detection over social media streams. Expert Syst Appl 2019;133:182–97.

[73] Rijen Pv, Teodoro D, Naderi N, Mottin L, Knafou J, Ruch P. Data-driven approach for measuring the severity of the signs of depression using reddit posts. In: Proceedings of CLEF (conference and labs of the evaluation forum) 2019 working notes, no. CONFERENCE, 9-12 September 2019, 2019.

[74] Abed-Esfahani P, Howard D, Maslej MM, Patel S, Mann V, Goegan S, et al. Transfer learning for depression: Early detection and severity prediction from social media postings. In: CLEF. 2019.

[75] Burdisso SG, Errecalde ML, Montes y Gómez M. Using text classification to estimate the depression level of reddit users. J Comput Sci Technol 2021;21.

[76] Maupomé D, Armstrong MD, Belbahar RM, Alezot J, Balassiano R, Queudot M, et al. Early mental health risk assessment through writing styles, topics and neural models. In: CLEF (Working Notes). 2020.

[77] Oliveira L. BioInfo@ UAVR at eRisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases. 2020.

[78] Uban A-S, Rosso P. Deep learning architectures and strategies for early detection of self-harm and depression level prediction. In: CEUR workshop proceedings. Vol. 2696, Sun SITE Central Europe; 2020, p. 1–12.

[79] Steer R, Beck A, Garrison B. Applications of the beck depression inventory. In: Assessment of depression. Springer; 1986, p. 123–42.

[80] Joshi A, Kale S, Chandel S, Pal DK. Likert scale: Explored and explained. Curr J Appl Sci Technol 2015;396–403.

[81] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Now Publishers Inc; 2009.

[82] Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A. Advances in pre-training distributed word representations. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA); 2018.

[83] Trask A, Michalak P, Liu J. Sense2vec-a fast and accurate method for word sense disambiguation in neural word embeddings. 2015, arXiv preprint arXiv:1511.06388.

[84] Zhang Z, Yang J, Zhao H. Retrospective reader for machine reading comprehension. 2020, arXiv preprint arXiv:2001.09694.

[85] Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 conference on empirical methods in natural language processing. Austin, Texas: Association for Computational Linguistics; 2016, p. 2383–92. http://dx.doi.org/10.18653/v1/D16-1264.

[86] Cer D, Yang Y, Kong S-y, Hua N, Limtiaco N, John RS, et al. Universal sentence encoder for English. In: Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations, 2018. p. 169–74.

[87] Strong DR, Kahler CW. Evaluation of the continuum of gambling problems using the DSM-IV. Addiction 2007;102(5):713–21.

[88] Garner DM. Eating disorder inventory-3 (EDI-3). In: Professional manual. Odessa, FL: Psychological Assessment Resources; 2004.