Network Modeling Analysis in Health Informatics and Bioinformatics (2020) 9:22
https://doi.org/10.1007/s13721-020-0226-0

ORIGINAL ARTICLE

Multi-modal social and psycho-linguistic embedding via recurrent neural networks to identify depressed users in online forums

Anu Shrestha[1] · **Edoardo Serra**[1] · **Francesca Spezzano**[1]

Received: 4 January 2020 / Revised: 4 March 2020 / Accepted: 8 March 2020 / Published online: 31 March 2020

Abstract
Depression is the most common mental illness in the US, with 6.7% of all adults experiencing a major depressive episode. Unfortunately, depression extends to teens and young users as well and researchers have observed an increasing rate in recent years (from 8.7% in 2005 to 11.3% in 2014 in adolescents and from 8.8 to 9.6% in young adults), especially among girls and women. People themselves are a barrier to fighting this disease as they tend to hide their symptoms and do not receive treatments. However, protected by anonymity, they share their sentiments on the Web, looking for help. In this paper, we address the problem of detecting depressed users in online forums. We analyze user behavior in the ReachOut.com online forum, a platform providing a supportive environment for young people to discuss their everyday issues, including depression. We propose an unsupervised technique based on recurrent neural networks and anomaly detection to detect depressed users. We examine the linguistic style of user posts in combination with network-based features modeling how users connect in the forum. Our results on detecting depressed users show that both psycho-linguistic features derived from user posts and network features are good predictors of users facing depression. Moreover, by combining these two sets of features, we can achieve an F1-measure of 0.64 and perform better than baselines.

Keywords Depression detection · Online forums · Multi-modal user representation · Unsupervised classification · Recurrent neural networks

1 Introduction

Depression is a mental illness commonly seen in people (6.7% of all U.S. adults have experienced at least one major depressive episode), which negatively affects their thoughts and behaviors. Depression causes mood fluctuations and impermanent emotional responses to the challenges of everyday life. Especially when lasting for a while and with moderate or severe intensity, depression may become a serious health condition. It can cause the affected person to suffer greatly and perform poorly at work, at school, and in the family. It has been one of the common problems seen in tens of millions of people. At its worst, depression can lead to suicide. Close to 800,000 individuals die due to suicide every year. According to a 2015 report by the World Health Organization, more than 300 million people are affected by depression. Unfortunately, depression extends to teens and young users as well and researchers have observed an increasing rate in recent years (from 8.7% in 2005 to 11.3% in 2014 in adolescents and from 8.8 to 9.6% in young adults), especially among girls and women. Very few people in the world receive the treatments provided for depression. In many countries, fewer than 10% of people in need receive such treatments. One of the barriers to this is the people themselves. They tend to hide their symptoms to avoid being known as psychiatric patients or because people are unaware of the condition and what is happening with them. Online forums and social media are platforms where

This paper is an extended version of the conference paper "Anu Shrestha and Francesca Spezzano, Detecting Depressed Users in Online Forums. In Proceedings of the International Symposium on Network Enabled Health Informatics, Biomedicine and Bioinformatics (HI-BI-BI 2019)", in conjunction with ASONAM 2019. Shrestha and Spezzano (2019).

✉ Francesca Spezzano
francescaspezzano@boisestate.edu

Anu Shrestha
anushrestha@u.boisestate.edu

Edoardo Serra
edoardoserra@boisestate.edu

[1] Boise State University, Boise, ID 83702, USA

people, protected by anonymity, can share their thoughts freely and publicly and look for help. Thus, the content of online posts is a valuable source of information to analyze to infer the presence of mental illness in these users and take timely actions.

In this paper, we address the problem of detecting users at risk of depression in online forums. Indeed, online posts provide a means to infer an individual's mood and socialization behavior. Our research contributes to automatically retrieving forum users that are potentially at risk and suggest them to the forum administrators for further investigation so that they can promptly act to take care of these people, eventually.

We formulate the problem as a binary classification task and use unsupervised techniques with (1) psycho-linguistic features describing the linguistic style of the user posts and the emotions expressed in them and (2) user networking behavior in the "who replies to whom" network extracted from the forum posts. Specifically, we propose a multimodal methodology where a user embedding is first computed from the sequence of their posts via recurrent neural networks in an unsupervised fashion and, then, combined with user networking behavior. Finally, unsupervised anomaly detection is performed on these features to classify users as depressed or not.

We test our approach on a dataset extracted from ReachOut.com: an Australian non-profit online forum established in 1996 to support young people in addressing problems common to their generation, including alcohol and drug addiction, gender identity, sexuality, and mental health concerns. This dataset is made available by CLPsych'17 shared task. Related work on this dataset has analyzed user posts to automatically triage them by their risk of being written by users suffering from depression Cohan et al. (2016), Yates et al. (2017). Our results on detecting depressed users show that both psycho-linguistic features derived from user posts and network features are good predictors of users facing depression. Moreover, by combining these two sets of features, we can achieve an F1-measure of 0.64 and perform better than baselines.

The paper is organized as follows. Section 2 summarizes related work, Sect. 3 describes the dataset we used in this paper, Sect. 4 presents our proposed unsupervised technique to identify depressed users, Sect. 5 reports on our experimental evaluations and, finally, conclusions are drawn in Sect. 6.

2 Related work

Researchers have been analyzing the online behavior of users in social media to detect depression. Resnik Resnik et al. (2015) studied topic models in the analysis of linguistic signals for detecting depression. These depression detection efforts demonstrated that it is possible to analyze depressed users on social media on a large scale. Preliminary research done by Park et al. (2012) explored the use of language to describe depression utilizing real-time moods captured from Twitter users. Further, Park et al. (2013) conducted face-to-face interviews with 14 active Twitter users to explore their behavior. They found that depressed users perceive Twitter as a tool for social awareness and emotional interaction. Using social network and linguistic patterns, Xu and Zhang (2016) attempted to explain how Web users discuss depression-related issues. They found that depressed users have an intensive use of self-focus words and negative effect words. Zimmermann et al. (2017) looked at how first-person pronoun use might be a predictor of future depressive symptoms. Computerized analysis of written text through LIWC features has also been applied to understand predictors of neurotic tendencies and psychiatric disorders Rude et al. (2004).

De Choudhury et al. (2013) explored the potential of using Twitter to detect and diagnose major depressive disorders in an individual. Thus, to detect depressed users, they considered both linguistic and network features and achieved an F1-measure of 0.68. Similarly to our work, they found out that depressed users on social media exhibit lower reciprocity and a higher clustering coefficient than non-depressed ones (cf. Sect. 4.2), but observed a different posting activity as measured by the insomnia index (cf. Sect. 3.1.1). To predict depression, Eichstaedt et al. Eichstaedt et al. (2018) performed a linguistic analysis of the history of Facebook statuses posted by patients visiting a large urban academic emergency department.

De Choudhury et al. (2013) proposed a statistical metric named Social Media Depression Index (SMDI), which is used to predict indicative depressive posts on Twitter and also helps to categorized depression levels. MacAvaney et al. (2018) addressed the problem of detecting posts on a dataset of annotated Reddit posts by including temporal information about the diagnosis and achieved an F1-measure of 0.55.

Many researchers have used the CLPsych'17 dataset (the same we use in this paper) to perform linguistic analysis of the user posts and triage them by author level of risk. These works have used TF-IDF weighted unigrams, post embeddings using sent2vec Le and Mikolov (2014), LIWC lexicon, as a measure of emotion Cohan et al. (2016) and sentiment Malmasi et al. (2016). Other works leveraged DepecheMood Staiano and Guerini (2014) to identify emotions associated with a post and the MPQA subjectivity lexicon Wilson et al. (2005) to distinguish between objective and subjective posts.

Yates et al. (2017) addressed both the post and user detection problems via deep learning-based text classification. They used a Reddit dataset for user classification and achieved an F1-measure of 0.65. For the user post

detection problem, they used the ReachOut CLPsych'16 dataset (a previous version of the dataset we use in this paper) and achieved the best F1-measure of 0.61, while the F1-measure for other linguistic methods they compared with ranged between 0.53 and 0.5 Kim et al. (2016); Malmasi et al. (2016); Brew (2016); Cohan et al. (2016). Yates et al. also tested their proposed methodology on the ReachOut CLPsych'17 dataset (the same we use in this paper) and reported an F1-measure of 0.50 for post detection.

Overall, several works have addressed the problem of (1) detecting depressed users in social media platforms such as Twitter and Reddit and (2) identifying posts that may indicate a risk of depression. Experimental results reported in these works show that both user and post detection are challenging problems. Moreover, previous work has focused on supervised detection, while in this paper, we propose an unsupervised technique to identify depressed users in online forums using both psycho-linguistic and network features.

3 The ReachOut forum

ReachOut.com[1] is an Australian non-profit online forum available for free, which is well reached to all the common people in Australia [1.58 million of visitors each year Millen (2015)]. This forum provides mental health services along with information and environment to support the youth of age 14–25 so that they can share their mental issues and experiences anonymously. Based on the communications through posts, young people are provided with resources, help, and proper guidance from well-trained moderators. The practical support and tips provided by this organization make it easier for parents to help their children facing mental illness.

In 2013, a survey was conducted among the users of the forum, showing that 33% of Australian young people are aware of the site and proving that the forum was beneficial in supporting people with mental disorders Metcalf and Blake (2013). The survey results showed that "77% of participants reported experiencing high or very high levels of psychological distress" and that 46% of these distressed visitors "were more likely to seek help from at least one professional source after visiting ReachOut."

3.1 Dataset

We used the dataset provided by the 2017 CLPsych shared task http://clpsych.org/shared-task-2017/ containing labeled forum posts from the ReachOut.com platform. The dataset contains a total of 147,619 forum posts, out of which 1,588 were manually annotated by three separate judges according to the following categories (indicating how urgently the post requires moderator's attention):

1. ***Crisis*** indicates that the author is at imminent risk of being harmed, themselves, or others. It should be prioritized above all others.
2. ***Red*** indicates that a moderator should respond to the post as soon as possible.
3. ***Amber*** indicates that a moderator should address the post at some point, but they do not need to do so immediately.
4. ***Green*** identifies posts that do not require direct input from a moderator and can safely be left for the wider community of peers to respond.

Moreover, the annotators added further information regarding the motivation of why the post may or may not need attention, according to the flowchart shown in Fig. 1. We considered the types of annotated posts marked with the tick symbol in the above figure as posts dealing with users who have a mental disease. Therefore, as our task is user-oriented: among all the users who authored at least one annotated post, we consider a user as depressed if they have posted or commented at least one post that is annotated as *crisis, currentAcuteDistress, currentMildDistress, followupWorse, pastDistress, undeserved* and non-depressed, otherwise. Overall, we marked 65 users as depressed and 94 users as non-depressed. Table 1 details the size of our dataset.

3.1.1 Posting activity

Insomnia is one of the major symptoms of depression and literature on depression indicates that users showing depression signs tend to be more active during the evening and night, indicating insomnia as a promising feature for depression detection Lustberg and Reynolds (2000). Thus, we analyzed the posting behavior of the users as in De Choudhury et al. (2013). We divided the time into day and night and considered the 'night' window as '9p.m.–6a.m.' and the 'day' window as '6:01a.m.–8:59p.m.' (we used the local time of the user) and analyzed the average number of posts during these windows for depressed and non-depressed users.

De Choudhury et al. De Choudhury et al. (2013) showed that depressed users in Twitter tend to post more at night and have a higher insomnia index, defined as the normalized difference in the number of posts made during the night window and the day window. Conversely, in our dataset, we observe that, in general, all the users tend to post more during the day than at night and that depressed users tend to post more than non-depressed ones during

[1] https://au.reachout.com/.

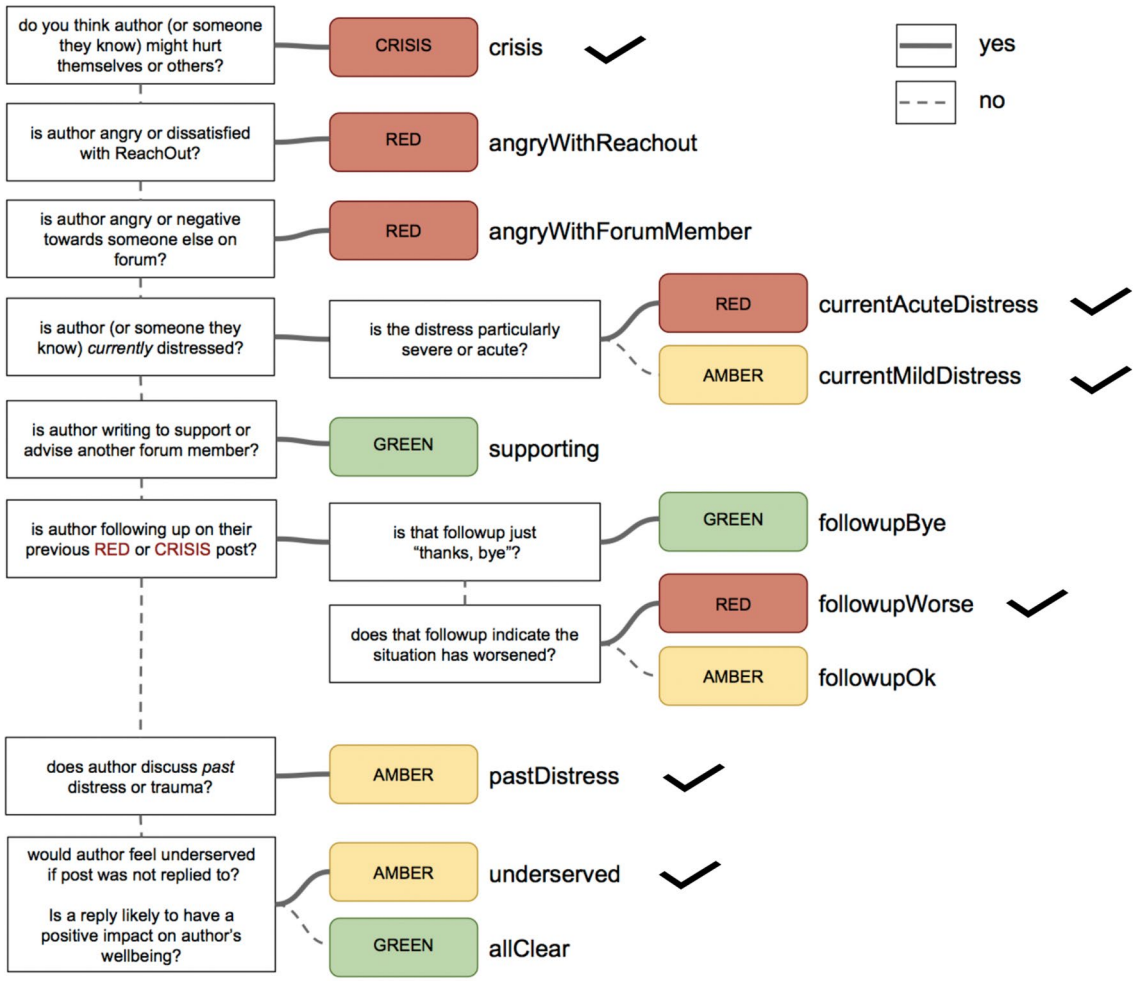

Fig. 1 The triage annotation decision tree. http://clpsych.org/shared-task-2017/

Table 1 Summary of the CLPsych 2017 dataset

Users	Posts	Depressed	Non-depressed	Unknown
1716	62,036	65	94	1557

both night and day time. On average, depressed users write 1.75 posts at night (vs. 0.76 posts for non-depressed users) and 4.79 posts during the day (vs. 2.79 posts for non-depressed users). One reason could be that the forum is specifically open to provide help, so depressed people feel free to post there at any time of the day, while they engage more with other online activities such as Twitter at night when their symptoms worsen Lustberg and Reynolds (2000). Thus, we did not find the insomnia index as an important feature in our dataset, and then, we did not use it for identifying depressed users.

4 Methodology

In this section, we describe our proposed methodology to identify depressed users in online forums. Given the scarcity of labeled users, we propose an unsupervised technique, as shown in Fig. 2. Given a user u, the first step is to compute a latent representation of u given the temporal sequence of posts they contributed to the forum. This user latent representation is learned in an unsupervised way using a long short-term memory (LSTM) autoencoder, as explained in Sect. 4.1. Next, we consider how forum users interact among them. Thus, we build a "who replies to whom" network and compute network-based features for each user as described in Sect. 4.2. These network features are concatenated to the user latent representation extracted from their post sequence and then used in input to an anomaly detection algorithm to identify depressed users. The different unsupervised algorithms we used and compared to perform the anomaly detection task are detailed in Sect. 4.3.

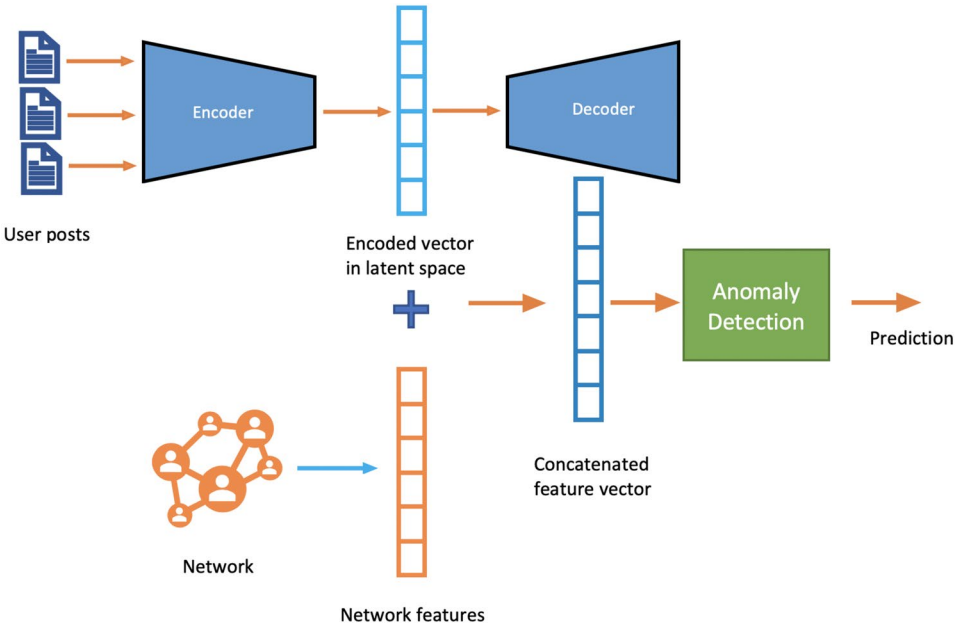

Fig. 2 Overview of the proposed unsupervised technique to identify depressed users in online forums

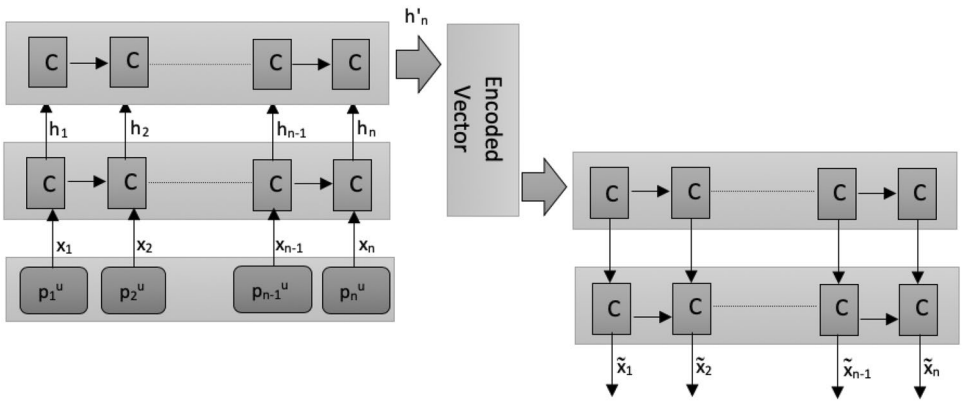

Fig. 3 Autoencoder architecture

4.1 Unsupervised learning of user representation from their posts

We propose to compute, for each forum user u, a set of latent features given the sequence of comments $\langle p_1^u, \ldots, p_n^u \rangle$ they posted in the forum. These latent features are learned by the stacked long short-term memory (LSTM) autoencoder shown in Fig. 3. An LSTM is a recurrent neural network Hochreiter and Schmidhuber (1997) where each cell C has the architecture shown in Fig. 4 (adapted from https://colah.github.io/posts/2015-08-Understanding-LSTMs/). Here, each LSTM cell outputs the next state h_t $(1 \leq t \leq n)$ by taking in input the previous state h_{t-1} and the next vector x_t. The operations done by the single LSTM cell C are described by the following equations:

$$a_t = \rho(W_a \cdot [h_{t-1}, x_t]) = 0 \quad (1)$$

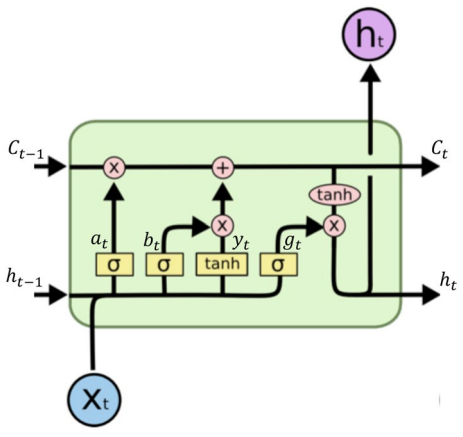

Fig. 4 Description of an LSTM cell C. Figure adapted from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

$$b_t = \rho(W_b \cdot [h_{t-1}, x_t]) \tag{2}$$

$$y_t = \tanh(W_y \cdot [h_{t-1}, x_t]) \tag{3}$$

$$g_t = \rho(W_g \cdot [h_{t-1}, x_t]) \tag{4}$$

$$c_t = c_{t-1} \cdot a_t + b_t \cdot y_t \tag{5}$$

$$h_t = \tanh(c_t) \cdot g_t \tag{6}$$

where the W_a, W_b, W_y and W_g are the weights representing the LSTM cell C and the entire LSTM neural network.

The encoder part in Fig. 3 takes in input the sequence of user posts, where each post p_t^u is represented using a vector x_t of linguistic features extracted from the post. As described in Sect. 4.1.1, we used the LIWC linguistic features. The subsequence $\langle x_1, \ldots, x_t \rangle$ is converted by the first LSTM into a single vector representation h_t of size k_1. The second LSTM takes in input the vectors $\langle h_1, \ldots, h_n \rangle$ from the first LSTM and further reduces their size to $k_2 < k_1$. The output h'_n of the last cell of the second LSTM gives the user u latent representation (or encoded vector) and represents the entire sequence of user u's posts.

The decoder part takes in input h'_n and reconstructs the input autoencoder sequence $\langle x_1, \ldots, x_t \rangle$ using the inverse of the architecture used by the encoder. We used the root mean square error (RMSE) loss function that measures the error between the input sequence $\langle x_1, \ldots, x_t \rangle$ and the reconstructed one $\langle \tilde{x}_1, \ldots, \tilde{x}_t \rangle$. We train our LSTM autoencoder by considering all the users in our dataset (depressed, non-depressed, and unknown).

4.1.1 Psycho-linguistic features for modeling user posts

The linguistic style captures how language is used by individuals and provides information about their behavioral characteristics subject to their social environment. Language can be quantified to unveil clues about the underlying psychology of the individual. Thus, to represent each post p_i^u in input to the LSTM autoencoder described in the previous section, we compute a vector x_i of Linguistic Inquiry and Word Count (LIWC) features from the post text. LIWC is a transparent text analysis tool that counts words in psychologically meaningful categories. It reads text files in batches and counts the percentage of words that belong to each category, which can be grouped as linguistic, punctuation, psychological, and summary features Pennebaker et al. (2015).

Linguistics features refer to features that represent the functionality of text, such as the average number of words per sentence and the rate of misspelling. This category of features also includes negations as well as parts of the speech (adjective, noun, verb, conjunction) frequencies. There are a total of 28 features under this category.

Punctuation features are used to dramatize or sensationalize a post that can be analyzed through types of punctuation used in the posts such as periods, commas, colons, semicolons, question marks, exclamation marks, dashes, quotation marks, apostrophes, parentheses, and other punctuation. There are a total of 11 features under this category.

Similarly, psychological features target emotional, social process, and cognitive processes. The affective processes (positive and negative emotions), social processes, cognitive processes, perceptual processes, biological processes, time orientations, relativity, personal concerns, and informal language (swear words, nonfluencies) can be used to scrutinize the emotional part of the posts. There are a total of 51 features under this category.

Summary features define the frequency of words that reflect the thoughts, perspective, and honesty of the writer. This category consists of features such as analytical thinking, clout, authenticity, emotional tone, words per Sentence (WPS), words with more than six letters, and dictionary words. There are a total of seven features under this category.

We used all the LIWC features for analyzing the cognitive, affective, and grammatical processes in the text, which helps in examining the difference between the writing style of posts among depressed and non-depressed users.

4.2 Network features

Since most of the work in depression or mental illness detection via social media has been done by analyzing user posts (especially on the ReachOut forum Cohan et al. (2016); Yates et al. (2017)), it would be interesting to analyze the users also from a networking point of view. Thus, to extract network-based features, we built a "who replies to whom" network as follows. We considered each user as a node in the network and added an edge from node u to node v if u wrote a post in reply to v's post. We denote this network as a directed graph $G = (V, E)$ where V is the set of nodes and E is the set of edges. In this paper, we use the following network features.

4.2.1 PageRank

The PageRank (PR) is a popularity measure for nodes in a network G and it is defined as

$$PR(u) = \frac{1-\beta}{|V|} + \beta \sum_{v \in M(u)} \frac{PR(v)}{|L(v)|} \tag{7}$$

where β is the damping factor usually set to 0.85 Brin and Page (1998), $M(u)$ is set of nodes that link to u, and $L(v)$ is the set of nodes pointed by v.

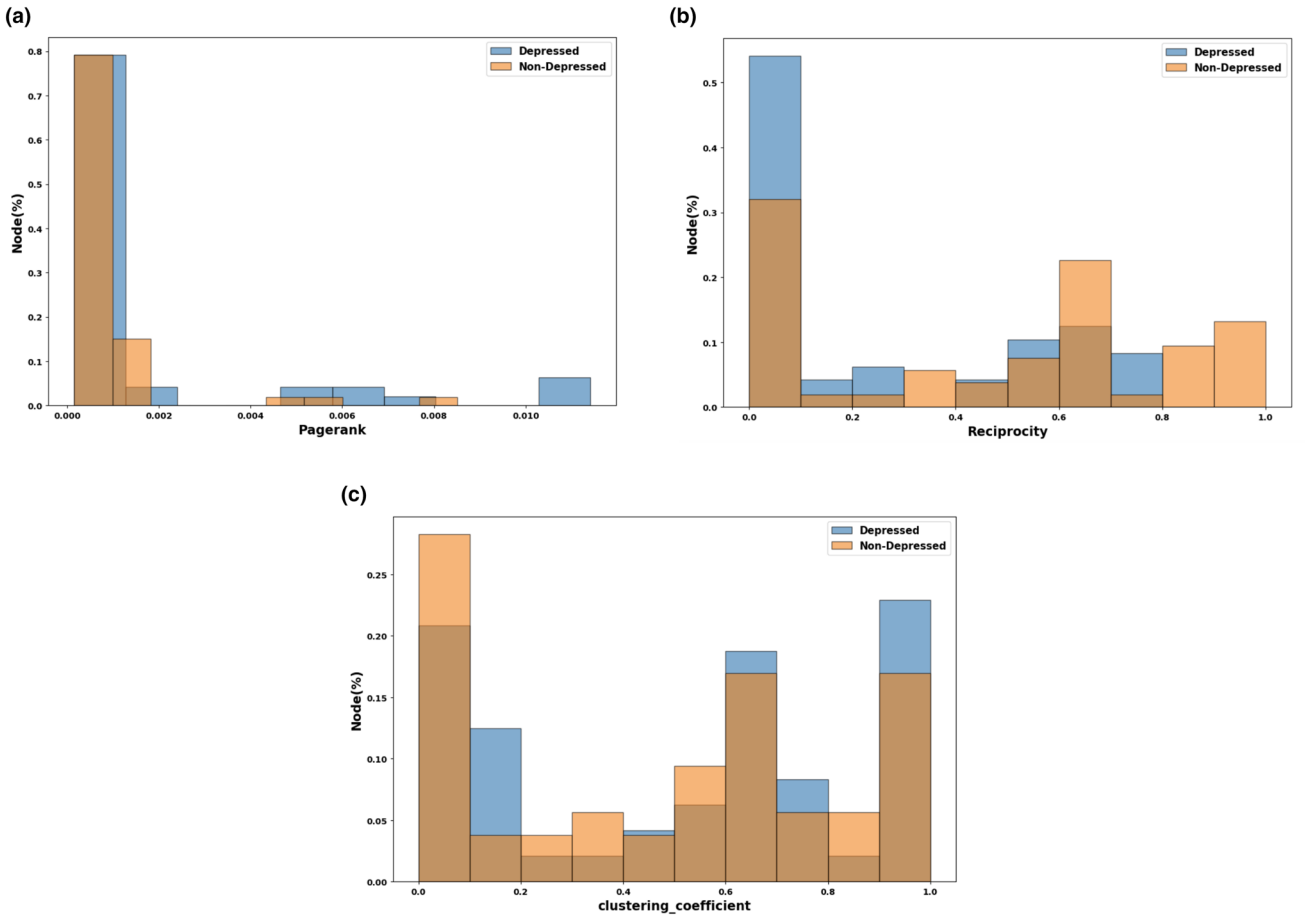

Fig. 5 PageRank (**a**), reciprocity (**b**), and local clustering coefficient (**c**) distribution of depressed (blue) and non-depressed (orange) users (Color figure online)

Figure 5a shows the distribution of PageRank for depressed and non-depressed users in our dataset. Here, we observe that depressed users tend to have a higher PageRank than non-depressed ones (0.0017 vs. 0.0008 on average).

4.2.2 Reciprocity

Reciprocity captures a basic way in which interaction on online sites takes place. When two users u and v interact, one expects that comments will be exchanged between them, i.e., users reply to each other. The reciprocity of a single node u is defined similarly. It is the ratio of the number of edges in both directions to the total number of edges involving the node u.

Figure 5b shows the distribution of reciprocity for depressed and non-depressed users in our dataset. As we can see, depressed users tend to have lower reciprocity than non-depressed ones (average reciprocity of 0.24 vs. 0.46), meaning that depressed users tend to reply (or their posts are replied) less than non-depressed users.

4.2.3 Clustering coefficient

It is observed that people who share connections in a social network tend to form clusters. The local clustering coefficient (LCC) measures the probability that the neighbors of a node are connected and it is equal to

$$\mathrm{LCC}(u) = \frac{2 \times |\{(v_1, v_2) \in E \mid v_1, v_2 \in \Gamma(u)\}|}{|\Gamma(u)| \times (|\Gamma(u)| - 1)} \tag{8}$$

where $\Gamma(u) = M(u) \cup L(u)$.

Figure 5c shows the distribution of the local clustering coefficient for depressed and non-depressed users in our dataset. We observe that depressed users have a higher local clustering coefficient value than non-depressed ones (average LCC of 0.52 vs. 0.47), meaning that depressed users' neighbors are more connected among them than the neighbors of non-depressed ones.

4.2.4 Node2Vec

Network embedding is a technique for mapping graph nodes in a geometric high dimensional space. Once the embedding is obtained for each node, its geometric representation can be used as features in input to machine learning algorithms. Node2Vec Grover and Leskovec (2016) is an embedding technique based on random walks. It computes the embedding in two steps. First, the context of a node (or neighborhood at a distance d) is approximated with biased random walks of length d that provide a trade-off between breadth-first and depth-first graph searches. Second, the values of the embedding features for the node are computed by maximizing the likelihood of generating the context by the given node.

4.3 Anomaly detection

Anomaly detection is the task of identifying the outlier or anomaly or the entity that does not comply with the normal behavior. The observation that significantly deviates from other observations is called an anomaly Chandola et al. (2009). The task of anomaly detection is not limited, for instance, to finding suspicious behavior in networks (intrusion detection) or finance applications but this technique can be leveraged for uncovering rare events such as symptoms of a new disease or unusual symptoms and rare diseases Hauskrecht et al. (2013). Thus, we use anomaly detection in our paper to identify depressed users by assuming that their behavior deviates from the one of normal users. As reported in Table 1, we have scarce labeled data, thus in this paper, we apply unsupervised anomaly detection techniques[2] to identify non-depressed (normal) and depressed (abnormal) users and use the available ground truth for evaluations purposes only. We apply and compare the following anomaly detection techniques.

Clustering-based anomaly detection techniques Clustering is an unsupervised machine learning technique that groups the observations into K clusters. Clustering can be used to performing anomaly detection under the assumption that "normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster." Chandola et al. (2009) In this paper we use K-means, Gaussian mixture model (GMM), and DBSCAN algorithms Han and Kamber (2012) to cluster the data. K-means and GMM have the problem of finding the optimal number of clusters and Gaussian components, respectively. To find these parameters, we used the elbow method for K-means and the Bayesian information criterion (BIC) for GMM.

Once the observations are clustered using K-means, we perform anomaly detection as follows. First, we compute the Euclidean distance between each data point and the centroid of the respective cluster. Second, we calculate the maximum cluster radius to identify the outliers. The maximum cluster radius determines the association of the data point to the particular cluster. For this, we use a percentile distance value as the threshold τ (i.e., for each cluster, the αth percentile of the distribution of all distances between data points and their respective centroid) to distinguish to which class the data point lies (if the distance is greater than τ, then we classify the user as depressed (anomaly), non-depressed otherwise).

Regarding GMM, once the model is fitted, it provides the weighted log of probability density function values for each data point. This probability density function can be used to understand which sample belongs to which class. For this, we used the percentile of the weighted log probability distribution values as the threshold τ (i.e., the αst percentile of the distribution of all weighted log probability) to determine in which class the data point lies: if the weighted log probability is less than τ then we classify the user as depressed (anomaly), non-depressed otherwise.

The DBSCAN algorithm is a density-based clustering algorithm that directly labels the data points as normal or anomaly, so no further steps are required to perform the anomaly detection task Ester et al. (1996).

One-class classification-based anomaly detection techniques These techniques assume that all the data instances have only one-class label. Under this category of algorithms, we used one-class SVM in our paper. The one-class SVM algorithm Schölkopf et al. (2001) learns the intrinsic properties of the normal cases (non-depressed users in our case) and uses these properties to understand which data point deviates from normal behavior (the known class). The data point that shows the abnormal behavior or that deviates from the normal are classified as anomalies (depressed users in our case) by the algorithm.

Ensemble-based anomaly detection techniques We considered isolation forest under this category. The isolation forest algorithm is based on the fact that the anomalous observations are very rare and have different properties than the normal ones, and using these properties, the anomalous observations can be isolated from the normal ones in a more effective way. The basic idea of isolation forest is to separate each data point by randomly creating a separation line between the data point and the others. Since the anomalous data points are different than normal ones and are few in number and scattered, they can be segregated in

[2] Unsupervised anomaly detection is used when the data are unlabelled, i.e., the class of an instance (normal or anomaly) is not known. This approach does not require the training or testing data, which makes it more flexible and widely applicable. The main idea of unsupervised anomaly detection is to provide a score for each instance by learning intrinsic properties such as distance or density. This score is called the anomaly score that determines whether the instance is normal or anomalous.

Table 2 Precision (Pr), recall (Re), and F1-measure (F1) of anomaly detection with social network features, psycho-linguistic features and combination

	K-means			GMM			DBSCAN			Isolation Forest			OC-SVM		
Features	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
PageRank + reciprocity + clustering coeff.	0.41	0.54	0.43	0.56	0.56	0.56	0.49	0.12	0.16	0.59	0.60	0.52	0.61	0.59	0.59
Node2Vec	0.58	0.60	0.49	0.56	0.56	0.56	0.55	0.31	0.32	0.56	0.59	0.51	0.60	0.58	0.59
LIWC	0.67	0.62	0.51	0.55	0.55	0.55	0.39	0.27	0.23	0.62	0.62	0.54	0.58	0.56	0.56
autoencoder	0.63	0.61	0.51	0.61	0.61	0.61	**0.78**	0.41	0.28	0.62	0.62	0.54	0.59	0.57	0.58
Proposed technique: Autoencoder + PageRank + reciprocity + clustering coeff.	0.63	0.61	0.51	0.64	**0.64**	**0.64**	0.77	0.42	0.27	0.62	0.62	0.54	0.60	0.58	0.59
Autoencoder + Node2Vec	0.61	0.60	0.50	0.57	0.57	0.57	0.77	0.42	0.27	0.59	0.60	0.52	0.58	0.56	0.56
Autoencoder + all networks	0.58	0.60	0.49	0.58	0.58	0.58	0.77	0.42	0.27	0.62	0.62	0.54	0.60	0.58	0.58

The best scores are given in bold

a few numbers of splitting. While, normal data points that are closer take a significant number of splittings Liu et al. (2008).

5 Experiments

This section reports on our experimental results of using linguistic and network-based features to identify depressed users in an unsupervised fashion.

5.1 Experimental setting

Since our methodology is based on unsupervised anomaly detection, we compute the user representation from their posts and the network features by considering all the users in the dataset (depressed, non-depressed, and unknown). Next, anomaly detection is performed using all the techniques presented in Sect. 4.3, namely K-means, Gaussian mixture model, DBSCAN, isolation forest, and one-class SVM. Once the users are labeled as normal (non-depressed) or abnormal (depressed) by any of the anomaly detection methods, we evaluate the prediction using the ground truth available in our dataset: 65 depressed users and 94 non-depressed users. As evaluation measures, we use precision (Pr), recall (Re), and F1-measure (F1), similarly to related work.

Parameter setting The parameters of the algorithms used in our experimental evaluation have been set as follows. For Node2Vec, we set the number of features to 32, the random walk length to 20, and the number of walks to 100. For DBSCAN, we set *eps* to 0.1 and *min_samples* to 2. For one-class SVM, we used RBF kernel with γ (kernel coefficient) equal to the inverse of the number of features and the ν parameter as the ratio of anomalous observations that we assume is present in the dataset. For isolation forest, we used 100 estimators and contamination as a proportion of outlier that we assume is present in the dataset. Finally, the autoencoder proposed in Sect. 4.1 learns a user representation h'_n of size 32.

5.2 Baselines for comparison

We compare our proposed method from Sect. 4 with the following network and linguistic-based baselines:

- *PageRank + reciprocity + local clustering coefficient* we perform the unsupervised anomaly detection task by considering these network features only.
- *Node2Vec* we perform the unsupervised anomaly detection task by considering the Node2Vec features only.
- *LIWC* for each user, we create a unique document by concatenating all their posts. Then, we compute the LIWC features of these documents (this is similar to the setting we had in our previous work Shrestha and Spezzano (2019)) and perform the unsupervised anomaly detection task with these features as input.

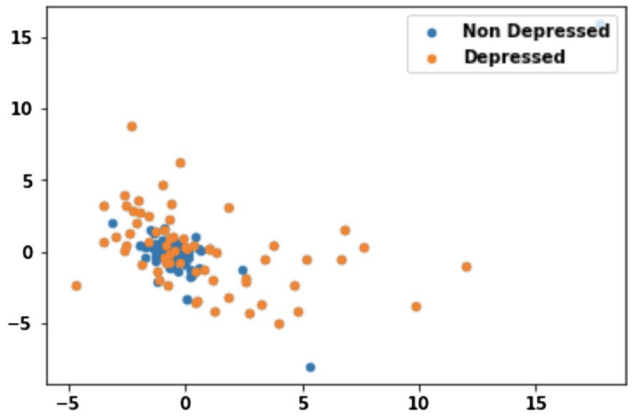

Fig. 6 Plot of the user embeddings computed with the autoencoder and reduced in a two-dimensional space via PCA

5.3 Results

Classification results are reported in Table 2. When considering a singular modality (user post content or network), we have our autoencoder-based approach with the Gaussian mixture model (based on user posts only) that provides the highest F1-score in comparison with the baselines PageRank + reciprocity + local clustering coefficient, Node2Vec and LIWC (cf. the first four rows of Table 2). The 5% minimum improvement (w.r.t. network-based baselines such as Node2vec and PageRank + reciprocity + local clustering coefficient) is due to the fact that the autoencoder is able to (1) consider the temporal characteristics of the user's mood through their posts (where the mood is computed by the LIWC features on each post), and (2) it is also able to recreate an embedding distribution that pairs well with standard anomaly detection techniques. Figure 6 plots the user embeddings computed with the autoencoder and reduced in a two-dimensional space via PCA. We clearly see that the majority of the orange points representing depressed users are far from the central cluster containing all the non-depressed users.

Our proposed methodology combines the user representation obtained by the autoencoder from the user post sequence with user network features. From the last three rows of Table 2 we observe that PageRank + reciprocity + local clustering coefficient are better than Node2Vec features when combined with the autoencoder features. In fact, this combination provides a further 3% improvement in the F1-score (0.64) w.r.t the autoencoder features only. Please note that in the case of anomaly detection computed with DBSCAN, even if the precision for autoencoder is 0.78, the recall is very low (0.41) by negatively impacting the F1-score that is 0.28. For this motivation, we do not consider the DBSCAN as a good option.

The deep learning-based approach we propose in our paper is more complex (in terms of execution time) than traditional methods from Sect. 5.2 we compare with. However, the additional complexity that exploits the temporal relationship among the user comments allows us to achieve better classification results than traditional methods. Moreover, once the model is trained, it can be used to infer the embedding for a new user without re-training (it is sufficient to pass in input the new user's sequence). Hence, our proposed approach can be applied to classify new users with the same complexity as traditional methods.

Qualitative analysis of unknown users To further strengthen our experimental results, we considered the unknown users in our dataset, sorted them by the score provided by our proposed technique,[3] and manually inspected the posts of the top 10 and bottom 10 unknown users. In the case of top 10 users, which are candidates for non-depressed users, we observed normal and positive comments similar to common comments regarding travel or movies users post on social media: "Is that Castle House for real? I'd seen it before but assumed it was total CGI magic. I've been wasting.. err I mean spending a lot of time on Airbnb lately and have found so many places I want to travel largely because of how cool the accom is. Here is some igloo acomm in Finland where you can see the Northern Lights through the ceiling!"; "My top five at the moment are: Game of Thrones...True Blood The Newsroom...Breaking Bad Also pretty excited about new season of Sons of Anarchy... Any fans of the above here?" Hence, we conclude these users do not show signs of depression.

Regarding the bottom 10 unknown users, which are candidates for depressed users, we found in the majority of them, at least one comment expressing discomfort. For instance, some comments were describing episodes where these people were crying without any reason: "when I'm alone again I get that 'empty' feeling and some days I feel kinda weak like I want to cry, and it's weird, this hasn't happened to me before." Other comments were about the fact that they noticed to be more aggressive than usual: "Hi sorry dumping All my problems on you again, but i have some, well, anger problems. Things piss me off easily and because i cant do vilance at school i take it out on my parents by yelling and fighting with them or crying for no reason."

This analysis confirms a good performance of our proposed technique also in the case of unlabeled users. In fact, crying without any reason and being more aggressive than usual are common symptoms of depression.

[3] The score is given by the weighed log probability obtained with GMM when this anomaly detection algorithm is applied to our autoencoder features plus PageRank, reciprocity, and local clustering coefficient network features.

6 Conclusion

We addressed the problem of identifying depressed users in online forums in an unsupervised fashion. We analyzed user behavior in the ReachOut.com online forum using psycho-linguistic features extracted from the sequence of user posts in combination with network-based features modeling how users connect in the forum. Our results showed the potential of these features in characterizing depressed users in online forums, especially user embedding extracted from user posts and network-based features such as reciprocity and local clustering coefficient. By combining both network and psycho-linguistic features, our proposed unsupervised approach achieved an F1-measure of 0.64 in detecting depressed users and performed better than baselines.

Future work will be devoted to (1) extending our results to the problem of early detection of depressed users in online forums and (2) exploiting depressed user detection techniques to enhance risky post detection by including author network features.

References

Brew C (2016) Classifying reachout posts with a radial basis function SVM. In: Proceedings of the 3rd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, CLPsych@NAACL-HLT, San Diego, California, pp 138–142

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw 30(1–7):107–117

Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv (CSUR) 41(3):15

Clpsych dataset. http://clpsych.org/shared-task-2017/

Cohan A, Young S, Goharian N (2016) Triaging mental health forum posts. In: Proceedings of the third workshop on computational lingusitics and clinical psychology, pp 143–147

De Choudhury M, Gamon M, Counts S, Horvitz E (2013) Predicting depression via social media. ICWSM 13:1–10

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

Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, Asch DA, Schwartz HA (2018) Facebook language predicts depression in medical records. Proc Natl Acad Sci 115(44):11203–11208

Ester M, Kriegel HP, Sander J, Xu X et al (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. KDD 96:226–231

Grover A, Leskovec J (2016) Node2vec: scalable feature learning for networks. In: SIGKDD, pp 855–864

Han J, Kamber M (2012) Data mining: concepts and techniques

Hauskrecht M, Batal I, Valko M, Visweswaran S, Cooper GF, Clermont G (2013) Outlier detection for patient monitoring and alerting. J Biomed Inf 46(1):47–55

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Kim SM, Wang Y, Wan S, Paris C (2016) Data61-csiro systems at the CLPSYCH 2016 shared task. In: Proceedings of the 3rd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, CLPsych@NAACL-HLT 2016, San Diego, California, pp 128–132

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: International conference on machine learning, pp 1188–1196

Liu FT, Ting KM, Zhou ZH (2008) Isolation forest. In: 2008 Eighth IEEE international conference on data mining, IEEE, pp 413–422

Lstm description. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Lustberg L, Reynolds CF III (2000) Depression and insomnia: questions of cause and effect. Sleep Med Rev 4(3):253–262

MacAvaney S, Desmet B, Cohan A, Soldaini L, Yates A, Zirikly A, Goharian N (2018) RSDD-time: Temporal annotation of self-reported mental health diagnoses. In: Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic, CLPsych@NAACL-HTL, New Orleans, pp 168–173

Malmasi S, Zampieri M, Dras M (2016) Predicting post severity in mental health forums. In: Proceedings of the 3rd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, CLPsych@NAACL-HLT 2016, San Diego, pp 133–137

Metcalf A, Blake V (2013) Reachout.com annual user survey results

Millen D (2015) Reachout annual report 2013/2014

Park M, McDonald DW, Cha M (2013) Perception differences between the depressed and non-depressed users in twitter. ICWSM 9:217–226

Park M, Cha C, Cha M (2012) Depressive moods of users portrayed in twitter. In: Proceedings of the ACM SIGKDD workshop on healthcare informatics (HI-KDD), vol 2012, ACM New York, pp 1–8

Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The development and psychometric properties of LIWC2015. In: Technical report

Resnik P, Armstrong W, Claudino L, Nguyen T, Nguyen VA, Boyd-Graber J (2015) Beyond LDA: exploring supervised topic modeling for depression-related language in twitter. In: Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality, pp 99–107

Rude S, Gortner EM, Pennebaker J (2004) Language use of depressed and depression-vulnerable college students. Cognit Emot 18(8):1121–1133

Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC (2001) Estimating the support of a high-dimensional distribution. Neural Comput 13(7):1443–1471

Shrestha A, Spezzano F (2019) Detecting depressed users in online forums. In: International symposium on network enabled health informatics, biomedicine and bioinformatics (HI-BI-BI 2019), in conjuction with ASONAM'19

Staiano J, Guerini M (2014) Depechemood: a lexicon for emotion analysis from crowd-annotated news. arXiv preprint arXiv:1405.1605

Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for computational linguistics, pp 347–354

Xu R, Zhang Q (2016) Understanding online health groups for depression: social network and linguistic perspectives. J Med Internet Res 18(3):e63

Yates A, Cohan A, Goharian N (2017) Depression and self-harm risk assessment in online forums. arXiv preprint arXiv:1709.01848

Zimmermann J, Brockmeyer T, Hunn M, Schauenburg H, Wolf M (2017) First-person pronoun use in spoken language as a predictor of future depressive symptoms: preliminary evidence from a clinical sample of depressed patients. Clin Psychol Psychother 24(2):384–391