Procesamiento del Lenguaje Natural, Revista nº 68, marzo de 2022, pp. 25-37 recibido 18-12-2021 revisado 20-01-2022 aceptado 26-01-2022

Depression Recognition in Social Media based on Symptoms' Detection

Reconocimiento de depresión en redes sociales basado en la detección de síntomas

Itzel Tlelo-Coyotecatl, Hugo Jair Escalante, Manuel Montes-y-Gómez
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, Mexico
{itlelo, hugojair, mmontesg}@inaoep.mx

Abstract: Depression is a common mental disorder that affects millions of people around the world. Recently, several methods have been proposed that detect people suffering from depression by analyzing their language patterns in social media. These methods show competitive results, but most of them are opaque and lack of explainability. Motivated by these problems, and inspired by the questionnaires used by health professionals for its diagnosis, in this paper we propose an approach for the detection of depression based on the identification and accumulation of evidence of symptoms through the users' posts. Results in a benchmark collection are encouraging, as they show a competitive performance with respect to state-of-the-art methods. Furthermore, taking advantage of the approach's properties, we outline what could be a support tool for healthcare professionals for analyzing and monitoring depression behaviors in social networks.
Keywords: Depression detection, social media, information retrieval.

Resumen: La depresión es un trastorno mental que afecta a millones de personas en todo el mundo. Recientemente, se han propuesto varios métodos que detectan personas que sufren depresión analizando sus patrones de lenguaje en las redes sociales. Estos métodos han mostrado resultados competitivos, sin embargo la mayoría son opacos y carecen de explicabilidad. Motivados por estos problemas, e inspirados en los cuestionarios utilizados por los profesionales de la salud para su diagnóstico, en este trabajo proponemos un método para la detección de depresión basado en la identificación y acumulación de evidencia de síntomas a través de las publicaciones de los usuarios. Los resultados obtenidos en una colección de referencia son prometedores, ya que muestran un desempeño competitivo con respecto a los mejores métodos actuales. Además, aprovechando las propiedades del método, describimos lo que podría ser una herramienta de apoyo para que los profesionales de la salud analicen y monitoreen las conductas depresivas en las redes sociales.
Palabras clave: Detección de depresión, redes sociales, recuperación de información.

1 Introduction

The World Health Organization (WHO) defines mental health as a state of emotional, psychological and social well-being that influences the way a person thinks, feels, acts or relates to others (World Health Organization, 2003). Accordingly, mental disorders refer to conditions that may affect the way of thinking, feeling or acting of persons. Among mental disorders, depression is one of the most common, affecting around 3.4% of the world's population (Saloni Dattani and Roser, 2021).

The diagnosis of depression is usually carried out by mental health professionals through the application of interviews or questionnaires focused on identifying the presence of certain symptoms (National Institute of Mental Health, 2021). These diagnostic methods are effective, but their coverage is limited mainly due to economic factors and social stigmatization (World Health Organization, 2003). These drawbacks, together with the need to address this growing problem, have motivated the development of computational tools for the automatic detection and monitoring of people suffering from depression. Particularly, the link between lan-

guage usage and the psychological state of people (Pennebaker, Mehl, and Niederhoffer, 2003) has led to the exploration of data from social networks for the automatic detection of depression, aiming to take advantage of the large amount of information generated by people through these media, in which they usually express their interests, concerns and feelings (Guntuku et al., 2017).

Current automatic methods usually address the depression detection task as a text classification problem, considering all the information shared by users, without necessarily adopting the traditional methodology that emphasizes the identification and measurement of symptoms. Most of these methods have achieved competitive results in benchmark collections (Losada, Crestani, and Parapar, 2018) (Coppersmith et al., 2015), but have also shown limitations in terms of transparency. Given the importance of the explainability of the decisions in this very sensitive task (Danilevsky et al., 2020) (Ríssola, Aliannejadi, and Crestani, 2020), we presume that the design of methods based on the identification of evidence of symptoms through the users' post, similar to the traditional diagnostic approach, could considerably improve the interpretation of the results. This paper describes a contribution in such direction.

In particular, in this work we propose an approach for depression detection in social media based on the identification and accumulation of evidence of symptoms. This approach has three main stages. In the first one users' posts are filtered, keeping only those that refer to or are related to any of the 21 symptoms declared on Beck's Depression Inventory (BDI)[1]. Then, in the second stage, 21 independent classifiers are built from the sets of filtered posts. The idea is that each classifier observes the target user from a different perspective, determining whether she or he suffers from depression or not according to the presence of only one of the symptoms. Finally, in a third stage, the decisions of the different classifiers are combined to generate a final unified prediction. Through this three-stage process, which gradually identifies and integrates evidence of the different symptoms, we move a step forward by facilitating the interpretation of decisions, and thereby enabling its usage in social media monitoring applications.

Summarizing, the main contributions of this work are:

- We propose a new approach for depression detection in social media based on the analysis of the presence of depression symptoms through users' posts.
- We carry out an in-depth analysis of the presence of the different symptoms in users' posts and their correlation with the classification errors, providing insights on their relevance for the detection of depression in social media.
- We outline a simple interface to support the detection and follow-up of users who suffer from depression.

The remainder of the paper is organized as follows. Section 2 presents a brief overview of related work on depression detection in social media. Section 3 describes the proposed approach for depression detection based on the symptoms identification and accumulation. Sections 4 and 5 reports the experiments, results, and their analysis. Section 6 outlines what could be a support tool for healthcare professionals to analyze and monitor depression behaviors in social networks. Finally, Section 7 points out our conclusions and future work.

2 Related Work

As we previously mentioned, the detection of depression in social media has been handled as a supervised learning problem, where the main goal is to build a model that distinguishes users suffering from depression from healthy users (Guntuku et al., 2017).

Most current methods rely on the use of traditional machine learning processes or deep learning techniques. On the one side, there are works like the ones from Nadeem (2016), Jamil et al. (2017) and Preoţiuc-Pietro et al. (2015) which consider a bag-of-words or word n-grams as the users' representation, and employ traditional learning algorithms to build the classifier. This kind of methods are very popular due to their low computational complexity and the easiness for interpreting their results related with depressive tendencies (Tsugawa et al., 2013).

[1]The considered version of the BDI we used corresponds to the provided at the CLEF eRisk2019 available at https://early.irlab.org/2019/index.html

High dimensionality is a known problem of bag-of-words approaches, which has prompted the use of manually or automatically-defined topic-based representations as an alternative. Examples of this are the works by Wolohan et al. (2018) that uses the LIWC categories as features, and by Loveys et al. (2018) that carry out an analysis of the language usage involving cultural differences for depression considering the LIWC categories, topic modeling, data visualization, and other techniques.

From another perspective, there are studies that have considered the use of information about sentiments and emotions to analyze depression behaviours in social media (Preoţiuc-Pietro et al., 2015), (De Choudhury et al., 2013), and (Aragon et al., 2021). These works have shown interesting results, indicating that negative posts, as well as certain emotions like anger and fear, are more abundant in people with depression than in users who do not suffer from this disorder.

On the other hand, some recent successful works have approached the combination of hand-crafted and automatically learned representations using deep learning models, such as CNNs or LSTMs (Liu et al., 2018), (Husseini Orabi et al., 2018), (Yates, Cohan, and Goharian, 2017), (Trotzek, Koitka, and Friedrich, 2018a). Despite their good results, this kind of methods have important drawbacks. For example, they require large amount of data to train their models, have high complexity, and consequently low explainability. Moving in the latter direction, the works by Mathur et al. (2020), Trifan et al. (2020) and Ríssola, Aliannejadi, and Crestani (2020) have integrated psycholinguistic features into the users' representations. Their common idea is to build effective methods, but also capable of providing understanding and descriptions of the decisions made (Burdisso, Errecalde, and Montes-y-Gómez, 2019), (Zogan et al., 2020).

Lastly, the eRisk evaluation forum[2] exemplifies the evolution that this task has had in the recent years. In its last editions (Losada, Crestani, and Parapar, 2020), (Parapar et al., 2021), it has included a subtask that consists of estimating the level of depression from a thread of user posts, in other words, of filling a standard depression questionnaire based on the evidence found in the history of postings. From these evaluations, its organizers have concluded that "*Although the effectiveness of the proposed solutions is still modest, the experiments suggest that evidence extracted from social media is valuable, and that automatic or semi-automatic screening tools could be designed to detect at-risk individuals*". Our work follows precisely this research direction. However, we approach it again as a binary classification problem, aimed at distinguishing users who suffer from depression from healthy users, but we adopt the idea of identifying and accumulating evidence of depression symptoms, and even more, we outline the proposal of a support tool for analyzing and monitoring depression behaviors in social networks.

3 Depression Detection based on Symptoms Evidence

This section describes our proposed method for depression detection based on social media content. Figure 1 shows its general diagram, which comprises three main stages:

1. *Symptoms evidence retrieval*, where the aim is to identify evidence associated to depression symptoms from users' posts.
2. *Symptom-based classification*, where we build predictive models for distinguishing healthy users from users suffering from depression based on the presence of each one of the symptoms.
3. *Depression status prediction*, where we combine the evidence of the identified symptoms to make a final prediction on the presence or absence of depression.

In the following subsections we detail each stage accordingly.

3.1 Symptoms Evidence Retrieval

The first stage of the proposed method consists of identifying candidate posts that can be considered evidence for the 21 BDI declared symptoms. BDI is a standard questionnaire composed by 21 questions, each one related to one possible depression symptom, applied on traditional depression diagnosis detection made by professionals[3] (Beck et al., 1961). Table 1 lists these symptoms.

[2]Early Risk Prediction on the Internet (eRisk) website: https://erisk.irlab.org/

[3]Alternative questionnaires based on symptoms identification for depression detection could also be considered (e.g., PHQ-9).

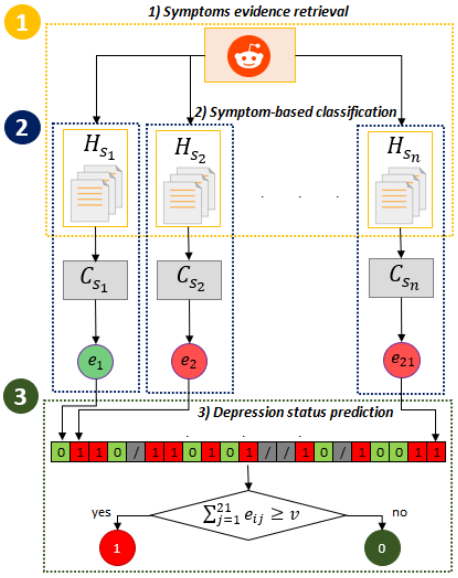

Figure 1: Diagram of the proposed method for depression detection based on symptoms identification.

We approach this task as one of information retrieval, where we want to retrieve relevant posts (*documents*) given the different symptoms (*queries*). More specifically, given a user's history made up of a set of posts $P = \{p_1, p_2, ..., p_n\}$, and the set of symptoms $S = \{s_1, s_2, ..., s_{21}\}$, the goal is to determine whether a post p_i is considered evidence (i.e., it is *relevant*) for each of the symptoms s_j with $j \in [1, 21]$.

For the retrieval process we relied on word embeddings and a similarity threshold. In preliminary work we evaluated other information retrieval models but the one reported herein resulted in better retrieval performance. This process is as follows:

Each symptom s_j is associated to a set of keywords $s_j = \{ws_1, ws_2, ..., ws_l\}$ with $j \in [1, 21]$ and $l \in [1, 3]$, according to Table 1[4]. In the same way, for each post p_i its set of words corresponds to $p_i = \{wp_1, wp_2, ..., wp_m\}$ where i indicates the number of post in the user history.

We say a post p_i is considered evidence for a symptom s_j (i.e., it is relevant) if there exist a tuple of words (wp_i, ws_j) whose cosine similarity in a particular embedding space is above a threshold β. The threshold β is a parameter that should be fixed depending on how strict one wants to be on the similarity of words for determining relevance between posts and symptoms. As a result of this retrieval process we can identify posts associated to the symptoms, where each post can be associated to none or any of the 21 symptoms. The process is applied to all the posts on the user history and all the symptoms on the BDI for all the users to analyze. In this way, each user u_i is represented by the evidence sets $\mathbb{H}_i = \{H_{i1}, H_{i2}, ..., H_{i21}\}$, one set H_{ij} per symptom.

In the next subsection we describe the way this information is used for building predictive models to automatically detect the presence of symptoms. One should note that the retrieved evidence on the presence/absence of symptoms could be also used for interpretability or explainability purposes, see Section 6.

3.2 Symptom-based Classification

This stage considers the previous retrieved evidence for each symptom. The aim is to build a predictive model per symptom using the identified evidence, that is, to build 21 classifiers that observe a target user from different perspectives and determine whether she/he suffers from depression or not.

The construction of each classifier c_j follows the traditional text-classification approach. That is, given a set of labeled users $U = \{(u_1, y_1), (u_2, y_2), ..., (u_m, y_m)\}$, and $y_i \in \{0, 1\}$, where $y_i = 0$ indicates a healthy user and $y_i = 1$ one suffering from depression, the goal is to learn a function $c_j : u \rightarrow \{0, 1\}$. For that purpose, we represent the users through a BOW model with $tfidf$ weights. All 21 classifiers are trained over the same set of labeled users, but differ in how these are represented. That is, for the construction of the classifier c_j only the posts corresponding to the evidence sets H_{kj} for all users $u_k \in U$ are used, and, therefore, it learns to recognize evidence associated to the symptom s_j.

Once the 21 classifiers are trained on the different symptoms, they can be used to make predictions about the presence or absence of depression in a new user u_i. In the case where there is no evidence about symptom s_j in the user's post history (i.e., $H_{ij} = \emptyset$), then the classifier c_k returns a void value (indicated as "/" in Figure 1.

The information provided by predictive models built in this stage is combined in order to give a final prediction for each subject,

[4]These keywords were manually chosen by considering those that represent at best the symptoms' contexts after narrowing down their lists of synonyms.

#	Symptom (BDI)	Keywords	#	Symptom (BDI)	Keywords
s01	Sadness	sadness	s12	Loss of interest	apathetic, worthless
s02	Pessimism	pessimism	s13	Indecisiveness	indecisiveness, indecisive
s03	Past failure	failure	s14	Worthlessness	worthlessness, worthless
s04	Loss of pleasure	displeasure, dissatisfaction	s15	Loss of energy	apathetic, dispirited
s05	Guilty feelings	guilty	s16	Changes in sleep pattern	sleep
s06	Punishment feelings	punishment	s17	Irritability	irritability
s07	Self-dislike	dislike, self	s18	Changes in appetite	appetite
s08	Self-criticalness	criticalness, critical, self	s19	Concentration difficulty	disconcerted, concentration
s09	Suicidal thoughts or wishes	suicidal	s20	Tiredness or fatigue	tiredness, fatigue
s10	Crying	crying	s21	Loss of interest in sex	sex, disinterest
s11	Agitation	agitation			

Table 1: 21 declared symptoms on the Beck's Depression Inventory (BDI).

as described in the next subsection.

3.3 Depression Status Prediction

This stage comprises the integration of the obtained predictions from the different symptom's classifiers. For that purpose, in a first step, these predictions are concatenated on a vector that is considered as the evidence-based representation for the user under analysis. After this process, each user u_i is represented by a vector $\mathbf{e}_i = \langle e_{i1}, e_{i2}, ..., e_{i21} \rangle$, where each e_{ij} indicates the prediction of classifier c_j on user u_i. Then, the final step of the proposed approach aims at providing a global prediction on the detection of depression for each user. Since the predictions vector $\mathbf{e}$ already captures the symptoms' presence, we process such a vector to obtain a final prediction. A number of ways for delivering a prediction from $\mathbf{e}$ were studied, including, standard ensembles, stacking generalization and meta-classifiers. However, we found that the most effective way was based on thresholding the number of symptoms detected by the distinct classifiers from stage 2.

To assign the final class label to a user u_i a vote counting is done over the values of the evidence vector $\mathbf{e}_i$. That is, if $\sum_{j=1}^{21} e_{ij} \geq v$, then the user u_i is classified as suffering from depression, otherwise he or she is marked as a healthy user. The votes are interpreted as: at least v symptom classifiers should be positive in order to classify a user in the depressive class. Therefore, the increasing of v means more symptoms should be positive in order to declare a user as depressed.

4 Experimental Settings

This section presents the corpus used in the experiments, and describes the experimental setup and the baseline results considered.

4.1 Dataset

The experiments were carried out on the eRisk-2018 corpus for Task 1 "Early Detection of Signs of Depression" (Losada, Crestani, and Parapar, 2018). This corpus contains English writings of Reddit users belonging to two categories, depressed and non-depressed users. The first group of users explicitly expressed in one of their posts that they were *diagnosed* with depression. On the other hand, the second group is formed by randomly selected users from the Reddit platform. Table 2 shows the distribution of the two categories of users in the given corpus. It is worth noting that the corpus presents a high class imbalance, which has motivated the use of the F_1 score over the positive class as the main evaluation measure.

Category	Train	Test
Depressed (D)	135	79
Non-depressed (ND)	752	741
Total	887	880

Table 2: Categories distribution in the eRisk-2018 corpus.

Besides being used in the aforementioned evaluation campaign (Losada, Crestani, and Parapar, 2018), several authors have reported results on it. In the next section we compare the performance of our method with some of those references.

4.2 Method Configuration

Text preprocessing. It was performed using NLTK packages; we normalized the texts by lowercasing all words, removing the stopwords, links, numbers and special characters.

Evidence retrieval. We employed pre-trained Twitter Glove embeddings[5] of 100-dimensions. In addition, we considered different values for the β-threshold (from 0.1 to 0.9), but the best results were obtained with $\beta = 0.6$, hence we use this value for all of our experiments.

Symptom-based classification. The 21 symptom classifiers were built in the same way. In all cases we used a SVM with a linear kernel, $C = 1$, L2 normalization, and weighted class imbalance.

Prediction threshold. For the third stage we followed an exploratory approach, considering different values for the v-parameter, which thresholds the number of votes from individual classifiers. Particularly, we used values from 1 to 11 (majority vote).

4.3 Reference Approaches

The following reference models were used for comparison.

Traditional. The *complete* user histories are represented using a BoW model with TF-IDF weights. These are fed to a SVM classifier with the same configuration as our symptom-based classifiers.

LIWC. It uses a representation that combines the LIWC categories and a BOW representation, with the 3,000 words with the greatest chi^2 values. It also uses a SVM classifier with the same configuration as our symptom-based classifiers (Aragon et al., 2021).

BiLSTM and CNN. Both neural networks used 100 neurons, the adam optimizer, and GloVe embeddings of 300-dimensions. For the CNN, 100 random filters of sizes 1, 2 and 3 were applied (Aragon et al., 2021).

BoSE. It uses a histogram of fine-grained emotions as representation, which captures the presence and variability of emotions through the users' posts. It employs a SVM as a classifier (Aragon et al., 2021).

DPP-EXPEI-SVM. It uses a BOW as representation, but considering a weighting scheme that rewards the presence of personal information in the posts. It employes a SVM as classifier (Ortega-Mendoza et al., 2022).

[5]Twitter Glove embeddings were chosen due to their orientation towards social network language, nonetheless alternative versions could be considered. GloVe embeddings were obtained from: https://nlp.stanford.edu/projects/glove/

Best results at eRisk2018. We consider the two best reported results at the eRisk-2018 forum (Losada, Crestani, and Parapar, 2018). Both results correspond to variations of the same method, named as FHDO-BCSGA and FHDO-BCSGB respectively (Trotzek, Koitka, and Friedrich, 2018b). They consider the results of machine learning models, together with an ensemble model that combined different base predictions. The models employ user-level linguistic metadata, bag of words, neural word embeddings, and convolutional neural networks.

5 Results

In this section we present the results of experiments that aim to evaluate the proposed method, and to compare its performance with that of the state of the art.

5.1 Symptom-based Evaluation

The goal of this experiment was to evaluate depression detection performance by considering evidence from a single symptom. That is, we evaluate the performance of the 21 per-symptom classifiers c_j. Results of this experiment are shown in Figure 2.

Obtained results show that evidence obtained from some symptoms is more discriminative than that from others. For example, the classifier for the symptom s_3 : *past failure* achieved an F_1 measure value close to 0.6 by itself. However, the performance for most symptoms is much lower, even a couple of them (s_{13}*: indecisiveness* and s_{11}*: agitation*) obtained a value of $F_1 = 0$.

The values at the right of Figure 2 show the average amount of information retrieved for each of the symptoms. There is a clear correlation between the amount of retrieved evidence and the performance of classifiers with few exceptions (e.g., s_9 : *suicidal thoughts or whishes* and s_6 : *punishmetn feelings*). In general, the results of this experiment indicate that the performance obtained with the different classifiers is related to the amount and quality of information available in the users' histories for training them.

Results shown in Figure 2 show that it is possible to detect depression by using information from a single symptom. However, the performance of such individual classifiers is still low when compared to reference methods. In the next section we show that by

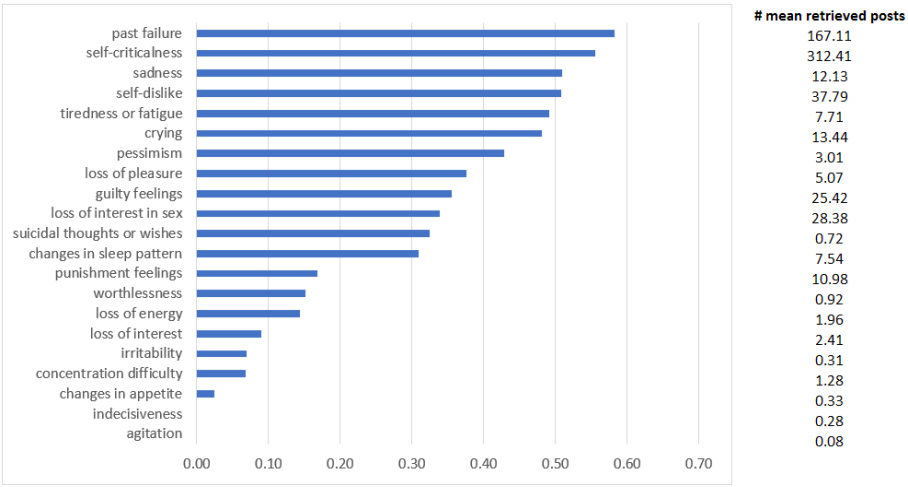

Figure 2: Evaluation F_1-score results for each one of the 21 BDI symptoms. The values on the right indicate the average retrieved posts for each symptom.

combining the information from these models it is possible to boost the performance.

5.2 Combining Evidences

Although the individual performance of symptom's classifiers is not that competitive, an hypothesis of this paper is that by combining the acquired evidence from all of the symptoms we could obtain better performance. This section aims to evaluate the combination of the predictions of individual models.

As described in Section 3.3, we aggregated the evidence obtained by the prediction vector $\mathbf{e}_i$, which is formed by the 21 predictions of individual classifiers for subject u_i, and used a threshold to determine whether the subject is detected as depressed (i.e., $\sum_{j=1}^{21} e_{ij} \geq v$) or not (otherwise). In the experiments we evaluated values for v lower than 11, since as reaching $v = 11$ votes from the individual classifiers means that more than half of the symptoms were detected from users' texts. We refer to this method as *ensemble of all symptoms* (SBC-ALL).

In addition, motivated by our previous results (see Section 5.1) that show that several symptoms did not present relevant or sufficient evidence for the detection of depression, we carried out an experiment combining only the decision of the best-half of the classifiers, selected according to their precision scores[6]. We named it as *ensemble of top symptoms* (SBC-TOP).

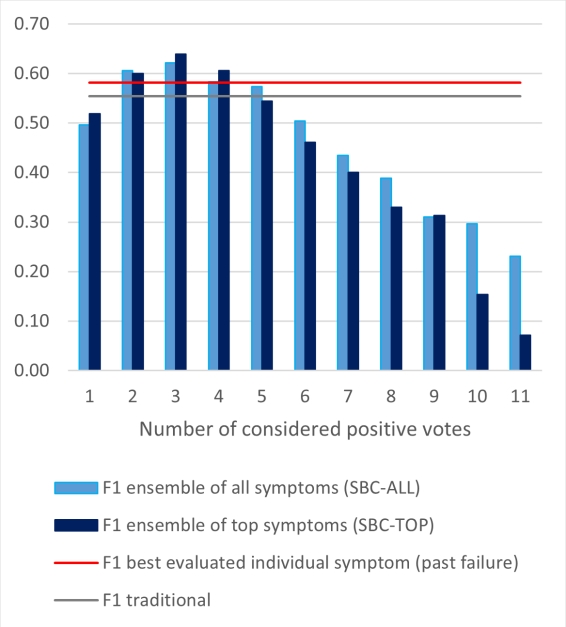

Figure 3: Reported F_1 results when combining the evidence of the individual models. We report the performance obtained when varying the threshold (x-axis) on the number of positively detected depression across the individual classifiers.

Figure 3 reports the results obtained with both ensemble approaches. In order to analyze the performance of these methods in detail, we report the obtained performance when varying the threshold on the number of positive outputs of the individual classi-

[6]The following symptom classifiers were selected for the SBC-TOP ensemble: *sadness, pessimism, past failure, loss of pleasure, guilty feelings, self-dislike, self-criticalness, crying, changes in sleep pattern, tiredness or fatigue* and *loss of interest in sex.*

fiers to detect depression with each ensemble. They indicate that it is slightly better to consider only the subset of the best individual symptoms than to consider all of them (i.e., SBC-TOP outperforms SBC-ALL). For both approaches detecting few positive symptoms ($v \leq 5$) allows achieving a high recall but at the expense of low precision levels, whereas increasing the number of symptoms identified as positive ($v \geq 6$) helps improving the precision but greatly affects the recall, which caused very low F_1 scores. Hence, the choice of v can be done according to what the final user judges as most important, either precision or recall.

Finally, it should be noted that the presence of only 3 symptoms was sufficient to obtain results that improved the results from the traditional baseline approach as well as from the best individual symptom classifier, gray and red horizontal lines respectively.

5.3 Comparison with the State of the Art

This section compares the performance of the best result obtained with our proposed method and reference techniques. For this experiment we consider the F_1-score on the positive class (i.e., depression) as evaluation measure. Table 3 compares the results of both ensemble approaches against all the methods described in Section 4.3.

Method	**P**	**R**	F_1
Traditional	0.565	0.544	0.554
LIWC	-	-	0.380
BiLSTM-GloVe	-	-	0.460
CNN-GloVe	-	-	0.510
BoSE	0.670	0.610	**0.640**
DPP-EXPEI-SVM	0.570	0.660	0.610
FHDO-BCSGB	0.640	0.650	**0.640**
FHDO-BCSGA	0.560	0.670	0.610
SBC-ALL	0.667	0.582	0.622
SBC-TOP	0.707	0.582	0.638

Table 3: Results for precision (P), recall (R) and F_1-score over the positive class of state of the art methods.

Results from Table 3 show some interesting points about the proposed approach. On the one hand, when comparing its results against the traditional baseline, it is evidenced the noisiness of the users' posts and, thus, the relevance of their filtering with respect to the presence of the depression symptoms. On the other hand, these results also show that methods based exclusively on neural network architectures are not the best alternative given the scarcity of training resources. In contrast, our approach better handles this complex scenario, by combining the decision of different independent classifiers, one for each symptom. Finally, it is observed that our two ensembles did not outperform the best result at the eRisk-2018 evaluation task, nonetheless, it is important to note that a comparable result was achieved using computationally less expensive methods, and even more important, that the proposed approach, due to its three-step architecture, facilitates the interpretation and explanation of the results, something critical in this type of applications.

5.4 Analysis and Discussion

This section presents additional experiments and results that aim to explain the performance obtained by the proposed solution, as well as highlighting its main benefits.

5.4.1 Maximum Possible F_1

The hypothesis behind our approach is that different users suffering from depression would show and express different symptoms through their posts. In consequence, by integrating the decisions of several independent symptom-based classifiers it would be possible to achieve high detection rates. The previous results were somewhat discouraging compared to this initial intuition, as the result of our best ensemble (SBC-TOP) only surpassed the best individual result (s_3 : *past failure*, see Figure 2) by around 5%.

In order to determine the potential of our initial idea, we calculated the performance obtained by a (hypothetical) perfect ensemble of the 21 symptom-based classifiers. To simulate this perfect ensemble we considered an user as correctly classified if at least one of the symptom-based classifiers did so (i.e., at least one of the c_j models returned a positive output). The results obtained by such a hypothetical ensemble was $F_1 = 0.66$, only 2% higher than that the result obtained by SBC-TOP, indicating that most of the symptoms allow to identify more or less the same group of depressed users, and they are not complementary to each other. From this result, we can conclude that the main direction of research should be in the improvement of the symptom-based classifiers and not in the

integration of their decisions.

5.4.2 More and Less Depressed Users

In this section we evaluate the performance of a simple model that approaches the depression detection problem as one of information retrieval. The goal is to verify whether users that trigger more symptoms have more chances to be depressed than the ones that present evidence for less or non symptoms.

We ranked users in descending order of the number of detected symptoms and evaluated the precision as a retrieval task (i.e., evaluating the ranked list with relevancy given by the ground truth label). Specifically, the ranking was evaluated with the precision at the k^{th} position ($P@K$). The obtained precision values were as follows: P@10=0.70, P@20=0.75, P@30=0.73. We think these are positive results, as among the top 10, 20 and 30 ranked users the majority were labeled as depressed. This also reinforces evidence on that the higher the number of detected symptoms the higher the chances of the depression category[7].

5.4.3 More and Less Informative Symptoms

In an attempt to further analyze the relevance of each of the symptoms considered, we measured the occurrence of the symptoms in the posts from positive users (i.e., users classified as suffering from depression). Figure 4 summarizes the results of this analysis, showing for each symptom the difference of their occurrences in true-positive and false-positive instances. These results indicate that the symptoms that occur the most in users correctly classified as suffering from depression are s_3 : *past failure*, s_8 : *self-criticalness* and s_7 : *self-dislike*. On the other hand, the symptoms whose presence caused most classification errors are s_{15} : *loss of energy*, s_{19} : *concentration difficulty* and s_{12} : *loss of interest*. These results are not surprising at all, because the first group of symptoms refers to the perception of users about themselves, while the symptoms of the second group refer to more generic moods and problems, which are also widely mentioned by users who do not suffer from depression.

[7]For your reference, we also report the value $P@79 = 0.58$, with 79 being the number of positive cases in the test set.

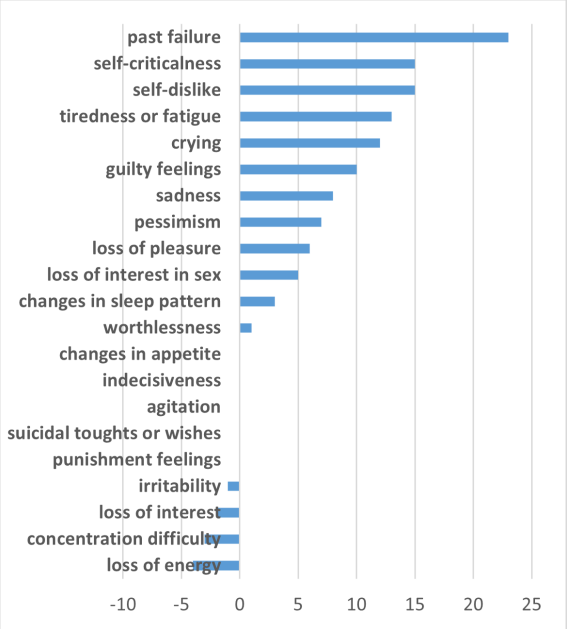

Figure 4: More to less informative symptoms arranged in descending order.

6 *Towards a Monitoring Tool*

The depression detection task has been approached with automatic methods achieving competitive performance. However, some of them are obscure, in the sense that it is not clear what motivates the recommendations of the models, or what information is the most informative for models. Our proposed solution is inspired by the traditional diagnosis process (i.e, through the application of questionnaires), and it is able to generate rich intermediate information, which gives our proposed solution a clear advantage when compared to other models. Here we outline some ways in which one could take advantage of the information that is generated by the proposed solution for explainability and interpretability purposes (see Figure 5) :

- **Thermometer of depression level.** Would allow us to visualize an estimate of the level of depression of a certain user according to the number of detected symptoms by the individual models.
- **List of identified symptoms.** Would allow us to visualize the list of symptoms that were identified as positive in a user.
- **List of evidence posts.** Would allow us to visualize posts that are considered evidence for a selected symptom.

These components and several others

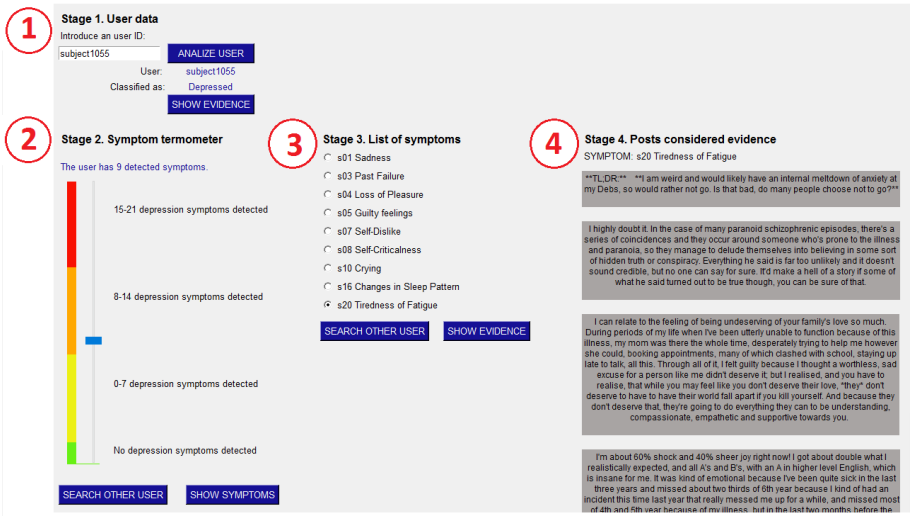

Figure 5: Example of the proposal of a support tool for decision making that takes advantage of the generated information by the proposed method.

could be put together in a decision support tool that could be helpful to psychologist or even the average user to have an idea on her/his potential level of depression. In the remainder of this section we provide a description of a possible monitoring support tool implementing the aforementioned components.

6.1 Running example

Figure 5 shows an screenshot on how the support tool may look when analyzing a particular subject. As a first stage the user data is introduced, for this example *subject1055* from the considered dataset is used, this user is declared as Depressed. The second stage shows the number of symptoms that the user has. In this case for *subject1055* nine symptoms were detected. According to the thermometer scale 9 symptoms correspond to an orange level that can be translated as a medium depression level. The third stage shows a list with the specific name and numbers of the previously detected symptoms. In this case the user presents the symptoms: *sadness, past failure, loss of pleasure, guilty feelings, self-dislike, self-criticalness, crying, changes in sleep pattern* and *tiredness or fatigue*. In the fourth stage, the most relevant posts to symptoms selected from the list are displayed.

Although this is just a sketch of a possible solution, we firmly believe that a tool like this could be very helpful to disentangle and understand the recommendations of the proposed model.

7 *Conclusion & Future Work*

We proposed a novel depression detection method based on the identification of the 21 declared symptoms from the BDI. The method first retrieves evidence related to the symptoms from the users' posts. These evidence is then passed through symptom-detectors, whose outputs are combined to provide a final prediction. The proposed method obtained competitive results when compared with the best eRisk2018 reported results. Even when the proposed method did not outperform the state of the art, its additional benefits compensate for the small performance difference with more sophisticated methods. In particular, the possibility of providing explanations on the predictions of the model, and to interpret the model functioning are its most notable advantages. Future work includes the improvement of the evidence retrieval stage by adopting more elaborated models and by improving the description of symptoms (e.g., exploring other embedding versions, using contextualized text representations or applying pseudo-relevance feedback). Likewise, we are implementing the decision support tool into a demo that could be publicly available to anyone. On the other hand, we plan to carry out a quantitative evaluation of the interpretability of the model applying models such as LIME or SHAP.

References

Aragon, M. E., A. P. Lopez-Monroy, L.-C. G. Gonzalez-Gurrola, and M. Montes. 2021. Detecting mental disorders in social media through emotional patterns - the case of anorexia and depression. *IEEE Transactions on Affective Computing*, pages 1–1.

Beck, A. T., C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. 1961. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571.

Burdisso, S. G., M. Errecalde, and M. Montes-y-Gómez. 2019. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197.

Coppersmith, G., M. Dredze, and C. Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Coppersmith, G., M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Danilevsky, M., K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen. 2020. A survey of the state of explainable ai for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

De Choudhury, M., M. Gamon, S. Counts, and E. Horvitz. 2013. Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.

Guntuku, S. C., D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Husseini Orabi, A., P. Buddhitha, M. Husseini Orabi, and D. 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*, pages 88–97, New Orleans, LA. Association for Computational Linguistics.

Jamil, Z., D. Inkpen, P. Buddhitha, and K. White. 2017. 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*, pages 32–40, Vancouver, BC. Association for Computational Linguistics.

Kuncheva, L. and C. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207.

Liu, N., Z. Zhou, K. Xin, and F. Ren. 2018. TUA1 at eRisk 2018. In *Working Notes of CLEF 2018—Conference and Labs of the Evaluation Forum*.

Losada, D., F. Crestani, and J. Parapar, 2020. *Overview of eRisk 2020: Early Risk Prediction on the Internet*, pages 272–287. 09.

Losada, D. E., F. Crestani, and J. Parapar. 2017. Clef 2017 erisk overview: Early risk prediction on the internet: Experimental foundations. In *CLEF*.

Losada, D. E., F. Crestani, and J. Parapar. 2018. Overview of erisk 2018: Early risk prediction on the internet (extended lab overview). In *CLEF*.

Losada, D. E., F. Crestani, and J. Parapar. 2019. Overview of erisk at clef 2019: Early risk prediction on the internet (extended overview). In *CLEF*.

Loveys, K., J. Torrez, A. Fine, G. Moriarty, and G. Coppersmith. 2018. Cross-cultural differences in language markers of

depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87, New Orleans, LA. Association for Computational Linguistics.

Lundberg, S. M. and S.-I. Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pages 4765–4774.

Mathur, P., R. Sawhney, S. Chopra, M. Leekha, and R. Ratn Shah. 2020. Utilizing temporal psycholinguistic cues for suicidal intent estimation. *Advances in Information Retrieval*, 12036:265–271.

Mowery, D. L., A. Park, C. Bryan, and M. Conway. 2016. Towards automatically classifying depressive symptoms from Twitter data for population health. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 182–191, Osaka, Japan. The COLING 2016 Organizing Committee.

Nadeem, M. 2016. Identifying depression on twitter. *ArXiv*, abs/1607.07384.

National Institute of Mental Health. 2021. Depression. NIH Publication No. 21-MH-8079. https://www.nimh.nih.gov/health/publications/depression.

Ortega-Mendoza, R. M., D. I. Hernández-Farías, M. M. y Gómez, and L. Villaseñor-Pineda. 2022. Revealing traces of depression through personal statements analysis in social media. *Artificial Intelligence in Medicine*, 123:102202.

Parapar, J., P. Martín-Rodilla, D. Losada, and F. Crestani, 2021. *Overview of eRisk 2021: Early Risk Prediction on the Internet*, pages 324–344. 09.

Pennebaker, J. W., M. R. Mehl, and K. G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577.

Preoţiuc-Pietro, D., J. Eichstaedt, G. Park, M. Sap, L. Smith, V. Tobolsky, H. A. Schwartz, and L. Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 21–30, Denver, Colorado. Association for Computational Linguistics.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Ríssola, E., M. Aliannejadi, and F. Crestani. 2020. Beyond modelling: Understanding mental disorders in online social media. *Advances in Information Retrieval*, 12035:296 – 310.

Saloni Dattani, H. R. and M. Roser. 2021. Mental health. *Our World in Data*. https://ourworldindata.org/mental-health.

Trifan, A., R. Antunes, S. Matos, and J. Oliveira. 2020. Understanding depression from psycholinguistic patterns in social media texts. *Advances in Information Retrieval*, 12036:402 – 409.

Trotzek, M., S. Koitka, and C. Friedrich. 2018a. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32:588–601.

Trotzek, M., S. Koitka, and C. Friedrich. 2018b. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. In *CLEF*.

Tsugawa, S., Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3187–3196, New York, NY, USA. Association for Computing Machinery.

Tsugawa, S., Y. Mogi, Y. Kikuchi, F. Kishino, K. Fujita, Y. Itoh, and H. Ohsaki. 2013. On estimating depressive tendencies of twitter users utilizing

their tweet data. In *2013 IEEE Virtual Reality (VR)*, pages 1–4, Los Alamitos, CA, USA. IEEE Computer Society.

Wolohan, J., M. Hiraga, A. Mukherjee, Z. A. Sayyed, and M. Millard. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with NLP. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 11–21, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

World Federation for Mental Health. 2012. Depression: A global crisis. *Health WF for M, editor. Occoquan*, pages 1–32.

World Health Organization. 2003. Investing in mental health. https://apps.who.int/iris/handle/10665/42823.

World Health Organization. 2017. Depression and other common mental disorders: global health estimates. Technical documents.

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

Zogan, H., X. Wang, S. Jameel, and G. Xu. 2020. Depression detection with multi-modalities using a hybrid deep learning model on social media. *CoRR*, abs/2007.02847.