Future Generation Computer Systems 124 (2021) 480–494

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

An emotion and cognitive based analysis of mental health disorders from social media data

Ana-Sabina Uban *, Berta Chulvi, Paolo Rosso

Pattern Recognition and Human Language Technology (PRHLT), Universitat Politècnica de València, València, Spain

ARTICLE INFO

Article history:
Received 28 December 2020
Received in revised form 26 April 2021
Accepted 21 May 2021
Available online 11 June 2021

Keywords:
Mental health disorders
Early risk prediction
Emotions
Cognitive styles
Deep learning
Social media

ABSTRACT

Mental disorders can severely affect quality of life, constitute a major predictive factor of suicide, and are usually underdiagnosed and undertreated. Early detection of signs of mental health problems is particularly important, since unattended, they can be life-threatening. This is why a deep understanding of the complex manifestations of mental disorder development is important. We present a study of mental disorders in social media, from different perspectives. We are interested in understanding whether monitoring language in social media could help with early detection of mental disorders, using computational methods. We developed deep learning models to learn linguistic markers of disorders, at different levels of the language (content, style, emotions), and further try to interpret the behavior of our models for a deeper understanding of mental disorder signs. We complement our prediction models with computational analyses grounded in theories from psychology related to cognitive styles and emotions, in order to understand to what extent it is possible to connect cognitive styles with the communication of emotions over time. The final goal is to distinguish between users diagnosed with a mental disorder and healthy users, in order to assist clinicians in diagnosing patients. We consider three different mental disorders, which we analyze separately and comparatively: depression, anorexia, and self-harm tendencies.

1. Introduction

Mental health disorders are an important and pervasive public health issue. Depression in particular affects approximately 300 million people worldwide [1], and is a major factor leading to suicide, constituting the 3rd leading cause of death for 10–24 year-olds. Moreover, depression is massively underdiagnosed and undertreated, with more than half of the people suffering from depression not receiving any treatment. People affected by mental disorders are often reluctant to approach a specialized clinician to seek help with treating the disorder. However, more and more frequently people turn to social media to discuss their issues and to seek emotional support. This opens up an important opportunity for automatic processing of social media data in order to identify changes in mental health status that may otherwise go undetected before they develop more serious health consequences. Identifying people who start to develop signs of a mental illness in the early stages is very important to managing its evolution, and in certain cases it can be life-saving. The CLEF eRisk Lab,[1] organized every year since 2017, is dedicated specifically to identifying early signs of mental disorders from a user's social media posts, before the user was diagnosed with the disorder, for disorders including depression, anorexia and thoughts of self-harm [2–4].

Recently, the COVID-19 pandemic is expected to exacerbate this problem, affecting mental health as well as physical health [5], through generating feelings of anxiety, depression and even trauma [6–11].

The way mental disorders manifest and can be recognized is primarily through everyday communication. It has been shown in previous computational studies that individuals suffering from mental disorders manifest changes in their language and their behavior, for example, greater negative emotions and high self-attentional focus [12,13]. Whether the topic of their disorder is mentioned or not, the language used by a speaker can contain strong indicators of an altered mental state. These can manifest both explicitly, at the level of the topics approached, or implicitly, at the level of the emotional charge of the text, or more subtle stylistic indicators (such as the increased use of personal pronouns [14]).

Thus, the need for automation is motivated not only by the intractability of manually analyzing the high quantity of data produced daily on social media platforms, but also by the potential of artificial intelligence to exploit large quantities of data in order to uncover implicit markers of mental disorder risk, that can sometimes be difficult to notice even by the patient herself. According

* Corresponding author.
E-mail address: ana.uban@fmi.unibuc.ro (A.-S. Uban).

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

https://doi.org/10.1016/j.future.2021.05.032

to [13], automatic social media-based screening of depression reported so far in literature may reach prediction performance somewhere between unaided clinician assessment and screening surveys. Data from social media, as a very rich and relatively easy to obtain type of data, as well as continuously growing source of real-time information, can thus be leveraged to gain many valuable insights into an individual's behavior and mental state and its evolution. Computational models capable of accurately predicting the development of mental disorders could then be integrated into applications with major social impact such as tools to alert users that they are at risk of developing a disorder or conversational bots. For clinicians, the computational analysis of mental health as seen in social media data could provide data for a deeper understanding of the mental disorders, and possibly lead to the development of improved instruments for diagnosis.

Research in psychology suggests that emotions play a central role in people's cognitive styles.*Cognitive style* or thinking style is a concept used in cognitive psychology to describe the way individuals think, perceive and remember information. Some of these stable ways of thinking have been detected as more prevalent in some patients suffering from depression [15] and anorexia [16]. We are interested in verifying the hypothesis that differences in the use of language connect with some cognitive styles that research has related to mental disorders. For example we know from attribution theory research, that people suffering from depression see the cause of the negative events as more related to themselves and they feel more responsibility for the negative outcomes [17]. Also some work has been done in cognitive styles and anorexia suggesting that measures of cognitive styles can be powerful predictors of treatment response [18]. Our question, from this perspective, is to what extent using social media posts makes it possible to detect different cognitive styles that co-occur with the communication of emotions over time and that vary significantly if we compare people with a mental disorder with negative cases.

Acknowledging the important social impact of modeling mental disorders with computational methods, and taking into account the knowledge provided by psychology in this area, the research questions we aim to answer in this work are the following:

RQ1. *How could monitoring the language used by social media users help with early detection of mental disorders?*

RQ2. *To what extent deep learning techniques can help to interpret the signs of mental disorders?*

RQ3. *Can a connection between cognitive styles and the expression of emotions reveal a specific pattern in users diagnosed with a mental disorder?*

In this study, we bring several contributions to research into the automatic detection of mental disorders, at different levels, with a focus on three particular disorders: depression, anorexia and self-harm. We design models to automatically predict disorders from social media data and study their performance in parallel; to this effect we explore the use of deep learning for detecting mental disorders from text data, and compare various architectures, including hierarchical attention networks and transformers. We model our text data using a multi-aspect representation, through using features that reflect various complementary levels of the language, including content, style and emotions. Our choice of model architecture and features is motivated on one hand by its prediction power (through the complex neural architecture and the capacity to model complex patterns in linguistic data at different levels), and on the other hand by its high interpretability (through the use of features such as emotions and psycho-linguistic categories, and through its hierarchical attention mechanisms). We attempt to interpret the decisions of the developed classifiers, as well as perform feature analysis, in order to obtain a deeper understanding of mental disorder signs. We finally go beyond the static approach, and propose that in order to properly understand the manifestations of mental disorders, we need to approach them from a dynamical perspective, accounting for the evolution of symptoms over time, particularly looking at the way emotions evolve in relation to cognitive styles. Starting from theories in psychology related to mental disorders and emotions, we propose a more sophisticated model of representing manifestations of mental disorders in language, through tracking the evolution of cognitive styles in conjunction with emotions, and reveal association patterns specific to subjects suffering from a mental disorder, as opposed to healthy users.

The paper is structured as follows: Section 2 contains a summary of the existing literature on the computational analysis of mental disorders, and in Section 3 we describe the datasets we used and introduce the set of experiments we performed. In Section 4 we discuss the different features we extract from the data, and explain how we use them to encode the texts for our several experiments. Section 5 describes the classification experiments for mental disorder prediction, including models used and results, and in Section 6 we attempt to explain the model's predictions through examining attention weights, and through feature analysis. Section 7 is dedicated to the study of emotion evolution in conjunction with cognitive styles, and includes the experimental setup and results, as well as detailed interpretations of our findings. We continue with a discussion of our results and their relevance in the wider context of mental health research in Section 8. Finally, in Section 9 we discuss the limitations of our work and suggest directions for future development ; and in Section 10 we present our conclusions, referring back to our original research questions listed in the introduction.

2. Previous work

There is an extensive body of research related to automatic risk detection for mental disorders from social media data [13,19]. The majority of research has focused on the study of depression [20–23], but other mental illnesses have also been studied, including generalized anxiety disorder [24], schizophrenia [25], post-traumatic stress disorder [26,27], risks of suicide [28], anorexia [3] and self-harm [3,29]. Different social media platforms have been considered, such as Twitter [30], Facebook [12], or Reddit [2]. Research works related to depression have considered monitoring already diagnosed patients [23] as well as the detection of users showing signs of depression.

The vast majority of studies provide either quantitative analyses, or predictors built using simple machine learning models [12, 20]. Few studies have made use of more complex deep learning methods such as different types of convolutional (CNN) or recurrent neural networks (RNN) [31–33] or word embeddings [14, 34]. At the level of features, most previous works have used traditional bag of word n-grams [26], while some have also applied more domain-specific representations, such as hand-crafted lexicons [35], linguistic inquiry and word count (LIWC) [36] features [12], or latent semantic analysis (LSA) [35,37].

Among the features used for mental disorder detection in previous literature, LIWC categories, emotions and personal pronouns usage have consistently been shown to be relevant for this task [12–14,20,21,37,38]. Most studies using complex linguistic features such as emotion expression or psycho-linguistic categories based on the LIWC lexicon are either simple quantitative analyses or use classical classifiers such as support vector machines or logistic regression [12,20,37]. To our knowledge, we are the first to use a deep neural architecture including multiple different types of linguistic features in order to model mental disorder manifestations in social media data. Recently, more studies have started employing deep learning for mental disorder

detection, generally using word sequences as features [31–33]. Among these, few studies use hierarchical attention networks, showing their usefulness for this task. In [39], the authors use a hierarchical attention network for the prediction of several mental disorders from a dataset of psychiatric notes, with good results. One study on anorexia detection [40] uses a hierarchical attention network similar to ours (although using only word embedding features) obtaining the best results in the eRisk shared task on anorexia detection [3]. Our choice of models is thus motivated not only by their high complexity and proven success in this area, but also by their higher interpretability through the use of attention, through the hierarchical structure and through the inclusion of multiple types of linguistic features.

Emotions have been previously shown to be relevant for modeling mental disorders, but not many go beyond simple quantitative analyses. We find another approach in previous research that focuses on an in-depth analysis of emotions in order to model mental disorders in language, published in three studies on depression, anorexia and self-harm detection [41–43]. Starting from Plutchik's eight basic emotions [44], the authors use word embedding spaces to automatically identify sub-emotions, which they use as features for their classifiers, trained to detect depression [41], anorexia [42] and self-harm [43] respectively. In our model of emotions, we use the same eight emotions as features, and instead we study their evolution, not in isolation, but in conjunction with cognitive styles, revealing correlation patterns specific to mental disorders.

There are few studies which jointly include in their models different aspects of the language for assessing risk of mental disorders [24,32]. In a previous study [45] on the early detection of self-harm tendencies, we propose using a hierarchical neural network composed of long short-term memory (LSTM) and CNN layers, trained on social media texts represented using content features, as well as emotion and style features. Few studies consider several disorders at the same time [46].

Quantitative analyses in existing research on mental disorders have found that people suffering from depression manifest changes in their language, such as a greater usage of negative emotions and high self-attentional focus [12], or increased first-person pronoun usage [14]. Other computational studies have used topic modeling or word usage analyses to discover increased prevalence of certain topics among depressed users, such as discussing medications or bodily issues such as lack of sleep, or expressing hopelessness or sadness [35,47]. Nevertheless, correlation studies are limited in discovering more complex connections between features of the text and mental disorder risk.

Moreover, most computational studies model mental disorder symptoms as static phenomena, whereas the evolution of mental disorder markers, as well as their prevalence in texts posted by a user, is an important indicator of mental disorder risk. We mention one previous study [48] in which the authors attempt to classify time series representing the mood of social media users in order to predict occurrence of anorexia. The CLEF eRisk Lab [2–4] focuses on early detection of mental disorders, and in the associated shared task, the participating systems are required to process a stream of social media posts ordered chronologically, and are evaluated taking into account the delay with which they manage to identify users who have been diagnosed with a disorder. However, annotations are provided at user level and are static (do not identify the moment of the onset of the disorder), and few submissions attempt to model the input data as a reflection of a dynamical process.

From a practical perspective, models based on neural networks are vastly successful for most NLP applications, even though they have been only briefly explored in existing computational studies on mental disorders. Nevertheless, neural networks are notoriously difficult to interpret. Recently, there is increasing interest in the field of explainability of machine learning methods including in NLP [49], which aims for providing interpretations of the decisions of neural networks. For a mental disorder detection tool whose aim is to assist social media users, it is essential that its diagnosis process is understandable. Moreover, explaining the behavior of powerful classifiers modeling complex patterns in the data has the potential to help uncover signs of the disease that are difficult to observe at a simple glance, and thus assist clinicians in the diagnosis process.

In the field of mental disorder detection, there are not many studies attempting to explain the behavior of classifiers. We note one such example [50], where the authors analyze attention weights of a neural network to show that attention at the user level correlates with the importance of individual texts for classification performance in automatic anorexia detection.

Finally, research on mental disorders from a computational perspective has been generally disconnected from mental health research in psychology, with virtually no computational studies providing in-depth interpretations of their findings from a psychological perspective.

Table 1
Datasets statistics.

Dataset	Users	Positive%	Posts	Words
eRisk depression	1304	16.4%	811,586	25M
eRisk anorexia	1287	10.4%	823,754	~23M
eRisk self-harm	763	19%	274,534	~6M

3. eRisk Reddit datasets on depression, anorexia and self-harm

We perform all our analyses on textual data extracted from social media, containing users suffering from mental disorders, as well as healthy users, annotated for several disorders and manifestations thereof: depression, anorexia and self-harm.

eRisk is a Lab with shared tasks on early prediction of mental disorder risk from social media data. Each year a new task is organized around a specific disorder: in 2017 and 2018 the shared task focused on depression [2], in 2019 and 2020 there were tasks on the prediction of anorexia and self-harm tendencies [3,4]. Each task comes with a new dataset, collected from Reddit posts and comments selected from specific relevant sub-reddits. Users suffering from a mental disorder are annotated by automatically detecting self-stated diagnoses, followed by a manual curation step (not involving clinical experts). Healthy users are selected from participants in the same sub-reddits (having similar interests), thus making sure the gap between healthy and diagnosed users is not trivially detectable. Moreover, a long history of posts are collected for the users included in the dataset, up to years prior to the diagnosis. We use in our experiments all three datasets, corresponding to different disorders: depression, anorexia and self-harm tendencies. Table 1 contains statistics describing all datasets considered.

We propose to conduct a series of experiments in order to answer the questions of predicting and interpreting mental disorders in social media data.

We start by modeling the prediction of mental disorders as a classification problem, where the value to be predicted is a binary label, i.e., whether ot not a certain user has, or she will likely develop, a mental disorder (separately for each disorder we considered: depression, anorexia and self-harm). We build deep learning models and train them on the eRisk datasets to predict mental disorder diagnoses from social media activity, using different neural network-based models. Since the signs of

mental health problems can be complex and manifest in language in different ways, we propose a multi-aspect approach, where we include as input to our model linguistic features at different levels of the language: content, as well as style and emotions.

Beyond showing whether or not it is possible to automatically predict mental disorders on the basis of linguistic data, we try to obtain a deeper understanding of how exactly mental health is reflected in language, and continue our analysis into interpreting the decision-making process of the model, by analyzing the model's behavior (including ablation experiments and attention mechanism analysis), as well as performing a fine-grained feature analysis.

Finally, since a static view of features seen independently is not sufficient to fully understand the complex linguistic and psychological phenomena that occur in problems of mental health, we perform an in-depth qualitative analysis. We carried out further experiments where we correlate the usage over time of emotions and linguistic styles to identify patterns that help to distinguish between users diagnosed with a mental disorder (depression, anorexia, self-harm) and healthy users.

4. Linguistic markers of mental health

As previous studies have shown, mental disorders manifest in language at different levels: topics discussed, emotions conveyed, as well as the author's style. In psychology, the notion of "cognitive style" has been used to explain that it is possible to identify some dimensions in a person's reasoning style that reflects how this individual organizes and structures information. Some research [51] has pointed out that thinking styles are associated with emotions and also that thinking styles had predictive power for emotions beyond age.

At a computational level, we aim to build multi-dimensional representations of the texts in our datasets to account for the different levels of the language where markers of mental illnesses can manifest. We extract various different features from the texts, which will be used as input to our subsequent text classification models, as well as in our other analyses in the following sections.

Content features. We include a general representation of text content by transforming each text into word sequences. The obtained sequences will constitute the main input of the recurrent and convolutional layers of our neural networks. Preprocessing of texts includes lowercasing and tokenizing, and removing punctuation and numbers. Stopwords are not excluded. The most frequent 20,000 words in all datasets were selected to form a common vocabulary. When passed as input to the neural networks, words within a sequence were encoded as embeddings of dimension 300. In order to initialize the weights of the embedding layers, we started from pre-trained GloVe [52] embeddings. In the Appendix we include additional results for our best model using alternative pre-trained embeddings.

Style features. We aim at representing the stylistic level of texts through including stopwords and pronoun features. The usage pattern of stopwords is known to be reflective of an author's style, at a subconscious level. In the study of mental disorders, the increased use of pronouns, especially first person pronouns, has been shown to correlate with mental disorder risk [35]. We include two separate stylistic features: firstly, we extract from each text a numerical vector representing stopword frequencies as bag-of-words, normalized by text lengths. We complement these with other stylistic features extracted from the LIWC lexicon, including syntactic features such as usage of pronouns or other parts of speech, as described below.

LIWC features. The LIWC lexicon [36] has been widely used in computational linguistics as well as in some clinical studies for analyzing how suffering from mental disorders manifests in an author's writings. LIWC is a lexicon mapping words of the English vocabulary to lexico-syntactic features of different kinds, with high quality associations curated by experts in psycholinguistics having the potential to capture different levels of language: including style (through syntactic categories such as "pronoun", "verb", "future", "past" etc.,), emotions (through affect categories such as "sad", "anxiety", "affect", "positive emotion", "negative emotion" etc.) and topics (through content-oriented categories such as words referring to cognitive or analytical processes (e.g. "cogmech", "analytical", "because", "effect"), or to topics such as "money", "health" or "religion"). We use LIWC 2007 [53] and include all 64 categories in the lexicon in our classification experiments, which we represent as a numerical vector for each input example by computing for each category the ratio of words in a text that are related to the category, according to the lexicon (obtaining one vector for each input text, with elements of each vector representing the ratio corresponding to one category in the lexicon).

Moreover, in order to verify our hypothesis that some differences in the use of language are connected with cognitive styles that research has related to mental disorders, we have selected from the LIWC lexicon some categories that could be taken as markers of these cognitive styles. In this sense we use "hear" as a marker that is linked with a demanding communication style. Using words like "listen, hearing, speak" suggests the speaker is making an effort to catch the attention of others. We study "causation", which includes words such "because", "effect" or "hence" that play an important role in attributional style, that is to say, in the way people reason about the causes of events in everyday life. We also use the "I" pronoun category that is a marker of a self centered cognitive style.

Emotions and sentiment. We dedicate a few features to representing emotional content in our texts, since the emotional state of a user is known to be highly correlated with her mental health. Several of the LIWC categories aim to capture sentiment polarity and emotion content, from more general ones, such as *negative emotion*, *positive emotion*, *affective processes*, to more specific emotions, such as *sadness* and *anxiety*. We additionally include a second lexicon, the NRC emotion lexicon [54], which is dedicated exclusively to emotion representation, containing eight categories corresponding to a more fine-grained selection of emotions, based on Plutchik's [44] eight basic emotions: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *trust*, along with two additional categories corresponding to *negative* and *positive* sentiment, respectively. We represent NRC features similarly to LIWC features, as numerical vectors of ratios for each text, and include all NRC features in the classification experiments.

In our subsequent analyses, when we try to test how cognitive styles interact with emotions in users with some mental disorder, we have used only the NRC lexicon for emotions, and kept the LIWC lexicon only for marking the cognitive styles. With the NRC lexicon we work only with concrete emotions: four negatives ("anger", "disgust", "fear" and "sadness") and three positives ("joy", "trust" and "anticipation"), going beyond the simple "negative" and "positive" categories because in the context of mental health these big clusters give few orientations to clinical practice. For the same reason, we have also not used "surprise" as emotion because it does not have a clear valence in terms of positive and negative.

5. Automatically predicting mental disorders

In the following section we describe the models and features used for our classification experiments. We experimented with different architectures, which were trained and evaluated separately on each dataset.

5.1. Classification experiments

All tasks approached are modeled as binary classification tasks, with labels at the user level. We follow the typical machine learning workflow, and split each of the datasets into training, validation and test subsets. We maintain the original train/test split provided by the shared task organizers, to ensure consistency and comparability with previous results, even though the test sets are sometimes larger than the training sets. We make sure to split train and test sets at the user level: no texts in the training set have common authors with texts in the test set. In this way we ensure models are not learning any user-specific artifacts, and instead are modeling markers of the disorders themselves.

Social media posts are not considered individually as datapoints, since they are too short to be sufficiently predictive. Instead, we generate our datapoints by grouping sequences of 50 chronologically consecutive posts into larger chunks, to obtain more substantial samples of text. As we will show in the following subsection, we consider two types of architectures for our models: a sequential and a hierarchical setup. To each of these corresponds a different representation of the input features. In the sequential setup, chunks of user posts are simply concatenated and considered together as one long text. The obtained concatenated sequences are truncated at 512 words. Bag-of-word features are calculated at chunk-level, as are numerical features (LIWC, NRC emotions and first person pronoun relative frequencies). Previous studies [48] have shown that the evolution of mental disorder markers, as well as their prevalence in texts posted by a user, is an important indicator of mental disorder risk. Inspired by this observation, we consider a more sophisticated representation of user posts, by modeling them as a hierarchical structure, such that the post level and the user level are considered separately. This will allow us to include targeted attention mechanisms, to weigh separately words within a post and posts within a user's history. In the hierarchical setup, posts within a chunk are stacked to form a hierarchical structure: one datapoint will consist of on group of posts corresponding to one user, as a stack of word sequences of 256 words. Bag-of-words and numerical features also follow a hierarchical structure, with a set of features extracted for each post in the group, and stacked together into bi-dimensional vectors.

Being the eRisk datasets unbalanced in positive versus negative examples, we use a weighted loss function to compensate for the minority class; the weights are calculated according to the distribution of the class labels.

For evaluating the performance of our models, we compute precision, recall, and F1-scores, where a prediction is considered positive if the network output exceeds a threshold, established for each task by finding the value that maximizes F1 on the validation sets. Being the datasets unbalanced, we additionally compute the area under the ROC curve (AUC), which is less sensitive to imbalance, and will be the main metric we consider when comparing performance between different models and datasets.

5.2. Models

In previous literature on automatic prediction of mental disorders, when neural networks are employed their architectures are pretty simple [24,55]. We propose exploring more sophisticated deep architectures, including RNNs and CNNs, as well as hierarchical attention networks and transformers. Each of the architectures used are briefly described below; the detailed configurations of each architecture are included in the Appendix.

5.2.1. BiLSTM with attention

For the sequential setup, we use a traditional bidirectional LSTM network. Word sequences (truncated/padded at 512 words), with words encoded as embeddings, are passed as input to the BiLSTM layer, which is then fed to an attention layer. The bag-of-stopwords features are passed through a dense layer; and the remaining extracted features (including pronouns, emotions and LIWC categories usage) are concatenated into one vector. The output of the BiLSTM is concatenated along with the other features and the final representation passed through an output layer that generates the final prediction.

5.2.2. Hierarchical attention network

Hierarchical attention networks for text classification were introduced in [29] where they were used for review classification, by representing a text as a hierarchical structure. We propose that social media data in our setup are very well suited to such a hierarchical representation; in our case the hierarchy consists of user post histories, which are composed of social media posts, composed of word sequences. Especially since the evolution of the mental state of a user is in itself a relevant indicator for the development of a disorder [40,48], user-level representations are expected to be natural and useful for modeling this problem. This structure will also allow us to include hierarchical attention mechanisms, to weigh separately words within a post and posts within a user's history, which is useful especially since the evolution of the mental state of a user can be in itself a relevant indicator for the development of a disorder [40,48]. Another study has included post-level and user-level attention for their classifier's architecture, obtaining top results in the anorexia detection shared task [40].

The hierarchical network is composed of two components: a *post-level encoder*, which produces a representation of a post, and a *user-level encoder*, which generates a representation of a user's post history.

Post-level encoder. For encoding the word sequence at post-level, we experiment with two different neural architectures: LSTM with attention, and CNN.

We denote the layer encoding the word sequences (LSTM/CNN) as $\overrightarrow{f}$, the input word sequence as $w_{it}, t \in [1,T], i \in [1,C]$, where T is the number of words in a post, C is the number of a user's posts in an input chunk, and H is the dimensionality of the post-level encoder (number LSTM units or CNN feature maps). The embedding matrix is denoted as W_e, E is the dimensionality of the embedding space, and V is the size of the vocabulary.

In the LSTM setup, the post-level attention weights α_{it} are applied to the resulted hidden layer, obtaining the final representation (s_i) of the input word sequence:

$$
\begin{gathered}
x_{it} = W_e w_{it}, t \in [1,T], i \in [1,C], w_{it} \in \mathbb{N}^V, W_e \in \mathbb{R}^{E \times V}, x_{it} \in \mathbb{R}^E \\
\overrightarrow{h}_{it} = \overrightarrow{f}(x_{it}), \overrightarrow{h}_{it} \in \mathbb{R}^{T \times H} \\
v_{it} = tanh(W_w h_{it} + b_w), W_w \in \mathbb{R}^{H \times 1}, b_w \in \mathbb{R}^H, v_{it} \in \mathbb{R}^T \\
\alpha_{it} = \frac{exp(v_{it}^T v_w)}{\sum_t exp(v_{it}^T v_w)}, v_w \in \mathbb{R}^H, \alpha_{it} \in \mathbb{R}^{T \times H} \\
s_i = \sum_t \alpha_{it} h_{it}, s_i \in \mathbb{R}^H
\end{gathered}
\tag{1}
$$

The bag-of-stopwords feature f_{it} and the lexicon features l_{it} for each post are passed to dense layers (with parameters W_f, b_f and W_l, b_l respectively, where F denotes the number of stopwords considered, L denotes the number of lexicon features considered, and D denotes the number of units in the dense layers used to

encode the bag-of-stopwords and lexicon features), and concatenated to the word sequence representations to obtain the post encoding p_i:

$$\begin{aligned} hf_{it} &= W_f f_i + b_f, hf_{it} \in \mathbb{R}^D, f_i \in \mathbb{R}^F, W_f \in \mathbb{R}^{F \times D}, b_f \in \mathbb{R}^D \\ hl_{it} &= W_l l_i + b_l, hl_{it} \in \mathbb{R}^D, l_i \in \mathbb{R}^L, W_l \in \mathbb{R}^{L \times D}, b_l \in \mathbb{R}^D \\ p_i &= s_i \oplus hf_{it} \oplus hl_{it}, p_i \in \mathbb{R}^{H+2D} \end{aligned} \quad (2)$$

For the CNN post-level encoder, the pooling layer output is used to encode the word sequences, and further concatenated with the other features' representations, to obtain the post-level encoding.

User-level encoder. Each of the posts in the input datapoint is encoded with the post-level encoder, and then they are stacked to form the bi-dimensional representation s_i, which is then concatenated with encodings of the other features, and passed to the user-level encoder, modeled as an LSTM with attention. User-level attention weights are denoted by α_i and the final user representation is u, where U is the number of LSTM units in the user-level encoder. The output of the user encoder is connected to the final output layer which generates the prediction.

$$\begin{aligned} h_i &= \overrightarrow{LSTM}(p_i), h_i \in \mathbb{R}^{C \times U} \\ v_i &= tanh(W_p h_i + b_p), W_p \in \mathbb{R}^{U \times 1}, b_p \in \mathbb{R}^U, v_i \in \mathbb{R}^L \\ \alpha_i &= \frac{exp(v_i^T v_p)}{\sum_t exp(v_i^T v_p)}, v_p \in \mathbb{R}^U, \alpha_i \in \mathbb{R}^{C \times U} \\ u &= \sum_i \alpha_i h_i, i \in [1, C], u \in \mathbb{R}^U \end{aligned} \quad (3)$$

A depiction of the hierarchical architecture is shown in Fig. 1.

In all architectures we use batch normalization and L_2 regularization. Binary cross-entropy is used as a loss function. For training our models we use the Adam optimizer. Other hyperparameters are chosen through hyperparameter tuning on the validation set. We provide a full list of the hyperparameter values, hidden layer dimensions and training setup configurations in the Appendix.

5.2.3. Baselines

Transformers. We experiment with state-of-the-art language models based on transformer architectures, which have been shown to obtain high performances on a wide range of NLP tasks, with minimal task-specific training. We use pre-trained models for English (the "base" versions of the models) and fine-tune them for our task. For the transformer models we only use word sequences as features. The architectures we experiment with include BERT [56] with its newer variants RoBERTa [57] and AlBERT [58].

Logistic regression. As a baseline, we use a simple logistic regression model with bag-of-word features, using the same vocabulary and pre-processing as for the deep learning models.

We are making all the code used for our experiments public via a github repository.[2]

5.3. Results

We first assess the performance of each of our proposed models on each of the tasks and datasets individually. Table 2 shows a summary of all results, in terms of F1 and AUC score. Additional metrics are listed in the Appendix. For most tasks, the best performing model is the hierarchical attention network with LSTM post-level and user-level encoders. Even with little task-specific training, the high performance of the transformer models

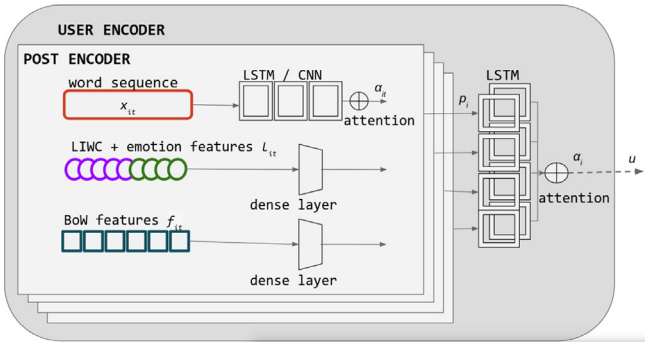


Fig. 1. Hierarchical attention network architecture.

Table 2
F1 and AUC scores for all datasets and models.

Model	Self-harm eRisk		Anorexia eRisk		Depression eRisk	
	F1	AUC	F1	AUC	F1	AUC
BiLSTM seq	.62	.84	.53	.90	.40	.82
LSTM hierarch (HAN)	.65	**.87**	.61	**.96**	.45	**.83**
CNN + LSTM hierarch	.44	.82	.76	.95	.35	.80
RoBERTa	.35	.60	.70	.83	.40	.71
AlBERT	.22	.55	.65	.77	.20	.61
LogReg	.45	.75	.49	.91	.36	.76

proves they are powerful for across tasks, including for mental disorder detection.

Anorexia in particular seems to be the easiest disorder to classify among the ones considered, with up to 0.96 AUC, while depression is the most difficult to detect. This might be due to the nature of the disorder and its symptoms, but could also be due to data-specific aspects: among the Reddit datasets, the anorexia one contains the most training data.

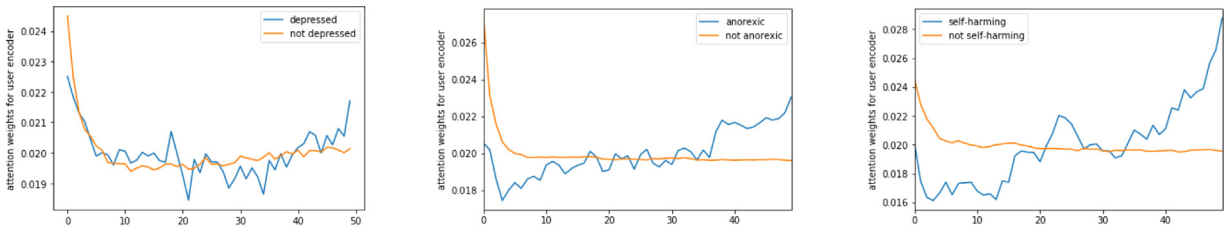
According to [13], previous studies in automatic detection of depression report AUC scores between 0.6 and 0.9. Among more recent studies, not many report AUC. Submissions to eRisk 2018 using the same dataset for depression detection report a best score of 0.64 F1-score. Anorexia detection was approached within the 2019 eRisk shared task [3], with the best team obtaining a 0.70 F1-score [40]. We obtain superior results in terms of F1 with the CNN+LSTM model. Most previous literature on self-harm detection consists of the solutions submitted in the eRisk 2019 and 2020 shared tasks [3,4], where the best team obtained 0.75 F1 scores using an augmented training dataset, and the second best team, using the simple eRisk dataset, obtained 0.62 F1, which we surpass with our HAN model.

While our models obtain competitive results on these datasets across the three mental disorders, our main focus in this paper is not surpassing state-of-the-art results in mental disorder detection, but rather providing a deeper understanding of the way they manifest in language, backed by theories in psychology. Our choice of model facilitates interpretation through its hierarchical attention mechanisms as well as the varied set of linguistic features, and the high classification performance suggests that a deeper analysis of its behavior could help uncover interesting patterns on mental disorder manifestations in language which the models were able to capture through training. We continue with our analyses of the signs of mental disorders in language in the following sections.

6. Explainability of predictions

We have shown in the previous sections that it is feasible to train a deep learning classifier to automatically detect mental

[2] https://github.com/ananana/mental-disorders/.



(a) Depression (b) Anorexia (c) Self-harm

Fig. 2. Attention weights for user encoder.

health disorders. If any automatic solution is to be used for this purpose, it is essential that its decision-making process is understandable in the name of transparency. Especially in the medical domain, using black-box systems can be dangerous for patients and is not a realistic solution [59,60]. Moreover, recently, the need of explanatory systems is required by regulations like the General Data Protection Regulation (GDPR) adopted by the European Union. Additionally, the behavior of powerful classifiers modeling complex patterns in the data has the potential to help uncover manifestations of the disease that are potentially difficult to observe with the naked eye, and thus assist clinicians in the diagnosis process.

In this section, we attempt to obtain a deeper understanding of the behavior of the trained models, through different explainability techniques including attention weight analysis, ablation experiments, and analysis of interpretable features such as emotions and psycho-linguistic categories. For these analyses, we focus on interpreting the behavior of our best-performing model in terms of AUC, that is the HAN.

6.1. Attention analysis

In neural networks with complex architectures, where multiple different components interact, it is difficult to explain the network's behavior through straightforward techniques such as comparing model weights or gradient values.

We have chosen to use a hierarchical attention model in order to separately represent words within posts and posts within user histories. The good results of this model compared to the simpler models suggest that indeed the model succeeds in a better representation of the data. This architecture gives us the opportunity to investigate this hypothesis, through analyzing the distribution of attention weights.

In order to get some insight into the inner workings of our model, we extract the attention weights in the attention layer, and interpret them as indicators of the importance of each corresponding dimension of the input. At the post encoder level, attention can be interpreted as word importance within individual posts, whereas at the user level, attention weights would approximate the importance of a given post among a chunk of several posts made by the same user. It should be noted that the user encoder includes multiple different features (word sequences, lexicon features and bag-of-stopwords), and all of them contribute to the attention weight assigned to a given post.

Figs. 3a, 3b, 3c show examples of comments posted by users suffering from one of the disorders, with words highlighted in blue according to the attention weights given by the trained networks, and with post-level attention marked in red. When plotting the word-level attention intensity, we weigh word-level attention by multiplying it with the post-level attention for the corresponding post. We compensate by normalizing the word-level attention scores by the text length when illustrating attention for words, in order to make word-level comparison more visible across different posts, since longer posts tend to be awarded more attention by the model (sometimes leading to longer posts being highlighted from beginning to end).

At the user level, we try to understand how attention weights distribute across the attention layer by plotting the weights along the length of each 50-post chunk, averaged for all chunks across all users. The resulted curves, seen in Fig. 2, show a slight increasing trend, suggesting later posts in the user history tend to be more useful for classification. For anorexia and self-harm, the increasing trend only appears in users with a positive label, which might suggest that they show increasing symptoms of the disorder (from the perspective of the features extracted by the neural network), in comparison with the healthy users who are more stable across time.

Leveraging attention mechanisms through using trained attention weights as a way to interpret feature importance has been used before in deep learning, including in one study related to anorexia [50]. Nevertheless, recent studies [61–63] have questioned whether attention mechanisms necessarily help with the interpretability of the relationship that exists between the weights and the final prediction, and suggested attention weights should not be strictly interpreted as measuring the importance of features in the model's decision, but merely as indicating a correlation. For this reason we argue a more in-depth analysis of the features at different levels is necessary for a reliable conclusion on linguistic markers of mental disorders.

Table 3
F1 and AUC scores for ablation experiments.

Dataset	Self-harm		Anorexia		Depression	
Features	F1	AUC	F1	AUC	F1	AUC
All - Word sequences	.34	.83	.48	.91	.34	.79
All - LIWC	.62	.87	.45	.91	.48	.82
All - NRC emotions	.59	.86	.45	.94	.43	.80
All - Stopwords	.55	.84	.47	.95	.50	.81
All features	.65	.87	.61	.96	.45	.83

6.2. Ablation experiments

In order to understand the relative importance of the features used to represent the text data, corresponding to the different levels of the language, we train and evaluate our classifier separately excluding each of the features, through ablation experiments. Through these experiments, we can gain insight into the relative importance of each feature for predicting mental disorders, and further, into the levels of the language (topics, style, emotions, etc.) where mental disorder symptoms manifest.

Results of the ablation studies are shown in Table 3, with performance measured in terms of F1-score AUC. Additional metrics are provided in the Appendix.

We find that, for all datasets, using all features leads to the best prediction performance, suggesting that each of the features we include has a relevant contribution to detecting signs of

Fig. 3. Texts posted by users suffering from a mental health disorder, highlighted according to attention weights.

mental disorders, including the lexicon features and the bag-of-stopwords feature, which are rarely used in deep learning models for mental disorder detection in previous research. Aside from word sequences, which have the biggest impact on performance, stopwords seem to play an important role on self-harm detection, suggesting that subtle stylistic markers, beyond explicit mentions of emotions or topics, have discriminative power. For anorexia and depression, LIWC features play a more important part compared to other features.

6.3. Lexicon-based feature analysis

In order to analyze the importance of the lexicon-based features (LIWC categories and NRC emotions), we compute for each feature the Spearman correlation between the numerical score obtained for the feature and the label, averaged across all posts for each user. In this way, the higher correlation scores can be interpreted as proportionally indicative of the relevance of the feature for the specific mental disorder. We compute the correlations at user-level: by aggregating the metrics for all posts belonging to a user into a single score for each feature, through simple averaging. Table 4 shows some of the categories with the highest (in absolute value) correlations with the positive label, for each dataset.

We find that people diagnosed with a mental health disorder tend to use more personal pronouns, especially first person pronouns, which is consistent with results reported in other works [35]. Additionally, we find that people suffering from a disorder speak less about categories such as *work*, *leisure*, *money* and *death*, consistently across all three disorders.

While these results allow us to already observe some different patterns between healthy subjects and ones suffering from mental disorders, we argue that such an analysis is limited from various perspectives. First of all, mental disorder symptoms develop over time, and the onset of a disorder can be better understood in terms of evolution of its signs rather than their overall occurrence, measured statically. Secondly, the signs of a disorder at the level of emotions and cognitive styles, is best characterized by the combination of these features. The prevalence of an emotion is not always in itself indicative of a disorder, but rather its

Table 4
Features highly correlated with mental disorder labels.

Depression	Feature	Work	Leisure	Money	Death	Health	Feel	You	Negate	Social	Cogmech	Verb	Present	Pronoun	I
	Correlation	−0.19	−0.19	−0.17	−0.12	0.10	0.15	0.20	0.21	0.22	0.32	0.36	0.36	0.39	0.46
Anorexia	Feature	Work	Leisure	Article	Money	Social	Disgust	Feel	Anxiety	Adverb	Funct	Bio	Health	Ppron	I
	Correlation	−0.20	−0.19	−0.19	−0.16	0.13	0.14	0.25	0.25	0.33	0.35	0.38	0.44	0.45	0.51
Self-harm	Feature	Work	We	Leisure	Positive	Negemo	Sadness	Health	Adverb	Present	Future	Cogmech	Funct	Pronoun	I
	Correlation	−0.15	−0.11	−0.10	−0.9	0.12	0.14	0.29	0.33	0.34	0.37	0.43	0.43	0.43	0.51

systematic association with certain topics and ways of thinking is how mental health issues are manifested.

To overcome these limitations, in the following section we propose a deeper analysis considering the joint evolution of emotions and cognitive styles in healthy and users diagnosed with a mental disorder.

7. Linguistic styles and emotions

We further move on from the static perspective and propose to look into the evolution of feature scores across time. This analysis is motivated by the hypothesis that the pattern of evolution of markers for mental disorders is an independent indicator of mental health issues. In this section, we present experiments designed with the aim to understand if the evolution patterns differ for users diagnosed with a mental disorder differ from healthy users.

In 7.1 we describe the technical details of the method used. In 7.2 we explain the criteria used to select some features to focus on, and discuss their correlations with emotions. Finally, in 7.3–7.5 we explain some patterns observed in these features.

7.1. Evolution of features over time

The datasets we used for this study facilitate this kind of temporal analysis, since, for each user, posts are collected for up to several years before the onset of the disease — thus making it possible to monitor how predictive symptoms develop up to and beyond the point of diagnosis.

We compute the evolution of markers of mental disorders, for the three datasets, with a focus on emotions. Since different users have activity over different periods of time, user histories ordered by the calendar date are not comparable, and averaging across users based on this timeline would potentially introduce noise or mask true patterns manifesting for users individually. We thus attempt to obtain a better representation of user histories and group posts according to the days passed since the first appearance of the user in the dataset, and use this number of days as the common reference between users when we average across users. We separate users into two groups according to their label: suffering from a mental disorder or healthy, and aggregate metrics across all users for each group, applying a moving average of 100 days to smooth out the measurements across time.

We model the interactions over time between emotions and linguistic styles by computing the Pearson correlation between the obtained evolution time series of emotions, on the one hand, and linguistic styles, on the other hand. We use the NRC lexicon to extract emotion prevalence, and LIWC categories to model cognitive styles. An analysis of the overlap between emotion words for categories common to the two lexicons (“negative emotion”, “positive emotion”, “anger”, “sadness”) shows that, on average across these categories, a proportion of 20% of the words in our vocabulary which appear the NRC lexicon are also found in the LIWC lexicon, and 44% of the words in the LIWC lexicon for the common categories also occur in the NRC lexicon.

In order to minimize dependence between features, we exclude the emotion features in LIWC from this analysis, and additionally only correlate features derived from NRC with features derived from LIWC (and avoid comparing two features extracted using the same lexicon).

We note that by correlating the time series across emotions and linguistic styles, we should also capture some information about co-occurrence: for example, if an emotion and a topic tend to co-occur in the same texts, this should also imply a higher correlation between their evolution patterns. The chosen method allows us to find more stable relationships that a simple co-occurrence analysis might not show due to the noise induced by the small size of individual texts.

For each emotion-style pair, we compute the correlation coefficient as well as the significance value. For each disorder, we compute all correlations separately for the healthy and diagnosed users, and then compare the correlations obtained in each group. We assess the significance of the difference between the groups using the z-test statistic. We select only the significant differences ($p < 0.05$) and show them in Table 5. The Appendix contains a full list of correlation coefficients and significance values for all pairs of emotions and linguistic styles.

The first thing that we observe is that there is an important number of linguistic features from LIWC that correlate with emotions in a significantly different way between people that suffer a mental disorder and people that do not. In Table 5 numbers indicate the amount of linguistic features from LIWC (excluding six LIWC categories that denote emotions or are related to emotions: *sadness*, *anxiety*, *positive emotion*, *negative emotion* and *affect*) that present a correlation with emotions which differs significantly for people with a mental disorder diagnosis and the negative cases.

As we can see in Table 5, different emotions are more relevant for each mental disorder. For depression, *anticipation* (57 features present significant differences) and *trust* (54 features) are the emotions that best capture the differences between positive and negative cases. For anorexia, it is *fear* (55 features) and *anger* (54 features); and the same emotions are the most relevant for self-harm tendencies: *fear* (47 features) and *anger* (45 features).

7.2. Language in use and evolution of emotions over time

Our aim is to find some common pattern in the communication of emotion that differs between people that suffer a mental disorder and the control cases. In order to find this common pattern among mental disorders, we disregard some features that are closely related to a specific mental disorder, e.g. “bio” with anorexia or “death” with self-harming tendencies. However, the number of features that correlate with emotions in a significantly different way for the two groups of users poses an added difficulty, then we decide to focus our attention on features that meet four criteria: (1) the correlation feature/emotion is distinctly different between the users that suffer a mental disorder and the negative cases ($p < 0.0001$); (2) there is a certain consistence of this pattern among the different emotions, and we find these differences in more that one emotion; (3) the correlation goes in the opposite direction between users suffering from a mental disorder and the control cases; and (4) psychological research offers a theoretical framework that points out the relevance of the psychological process involved in the use of this feature. In this last criterion we focus on communication styles and attribution

Table 5
Linguistic features from LIWC that correlate differently with emotions between people that suffer a mental disorder vs. negative cases.

	Anger		Disgust		Fear		Sadness		Trust		Anticip.		Joy	
Significant (S)/not(N)	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS	S	NS
Depression	34	24	47	11	48	10	44	14	54	4	57	1	49	9
Anorexia	54	4	51	7	55	3	52	6	45	13	48	10	50	8
Self-harm	45	13	45	13	47	11	41	17	44	14	43	15	43	15

Table 6
Correlation between "hear" and emotions in the three mental disorders, for positive users (diagnosed with a mental disorders) and negative ones (healthy). Only correlations which are significantly different between the positive and negative classes are shown.

	Anger		Disgust		Fear		Sadness		Trust		Anticip.		Joy	
pos(P)/neg(N)	P	N	P	N	P	N	P	N	P	N	P	N	P	N
Depression	−0.56	0.25	−0.45	0.48	−0.47	0.07	−0.29	0.08	0.25	0.35	0.35	−0.28	0.19	−0.13
Anorexia	0.19	−0.13	0.17	−0.08	0.33	−0.03	–	–	0.16	−0.42	0.28	−0.04	0.51	0.31
Self-harm	0.33	0.02	0.60	−0.26	–	–	0.08	−0.11	0.55	0.11	0.55	−0.15	0.70	−0.05

styles [15–17] and a self-centered cognitive style based on the use of first person pronoun [13,26,64], three psychosocial processes closely related to mental health.

In addition to the four criteria mentioned above, we have decided to focus the analysis on features whose definition in the LIWC dictionary is sufficiently clear to be related to an underlying psychological process. For example, the feature "inhibition" in the LIWC dictionary presents the same pattern in users suffering from depression as the feature "causation" but the meaning of "causation" is more directly related to a psychological process studied by the scientific literature [15,16] than that of "inhibition"; therefore, we decided to select "causation". The same occurs with "see" and "hear". The three features selected for a cross-sectional analysis between the different emotions and the three mental disorders are "hear", "causation" and the use of the pronoun "I" from LIWC.

We analyze the correlation of these features with negative (*anger*, *fear*, *disgust* and *sadness*) and positive emotions (*joy*, *anticipation* and *trust*). Although negative emotions are more relevant when we speak about depression, anorexia and self-harm tendencies, including the positive emotions as well is interesting because, as we will see, some linguistic features co-evolve in a different way with positive vs. with negative emotions.

7.3. Demanding communication style: Can you hear me?

To study a demanding communication style we used the *"hear"* category from LIWC that includes words like "listen, hearing, speak", and describes a user that is making an effort for drawing attention to themselves. This feature correlates with emotions in a significantly different way between people that suffer from a mental disorder and those who do not. They also vary with respect to the disorders (see Table 6).

Depression. In the case of depressed people, the less they use these categories, the more expressions of negative emotion we find. The opposite happens with not depressed people. We will go deeper into the interpretation of these results in Section 8, but it seems that depressed people give up trying to draw the attention to themselves when negative emotions are more present in their discourse. The same occurs with *trust*, but in the other two positive emotions (*anticipation* and *joy*) we find the opposite pattern: depressed people are those using this demanding communication style more than the healthy people when they are communicating positive emotions.

Anorexia. We find a consistently different pattern in the use of the feature *"hear"* and the expression of emotions in people with a diagnosis of anorexia: patients suffering from anorexia draw the attention to themselves more than healthy users, in both negative and positive emotions.

Self-harm tendencies. Among users with self-harm tendencies we find the same pattern as in anorexia: people with self-harm tendencies use *"hear"* in correlation with both negative and positive emotions while healthy users do not.

7.4. Attributional style and negative emotions

Another LIWC category related to "cognitive process" is *"causation"*, that plays an important role in cognitive attributional style. This category presents differences in correlation with emotions in people suffering from mental disorders and healthy users (Table 7).

Depression. Depressed people speak about causes more when more negative emotions are communicated and the opposite happens to not depressed people: the less they speak about causes, the more negative emotions they express. We find the same pattern in the expression of *trust*, but the opposite when expressing *joy*: depressed people speak less about causes when they are communicating joy than not depressed people do.

Anorexia. People with a diagnosis of anorexia speak about causes more when more negative emotions are communicated and the opposite happens to not diagnosed people: the less they speak about causes, the more negative emotions they express. It is the same pattern that we found in depressed people but in this mental disorder it additionally occurs with *sadness*. With *trust* we find the same pattern as with negative emotions, while we do not find a clear pattern with other positive emotions.

Self-harm tendencies. Among the patients with self-harm tendencies we find the same pattern that we observe in depressed and anorexic patients, but only in relation with anger: the more they speak about causes, the more anger they communicate, while the opposite occurs in the group of users without self-harm tendencies. On the contrary, this does not happen in the communication of disgust. With fear and sadness we find the opposite relation: the less people with self-harm tendencies speak about causes, the more fear and sadness they communicate, in comparison with those without self-harm tendencies.

7.5. Self-centered cognitive style

Although we can find a well established pattern in computational linguistics when we look at the use of the *"I"* pronoun [35], we believe that it is important to see how it is used in correlation with the expression of different emotions (see Table 8).

Depression. The use of *"I"* and personal pronouns in general present differences in the correlation with all positive emotions between depressed and not depressed people: the more depressed people use *"I"* and personal pronouns, the more they express positive emotions like *joy*, *anticipation* and *trust*, and the

Table 7
Correlation between "causation" and emotions in the three mental disorders for positive users (diagnosed with a mental disorders) and negative ones (healthy). Only correlations which are significantly different between the positive and negative classes are shown.

	Anger		Disgust		Fear		Sadness		Trust		Anticip.		Joy	
pos(P)/neg(N)	P	N	P	N	P	N	P	N	P	N	P	N	P	N
Depression	0.39	−0.36	0.33	−0.21	0.46	−0.43	–	–	−0.06	−0.31	–	–	−0.23	−0.03
Anorexia	0.48	0.07	0.40	0.02	0.39	0.04	0.45	−0.38	0.41	0.16	−0.13	−0.23	−0.08	−0.55
Self-harm	0.25	0.12	–	–	0.15	0.27	−0.15	0.03	–	–	0.23	−0.17	0.26	−0.19

Table 8
Correlation between the use of "I" and emotions in the three mental disorder for positive users (diagnosed with a mental disorder) and negative ones (healthy). Only correlations which are significantly different between the positive and negative classes are shown.

	Anger		Disgust		Fear		Sadness		Trust		Anticip.		Joy	
pos(P)/neg(N)	P	N	P	N	P	N	P	N	P	N	P	N	P	N
Depression	–	–	–	–	−0.11	−0.29	0.25	−0.04	0.15	−0.15	0.58	−0.62	0.50	−0.06
Anorexia	0.12	−0.16	0.08	−0.22	0.30	−0.16	0.24	0.06	0.27	−0.25	0.53	0.24	0.72	0.55
Self-harm	0.42	0.01	0.34	0.13	0.21	−0.28	0.34	−0.06	–	–	−0.16	0.31	−0.05	0.40

opposite happens with not depressed people. However, there is not a significant difference in the expression of this self-centered pattern in relation with negative emotions, with the exception of *fear*.

Anorexia. In patients with a diagnosis of anorexia we find a consistent difference across the expression of all emotions: if we compare them with not diagnosed people, the more they use the *"I"* pronoun, the more they express their emotions.

Self-harm tendencies. Among patients with self-harm tendencies we find the same pattern as in anorexia, but only in relation to negative emotions: the more they use *"I"*, the more they express negative emotions. However, the opposite happens with positive emotions: in this case the correlation of *"I"* with the expression of positive emotions is more characteristic to people without self-harm tendencies.

8. Discussion

The success of the hierarchical attention network for the prediction of all three disorders suggests that a hierarchical representation of social media activity is useful for this task (**RQ1**). While the performance of transformers is high given the small amount of parameters trained for this task, for most experiments we obtain superior results with our own models, implying the additional features we use for them (stopwords, emotions and LIWC features) have an important contribution for correctly classifying users, as confirmed by the ablation experiments. One factor which could contribute to the difficulty of an accurate detection is that, in the datasets we used, the set of healthy subjects is selected among active users in the same sub-reddits as the users diagnosed with the mental disorders, that is, discussing the same topics. This could mean that we cannot rely on the topics discussed by the users to easily distinguish between the two groups, and that capturing the implicit markers of the development of the disorders is indeed necessary.

The datasets we use contain long histories (up to several years) of social media activity for each user, possibly including texts written before the users were diagnosed with the mental disorder, which makes it suitable for evaluating early detection and for analyses of the evolution of symptoms over time. At the same time, it poses a challenge for the accurate classification of early posts, and possibly affects the performance of the classifiers. The analysis of user-level attention can be seen as a confirmation that later posts are considered by the classifier more representative of the true label of the user (**RQ2**).

The values of attention weights at the user level show an increasing trend over time for users suffering from anorexia and self-harm, in contrast to healthy users, suggesting a potential progression of the disorder that is captured by the trained hierarchical networks. We notice a different pattern in depression, where we see some weight accumulated towards the beginning of the post history as well, and a less clear distinction between depressed and healthy users.

This coincides with a lower classification performance in the case of depression as compared to the other disorders, which suggests the automatic discrimination between depressed and healthy people (especially when they are discussing the same topics, by design of the dataset) is a more difficult task. Compared to other disorders, the language used to refer to depression tends to be more common in everyday speech, and is used more frequently in a casual way. Statements about "being depressed" and other depression-related vocabulary can often be used improperly by people who were not diagnosed with depression, which can cause ambiguity and possibly lead to misclassifications in classifiers trained for this task.

The initial analysis of lexicon features presented in Section 6 shows a consistent pattern that crosses the three pathologies: people who suffer from these mental disorders write less about categories that LIWC classifies as personal concerns: *work*, *leisure*, *money* and *death*. These categories refer to acting in the exterior world and this is consistent with results in psychological research on attributional styles and depression [17]. Attributional style is a concept used in psychology to analyze the way people explain events in everyday life. Results from this area identify external attribution (finding the causes of events in the outside world) as less related to depression than internal attribution (finding the causes of events in oneself). We also found that the LIWC categories that have a significant correlation with a positive label in all three mental health disorders are *health*, *future*, *pronoun* and *I* (first person pronouns), four dimensions that refer more to internal attribution than to the external one. According to these results, in terms of improving early automatic detection of sings of mental disorders, the proper question might be what people do *not* talk about. In order words, which is the everyday language that we do not find in people who have been diagnosed with depression, self-harm or anorexia.

In psychology, in the 90's emotions begin to be conceptualized not as an internal, subjective experience but as a type of interpersonal and social phenomenon that occurs in communication [65]. We think our results support this view of emotions. As we show in Table 5 we have found that between 58% and 98% of 58 LIWC features interact with emotions in a significantly different way for people suffering from a mental disorder versus the negative cases (**RQ3**). This has a major relevance in therapeutic contexts that try to help people to deal with mental disorders in everyday life. As we already know that mental health benefits from

narrative [66], finding common patterns in the communication of emotions among people that suffer from some mental disorder is crucial to improve diagnosis, but also to understand how the mental pathology evolves in order to design therapeutic strategies that could help people with their suffering.

According to cognitive styles, we think that it is interesting that for people with a diagnosis of anorexia and with self-harm tendencies, in comparison with healthy users, when they communicate the majority of emotions, they tend to use more words that are meant to draw the attention of the other people. This result provides us with valuable information about how important active listening is in the treatment of these mental disorders. Among depressed people we find this pattern of a demanding communication style only in positive emotions. In our opinion an explanation of this is that depression is sometimes accompanied by bipolar disorders and in the euphoric phases, people suffering from this mental disorder try to call the attention of others to their achievements and projects. This gives us also a useful insight about how difficult it is to have an active communication with depressed people when they are expressing negative emotions.

Our results on attributional style and negative emotions confirm previous psychological research in attributional theory and depression (**RQ3**): users speaking about causes in relation with negative emotions endorse what has been defined as the "depressed attributional style" [67]. Explaining events is a common way to overcome difficulties in everyday life, and this could be the explanation of the results obtained for not depressed people. When not depressed people do not speak about causes, they express more negative emotions because they do not find the way to explain what is going on. With depressed people a major therapeutic difficulty is that they do not use attributional explanations in a healthy and self-protective way, but in the opposite sense. An unexpected result is that, to some extent, people suffering from anorexia show a similar pattern as depressed people with regards to causation and negative emotions, and this is interesting because it suggests that more research in attributional styles and negative emotions could give some new insights for the diagnosis and treatment of this mental disorder. We do not find a clear pattern in relation to self-harm tendencies.

In relation with self-centered cognitive style and emotions, our results confirm that the use of *"I"* in interaction with the expression of emotions could provide a deep understanding of mental disorders, because there are different patterns for positive and negative emotions that need to be taken into account.

9. Limitations and ethical concerns

Some limitations we find in the datasets we employed in this work is the lack of information about gender and age of the users.

Another clear limitation imposed by the datasets is the semi-automatic procedure used for annotation, based on self-stated diagnoses, without confirmation from medical experts. Using a dataset which relies directly on medical records would provide more reliable labels, although this might lead to additional limitations on the available amount and diversity of subjects.

Our analysis of emotion evolution in relation to linguistic styles relies on two lexicons (NRC and LIWC). For future developments we think that it could be interesting to include another lexicon for emotion extraction as a way to confirm we find the same patterns, in order to demonstrate a consistent clear difference in the way that cognitive styles interact with emotions in people with a mental disorder in comparison with the negative cases.

In this work, we use LIWC categories to model cognitive styles, and to represent certain topics discussed in the texts (such as *work*, *leisure*, *health* etc.), some of which show interesting correlations with emotions (cf. the Appendix). Another possible choice for identifying topics would be using topic modeling techniques such as Latent Dirichlet Allocation; we will explore this possibility in the future.

While in the final part of this work we propose a novel approach based on a dynamic view of a mental disorder development by tracking the evolution of symptoms over time, this analysis is limited by the static nature of the labels. Having available information on the specific moment of the diagnosis along the history of activity would allow us to develop experiments attempting to predict the moment of the onset of the disorder, and potentially identify the specific patterns leading to the diagnosis. Moreover, our emotion analysis relies on average values across users from a certain group, which might reduce the observed effects if the users in the group show different patterns of evolution (for example if they are at different stages of their disorder). Performing the same analysis at the level of individual subjects, and exploring whether we can cluster users on the basis of their activity in an unsupervised way might be an interesting future direction for our research. Additionally, it would be interesting to model the evolution of emotions as a time series, and explore whether we can automatically detect significant trends or seasonality patterns for users suffering from mental disorders, as opposed to healthy ones.

In Section 7 we mentioned that the results of our correlation analysis reveal a considerable amount of linguistic styles which manifest differently in users suffering a mental disorder and the negative cases, as shown in full in the Appendix. As previously explained, we have used statistical and theoretical criteria to select which correlations between features and emotions to analyze, although more research is needed in this regard.

Our analysis of emotions is novel because of the way we view them in conjunction with cognitive styles, as a dynamical process. The amount of significant findings encourage us to continue in future work with a more in-depth analysis that needs to look at the different mental disorders separately. In our opinion the important contribution of these experiments is to show that the correlation between features and emotions gives some valuable information when we try to use social media data for early detection of mental disorders. We think also that the correlations that differ from positive cases and negative ones could help clinicians to understand better the nature of each mental disorder and to encourage the development of new insights in psychology.

Previous work [41] shows that the eight basic emotions in NRC could further be split into more fine-grained sub-emotions. This previous study reports automatically discovering hundreds of sub-emotions. Considering this high number we did not attempt to include them in our current analysis; nevertheless, we consider the study of sub-emotions an interesting avenue; as future work we intend to investigate the feasibility of including them in our model of emotions to detect mental disorders.

Finally, as previously discussed, we believe our models and findings can be very valuable to potential patients suffering from mental disorders, but any deployment of a tool for mental disorder detection should take into account potential ethical concerns. If such tools are used by third parties (such as employers seeking to filter candidates based on their mental health profile), this could compromise the privacy of the subjects. We suggest that the development of an ethical standard is necessary, such that launching such tools could be accompanied by an ethical statement to constrain its users.

10. Conclusions

Mental disorders are an important and sensitive issue of public health. Using sophisticated computational solutions which leverage large amounts of data to assist with early detection of mental

health issues can be very valuable, but such systems must be developed with corresponding attention to the complexity of the signs of mental health disorders, and to the impact they can have on the life of individuals. This is why a deep understanding of how mental disorders develop is necessary, from the early stages, accounting for explicit and implicit signs of the disorder, for their evolution and interactions over time. The way emotions are communicated, in their interaction with cognitive styles of people suffering from mental disorders, can be particularly important for understanding mental health and the way it is mirrored in language, and particularly in the language used in social media.

We have presented a study on the automatic detection of mental disorders based on linguistic markers extracted from social media, targeting three different disorders (depression, anorexia and self-harm). We showed that deep learning models can be useful for successfully detecting social media users who risk developing a mental disorder (**RQ1**), where the texts they post online are represented with linguistic features at different levels, including the content of the message, but also the user's writing style and the emotions conveyed. We provided interpretations of the model's decisions through the analysis of the attention mechanism together with the features used (**RQ2**); particularly an in-depth analysis of how emotions are used in relation to cognitive styles. We have shown that a hierarchical attention network using features representing different levels of the language can out-perform transformer-based baselines; it would be interesting in future work to explore an architecture which combines the power of our network with pre-trained transformers, for example by encoding sentences or words using representations extracted from pre-trained transformer models such as BERT [56].

In order to understand mental disorders and their early signs in language, we showed the importance of monitoring the language of the affected subjects from the early stages of the onset of the disease, and model the evolution of linguistic markers over time. Emotions are particularly relevant for mental health, but beyond simply focusing on negative sentiment, which is a common approach in NLP, we propose it is more meaningful to treat individual emotions separately, and relate them to specific cognitive styles of subjects, as reflected in linguistic markers. Through the computational analysis of these multi-dimensional features of language over time, we discovered distinct patterns that separate healthy users from users suffering from or developing a mental disorder (**RQ3**). Moreover, we interpreted our findings in an in-depth qualitative analysis grounded in well-established theories in psychology. We suggest that linking computational findings to theoretical grounding provided by research in psychology is essential for more refined computational and linguistic models of mental health, and that such approaches have high potential for improving the performance of automatic tools for monitoring mental health, as well for aiding clinicians to better understand mental disorders.

Our analysis of emotion evolution has confirmed that the temporal aspect of mental disorder symptoms manifested in language is an important one in order to understand linguistic markers of mental health in more depth. Building a dataset with temporally-aware annotations would be a valuable direction for future research.

Beyond the field of mental disorder detection, our model of emotion evolution over time in relation to different psycho-linguistic categories could be relevant for sentiment analysis and for the general study of semantic change in language. Previous studies on language change have analyzed the evolution of word valence over time [68], showing that some words can change polarity over time, and that negative words tend to change their meaning faster than positive words (such as in the case of the word "terrific"). Including fine-grained emotions in relation to other psycho-linguistic categories in such an analysis could provide additional insights into the use of emotion words in language and their evolution.

CRediT authorship contribution statement

Ana-Sabina Uban: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Visualization. **Berta Chulvi:** Conceptualization, Methodology, Formal analysis, Writing - original draft. **Paolo Rosso:** Conceptualization, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

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

The authors thank the EU-FEDER Comunitat Valenciana 2014–2020 grant IDIFEDER/2018/025. The work of Paolo Rosso was in the framework of the research project PROMETEO/2019/121 (DeepPattern) by the Generalitat Valenciana.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.future.2021.05.032.

References

[1] WHO World Health Organization, Depression: A Global Crisis. World Mental Health Day, October 10 2012, World Federation for Mental Health, Occoquan, Va, USA, 2012.
[2] D.E. Losada, F. Crestani, J. Parapar, Overview of erisk: early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2018, pp. 343–361.
[3] D.E. Losada, F. Crestani, J. Parapar, Overview of erisk 2019 early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 340–357.
[4] D. Losada, F. Crestani, J. Parapar, Overview of erisk 2020: Early risk prediction on the internet, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, Vol. 2696, CEUR-WS.Org, 2020.
[5] S.A. Lee, A.A. Mathis, M.C. Jobe, E.A. Pappalardo, Clinically significant fear and anxiety of COVID-19: A psychometric examination of the coronavirus anxiety scale, Psychiatry Res. (2020) 113112.
[6] C.K.T. Lima, P.M. de Medeiros Carvalho, I.d.A.S. Lima, J.V.A. de Oliveira Nunes, J.S. Saraiva, R.I. de Souza, C.G.L. da Silva, M.L.R. Neto, The emotional impact of coronavirus 2019-nCoV (new coronavirus disease), Psychiatry Res. (2020) 112915.
[7] K. Shah, D. Kamrai, H. Mekala, B. Mann, K. Desai, R.S. Patel, Focus on mental health during the coronavirus (COVID-19) pandemic: applying learnings from the past outbreaks, Cureus 12 (3) (2020).
[8] J. Shigemura, R.J. Ursano, J.C. Morganstein, M. Kurosawa, D.M. Benedek, Public responses to the novel 2019 coronavirus (2019-nCoV) in Japan: Mental health consequences and target populations, Psychiatry Clin. Neurosci. 74 (4) (2020) 281.
[9] J. Torales, M. O'Higgins, J.M. Castaldelli-Maia, A. Ventriglio, The outbreak of COVID-19 coronavirus and its impact on global mental health, Int. J. Soc. Psychiatry (2020) 0020764020915212.
[10] Y.-T. Xiang, Y. Yang, W. Li, L. Zhang, Q. Zhang, T. Cheung, C.H. Ng, Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed, Lancet Psychiatry 7 (3) (2020) 228–229.
[11] J. Qiu, B. Shen, M. Zhao, Z. Wang, B. Xie, Y. Xu, A nationwide survey of psychological distress among chinese people in the COVID-19 epidemic: implications and policy recommendations, Gener. Psychiatry 33 (2) (2020).
[12] M. De Choudhury, S. Counts, E.J. Horvitz, A. Hoff, Characterizing and predicting postpartum depression from shared facebook data, in: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, 2014, pp. 626–638.

[13] S.C. Guntuku, D.B. Yaden, M.L. Kern, L.H. Ungar, J.C. Eichstaedt, Detecting depression and mental illness on social media: an integrative review, Curr. Opin. Behav. Sci. 18 (2017) 43–49.
[14] M. Trotzek, S. Koitka, C.M. Friedrich, Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia., in: L. Cappellato, N. Ferro, J. Nie, L. Soulier (Eds.), CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, Vol. 2125, CEUR-WS.Org, 2018.
[15] G.J. Haeffel, B.E. Gibb, G.I. Metalsky, L.B. Alloy, L.Y. Abramson, B.L. Hankin, T.E. Joiner Jr, J.D. Swendsen, Measuring cognitive vulnerability to depression: Development and validation of the cognitive style questionnaire, Clin. Psychol. Rev. 28 (5) (2008) 824–836.
[16] M. Wolf, J. Sedway, C.M. Bulik, H. Kordy, Linguistic analyses of natural written language: Unobtrusive assessment of cognitive style in eating disorders, Int. J. Eat. Disord. 40 (8) (2007) 711–717.
[17] P.D. Sweeney, K. Anderson, S. Bailey, Attributional style in depression: A meta-analytic review., J. Personal. Soc. Psychol. 50 (5) (1986) 974.
[18] P. Butow, P. Beumont, S. Touyz, Cognitive processes in dieting disorders, Int. J. Eat. Disord. 14 (3) (1993) 319–329.
[19] R.A. Calvo, D.N. Milne, M.S. Hussain, H. Christensen, Natural language processing in mental health applications using non-clinical texts, Nat. Lang. Eng. 23 (5) (2017) 649–685.
[20] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: Seventh International AAAI Conference on Weblogs and Social Media, 2013.
[21] J.C. Eichstaedt, R.J. Smith, R.M. Merchant, L.H. Ungar, P. Crutchley, D. Preoţiuc-Pietro, D.A. Asch, H.A. Schwartz, Facebook language predicts depression in medical records, Proc. Natl. Acad. Sci. 115 (44) (2018) 11203–11208.
[22] N.F. Abd Yusof, C. Lin, F. Guerin, Analysing the causes of depressed mood from depression vulnerable individuals, in: Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017), 2017, pp. 9–17.
[23] A.H. Yazdavar, H.S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, A. Sheth, Semi-supervised approach to monitoring clinical depressive symptoms in social media, in: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, 2017, pp. 1191–1198.
[24] J.H. Shen, F. Rudzicz, Detecting anxiety through reddit, in: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—from Linguistic Signal to Clinical Reality, 2017, pp. 58–65.
[25] M. Mitchell, K. Hollingshead, G. Coppersmith, 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, pp. 11–20.
[26] G. Coppersmith, M. Dredze, C. Harman, Quantifying mental health signals in Twitter, in: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2014, pp. 51–60.
[27] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, M. Mitchell, 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, pp. 31–39.
[28] B. O'dea, S. Wan, P.J. Batterham, A.L. Calear, C. Paris, H. Christensen, Detecting suicidality on Twitter, Internet Interv. 2 (2) (2015) 183–188.
[29] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.
[30] X. Chen, M.D. Sykora, T.W. Jackson, S. Elayan, What about mood swings: Identifying depression on twitter with temporal measures of emotions, in: Companion Proceedings of the the Web Conference 2018, 2018, pp. 1653–1660.
[31] F. Sadeque, D. Xu, S. Bethard, Uarizona at the CLEF erisk 2017 pilot task: linear and recurrent models for early depression detection, in: CEUR Workshop Proceedings, Vol. 1866, NIH Public Access, 2017.
[32] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, W. Zhu, Depression detection via harvesting social media: A multimodal dictionary learning solution., in: IJCAI, 2017, pp. 3838–3844.
[33] Y.-T. Wang, H.-H. Huang, H.-H. Chen, A neural network approach to early risk detection of depression and anorexia on social media text, in: L. Cappellato, N. Ferro, J. Nie, L. Soulier (Eds.), CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, Vol. 2125, CEUR-WS.Org, 2018.
[34] A.H. Orabi, P. Buddhitha, M.H. Orabi, D. Inkpen, Deep learning for depression detection of twitter users, in: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, 2018, pp. 88–97.
[35] M. Trotzek, S. Koitka, C.M. Friedrich, Linguistic metadata augmented classifiers at the CLEF 2017 task for early detection of depression, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, Vol. 1866, CEUR-WS.Org, 2017.
[36] J.W. Pennebaker, M.E. Francis, R.J. Booth, Linguistic inquiry and word count: LIWC 2001, Mahway: Lawrence Erlbaum Assoc. 71 (2001) (2001) 2001.
[37] P. Resnik, A. Garron, R. Resnik, Using topic modeling to improve prediction of neuroticism and depression in college students, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1348–1353.
[38] M. Conway, D. O'Connor, Social media, big data, and mental health: current advances and ethical implications, Curr. Opin. Psychol. 9 (2016) 77–82.
[39] T. Tran, R. Kavuluru, Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks, J. Biomed. Inform. 75 (2017) S138–S148.
[40] E. Mohammadi, H. Amini, L. Kosseim, Quick and (maybe not so) easy detection of anorexia in social media posts, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, Vol. 2380, CEUR-WS.Org, 2019.
[41] M.E. Aragón, A.P. López-Monroy, L.C. González-Gurrola, M. Montes, 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, pp. 1481–1486.
[42] M.E. Aragón, A.P. López-Monroy, M. Montes-y Gómez, INAOE-CIMAT at eRisk 2019: Detecting signs of anorexia using fine-grained emotions, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, Vol. 2380, CEUR-WS.Org, 2019.
[43] M.E. Aragón, A.P. López-Monroy, M. Montes-y Gómez, INAOE-cimat at erisk 2020: Detecting signs of self-harm using sub-emotions and words, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, Vol. 2696, CEUR-WS.Org, 2020.
[44] R. Plutchik, A general psychoevolutionary theory of emotion, in: Theories of Emotion, Elsevier, 1980, pp. 3–33.
[45] A.S. Uban, P. Rosso, Deep learning architectures and strategies for early detection of self-harm and depression level prediction, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, Vol. 2696, CEUR-WS.Org, 2020.
[46] A. Yates, A. Cohan, N. Goharian, Depression and self-harm risk assessment in online forums, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 2968–2978.
[47] H.A. Schwartz, J. Eichstaedt, M. Kern, G. Park, M. Sap, D. Stillwell, M. Kosinski, L. Ungar, Towards assessing changes in degree of depression through facebook, in: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2014, pp. 118–125.
[48] W. Ragheb, J. Azé, S. Bringay, M. Servajean, Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, Vol. 2380, CEUR-WS.Org, 2019.
[49] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2018, pp. 80–89.
[50] H. Amini, L. Kosseim, Towards explainability in using deep learning for the detection of anorexia in social media, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2020, pp. 225–235.
[51] L.-F. Zhang, Thinking styles and emotions, J. Psychol. 142 (5) (2008) 497–516.
[52] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[53] J.W. Pennebaker, R.L. Boyd, K. Jordan, K. Blackburn, The Development and Psychometric Properties of LIWC2015, Technical Report, 2015.

[54] S.M. Mohammad, P.D. Turney, Nrc Emotion Lexicon, Vol. 2, National Research Council, Canada, 2013.
[55] A. Benton, M. Mitchell, D. Hovy, Multi-task learning for mental health using social media text, 2017, arXiv preprint arXiv:1712.03538.
[56] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT (1), 2019.
[57] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv preprint arXiv:1907.11692.
[58] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, 2019, arXiv preprint arXiv:1909.11942.
[59] C. Zucco, H. Liang, G. Di Fatta, M. Cannataro, Explainable sentiment analysis with applications in medicine, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2018, pp. 1740–1747.
[60] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable AI systems for the medical domain?, 2017, arXiv preprint arXiv:1712.09923.
[61] S. Jain, B.C. Wallace, Attention is not Explanation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 3543–3556.
[62] S. Serrano, N.A. Smith, Is Attention Interpretable? in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 2931–2951.
[63] S. Wiegreffe, Y. Pinter, Attention is not not Explanation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 11–20.
[64] S. Rude, E.-M. Gortner, J. Pennebaker, Language use of depressed and depression-vulnerable college students, Cogn. Emot. 18 (8) (2004) 1121–1133.
[65] L.K. Guerrero, P.A. Andersen, M.R. Trost, Communication and emotion: Basic concepts and approaches, in: Handbook of Communication and Emotion, Elsevier, 1996, pp. 3–27.
[66] J.W. Pennebaker, J.D. Seagal, Forming a story: The health benefits of narrative, J. Clin. Psychol. 55 (10) (1999) 1243–1254.
[67] J. Crocker, L.B. Alloy, N.T. Kayne, Attributional style, depression, and perceptions of consensus for events., J. Personal. Soc. Psychol. 54 (5) (1988) 840.
[68] W.L. Hamilton, K. Clark, J. Leskovec, D. Jurafsky, Inducing domain-specific sentiment lexicons from unlabeled corpora, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, Vol. 2016, NIH Public Access, 2016, p. 595.

Ana-Sabina Uban received her Ph.D. in computer science from the University of Bucharest in 2020, and she is currently a researcher at Universitat Politècnica de València, working on problems of natural language processing and machine learning. She is also a member of the Human Language Technologies Research Centre in Bucharest, where she has previously published papers on topics in computational linguistics such as stylistics, language change and abusive language in social media. Recently, she has participated in the eRisk 2020 shared tasks on self-harm detection, and on the prediction of the level of depression of social media users.

Berta Chulvi received a Ph.D. on Social Psychology from Universitat de València, Spain. She is assistant professor at the Department of Social Psychology of this University since 2009. Her first degree was in Communication Science and she has used this interdisciplinary approach to research the meeting points between everyday language, common sense theories, and social psychology. Her areas of research are racism, xenophobia, and occupational health. She has published peer reviewed papers in journals as British Journal of Social Psychology and International Review of Social Psychology. She joined recently the PRHLT research center at Universitat Politècnica de València.

Paolo Rosso is full professor at the Universitat Politècnica de València (Spain) where he is also member of the Pattern Recognition and Human Language Technology (PRHLT) research center. His research interests focus on natural language processing and information retrieval topics, related with the analysis of information in social media. Since 2009 he was involved in the organization of PAN benchmark activities, mainly at CLEF and at FIRE evaluation forums. He also helped in the organizations of shared tasks at SemEval (in 2015 and 2019), at Evalita (2014, 2018, 2020) and since 2017 at IberLEF (previously IberEval). Since 2019 he is co-ordinator of the benchmark activities of the IberLEF evaluation forum (previously in 2017 and 2018 of IberEval). He has been chair at Evalita (2018), and overall track co-ordinator at FIRE (2018 and 2019). He serves as deputy steering committee chair for the CLEF conference and as associate editor for the Information Processing & Management journal. He has been PI of several research projects, national and international, funded by EU, Army Research Office (US), and Qatar National Research Fund (QNRF) in collaboration with Carnegie Mellon University. At the moment he is the PI of the 3-year MISMIS-FAKEnHATE research project on MISinformation and MIScommunication in social media. He is the author of 400+ research papers.