A Deep Learning Ensemble to Classify Anxiety, Depression, and their Comorbidity from Texts of Social Networks

Vanessa Souza, Jeferson Nobre, Karin Becker

Universidade Federal do Rio Grande do Sul, Brazil
{vbsouza, jcnobre, karin.becker}@inf.ufrgs.br

Abstract. The use of social networks to expose personal difficulties has enabled works on the automatic identification of specific mental conditions, particularly depression. Depression is the most incapacitating disease worldwide, and it has an alarming comorbidity rate with anxiety. In this paper, we explore deep learning techniques to develop a stacking ensemble to automatically identify depression, anxiety, and comorbidity, using data extracted from Reddit. The stacking is composed of specialized single-label binary classifiers that distinguish between specific disorders and control users. A meta-learner explores these base classifiers as a context for reaching a multi-label, multi-class decision. We developed extensive experiments using alternative architectures (LSTM, CNN, and their combination), word embeddings, and ensemble topologies. All base classifiers and ensembles outperformed the baselines. The CNN-based binary classifiers achieved the best performance, with f-measures of 0.79 for depression, 0.78 for anxiety, and 0.78 for comorbidity. The ensemble topology with best performance (Hamming Loss of 0.29 and Exact Match Ratio of 0.47) combines base classifiers according to three architectures, and do not include comorbidity classifiers. Using SHAP, we confirmed the influential features are related to symptoms of these disorders.

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: Deep learning, Social networks, Mental health

1. INTRODUCTION

The disabling impact of depression on society is a major concern of public health entities. The World Health Organization (WHO) estimates that depression affects near 322 million worldwide, at an approximate cost of $2.5 trillion dollars[1]. Depression is characterized by the presence of a sad, empty or irritable mood, accompanied by somatic and cognitive changes that significantly affect the individual's ability to function, impairing their performance in daily tasks and social life [American Psychiatric Association 2013]. Anxiety is another prevalent disorder, characterized by feelings of tension, excessive fear, recurring intrusive thoughts, and physiological changes [American Psychiatric Association 2013]. Anxious individuals experience physical symptoms such as restlessness, tension, sleep disorders, difficulty in concentrating, irritability, and muscle tension.

WHO also reports an alarming comorbidity rate of anxiety and depression. Around 85% of patients with depression experience significant symptoms of anxiety, while depression occurs in more than 90% of patients who suffer from anxiety disorders [Tiller 2013]. Such comorbidity accentuates the clinical picture of depressed individuals, leading to a higher risk of suicide, worse social functioning, and resistance to treatment [Hirschfeld 2001]. The impact imposed by depression on society requires prevention and intervention strategies, particularly focused on screening and early diagnosis [Hamilton

[1] https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf

This research was partially funded by FAPERGS - Brazil (19/2551-0001862-2)

1967; Radloff 1977]. The task of diagnosing an individual suffering from one or more mental disorders involves different skills, ranging from the perception and interpretation of the patient's reports to the subtle distinction of symptoms between disorders that present common behaviors [Hirschfeld 2001].

Individuals have used social media platforms (e.g., Twitter, Reddit) to expose their difficulties, including the ones related to mental disorders. Motivations include obstacles in the access to treatment, loneliness, shame, or to avoid prejudice. This use of social networks allows the development of computational solutions to support the study of mental disorders, as there is a huge volume of high-value data, containing thoughts, sentiments, and behavioral expressions of these individuals. A systematic review [Wongkoblap et al. 2017] reveals that related work primarily focus on the automatic identification of specific disorders using supervised learning techniques over textual and social interaction data available on the internet.

Most works address the automatic classification of depression. A systematic review [Giuntini et al. 2020] summarizes the techniques deployed for the recognition of depressive mood disorders in social networks. In general, these works deploy supervised learning algorithms and extensive feature engineering to derive textual, social, and sentiment features (e.g., [Sharma and De Choudhury 2018; De Choudhury et al. 2017; Park et al. 2015; Tsugawa et al. 2015; De Choudhury et al. 2014]). Few works address anxiety [Dutta et al. 2018; Ireland and Iserman 2018; Gruda and Hasan 2019], or its comorbidity with other disorders, including depression [Cohan et al. 2018; Bagroy et al. 2017]. More recently, deep learning techniques have been explored for the classification of depression [Yates et al. 2017; Mann et al. 2020], chronic stress [Lin et al. 2014], and anxiety [Shen and Rudzicz 2017]. Common deep learning architectures are convolutional networks (CNN) and recurrent network variations, such as LSTM (Long-Short Term Memory) [Kowsari et al. 2019]. Deep learning has the benefit of including the extraction of data representations from input data as part of the learning process [Murphy 2012]. Our research leverages deep learning techniques for the automatic classification of depression, anxiety, and their comorbidity, with the aim of contributing with insights and patterns derived from textual social interaction.

In this article, we propose a stacking ensemble for the automatic classification of depression, anxiety, and their comorbidity, using a self-diagnosed dataset extracted from Reddit [Cohan et al. 2018]. The use of a stacking ensemble [Zhang and Ma 2012] aims to overcome the difficulties of dealing with a multi-class, multi-label classification problem involved in the scenario of comorbidity, where the distinguishing patterns may be harder to identify. At the lowest level, the ensemble is composed of single-label binary classifiers (base classifiers), which predict class probabilities related to diagnosed/control users of a specific target condition (anxiety, depression, or comorbidity). To develop the base classifiers, we experimented distinct deep learning architectures (CNN, LSTM, and their combination) and pre-trained word embeddings (general-purpose and domain-specific). At a higher level, these individual predictions are consolidated using a dense neural network, which handles the multi-label, multi-class problem of assigning control or diagnosed labels. We assessed the different components of our solution: a) the effect of deep learning architectures in the ensemble, by experimenting with ensemble topologies composed of homogeneous and heterogeneous architectures; b) the function of the base classifiers, i.e. specialized in identifying isolated disorders (depression or anxiety), or if patterns related to comorbidity should be considered, c) the effect of embeddings in the base classifiers and the ensemble, d) the relationship of the influential features in the classification of diagnosed users and disorders', as described in the DSM-5 psychological manual [American Psychiatric Association 2013]. The paper details the architectural choices, and the quantitative and qualitative assessments performed.

Our assessments revealed very encouraging results. All single-label binary base classifiers outperform existing solutions for depression, anxiety, and comorbidity [Yates et al. 2017; Cohan et al. 2018]. The best performance was achieved by CNN base classifiers, achieving F-measures of 0.79 for depression, 0.78 for anxiety, and 0.78 for their comorbidity. All ensembles also outperformed the baselines,

and best performance (Hamming Loss of 0.29 and Exact Match Ratio of 0.47) was achieved by a homogeneous topology composed to CNN base classifiers, and which did not include classifiers for comorbidity. To identify the influential features in the classification of specific disorders, we adopted SHAP (Shapley Additive Explanation) [Lundberg and Lee 2017], a method that explains the prediction of a given instance according to coalitional game theory, and which enables the global interpretation of influential features by the aggregations of Shapley values. The results were encouraging, as most influential features correspond to symptoms described in DSM-5 manual.

This article is an extension of our previously presented work [Souza et al. 2020]. We have significantly evolved it by: a) considering different deep learning architectures (LSTM, CNN and hybrid); b) formally assessing the effect on the base classifiers' performance of general-purpose and domain-specific pre-trained word embedding; and c) updating the theoretical background and related work. With regard to related work, our main contributions are:

a) a stacking ensemble approach that leverages deep learning architectures to solve the multi-label classification problem involving control and diagnosed users (anxiety, depression or both disorders).
b) experiments that assess the quantitative performance of deep learning architectures and classification functions in the composition of the ensemble. The experiments demonstrate that the best performance was achieved using an homogeneous topology of CNN base classifiers, followed by the heterogeneous topology. These results significantly outperform our original work [Souza et al. 2020], restricted to the LSTM architecture. All ensembles performed better when their topology did not include classifiers targeted at comorbidity.
c) experiments assessing the influence of general-purpose and domain specific domain word embeddings in the performance of the base classifiers;
d) a qualitative assessment of the base classifiers that relates SHAP influential classification features to symptoms of each target condition according to the DSM-5 psychological manual.

The remainder of the paper is organized as follows. Section 2 summarizes the theoretical background and Section 3 discusses related work. Section 4 describes the dataset used. Section 5 details the architectural elements of the proposed ensemble. Experiments assessing the proposed solution are presented in Section 6. Section 7 draws conclusions and discusses future work.

2. THEORETICAL BACKGROUND

The solution proposed in this article for the automatic identification of depression, anxiety, and its comorbidity leverages DL, model explainability and ensemble stacking. In the remaining of this section we briefly summarize the relevant concepts underlying these techniques.

2.1 Deep Learning

Deep learning has emerged as a powerful technique that allows computational models to learn representations of large sets of data using computing power. Deep learning architectures are different from traditional neural networks, as their structure is comprised of more layers and more units within a layer. In a nutshell, deep learning uses a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. The lower layers, closer to the data input, learn simpler features, while higher layers learn more complex features derived from lower layer ones [Zhang et al. 2018; Harb et al. 2019]. The architecture forms a hierarchical and powerful feature representation [Murphy 2012]. The result is a set of features that hierarchically grows in complexity. Defining the architecture of a deep neural work is a complex task, and the design decisions involve cell disposition in different layers (input, hidden layers, output), choice of activation and loss functions (e.g., sigmoid and binary crossentropy), regularization approaches to avoid overfitting (e.g., dropout), among others.

Deep learning has become increasingly popular for text classification [Kowsari et al. 2019; Minaee et al. 2020; Zhang et al. 2018]. Typically, deep learning models for textual data rely on *word embeddings* as input features. Word embeddings are low-dimensional dense vectors representations learned from data, such that words that frequently appear in similar contexts are close to each other. Word embeddings can be learned from the input *corpora* by an embedding layer in the neural network architecture, or produced independently using an unsupervised machine learning algorithm such as Word2vec [Mikolov et al. 2013] or GloVe [Pennington et al. 2014]. Different general-purpose pre-trained embeddings trained using a huge corpus to reflect general vocabulary usage are available, such as GloVe 6B[2], GloVe Twitter[2], and Google News[3]. When including an embeddings layer in a deep neural network, it is necessary to define if the learning approach is static or non-static. In the static learning approach, the layer referring to the word embeddings is frozen, preventing the update of their weights as a result of the learning process. In non-static models, word vectors are initialized according to the pre-trained embeddings, and these weights are updated during training through back-propagation during training.

Surveys [Minaee et al. 2020; Zhang et al. 2018] summarize the available deep learning architectures for text classification. Variations of LSTM (Long-Short Term Memory) and Convolutional Neural Networks (CNN) are among the most deployed architectures to learn the intrinsic semantic and syntactic relationships between words, and are adopted in this work.

As a recurrent neural network, LSTM architectures [Hochreiter and Schmidhuber 1997] are suitable for handling sequence elements since they maintain a state relative to what has been processed so far. LSTM architectures have been successfully applied to natural language processing applications, such as sentiment analysis [Amora et al. 2018; Becker et al. 2019], speech or hand writing recognition [Greff et al. 2017]. For text classification, LSTM are successful under the premise that the order of words in documents are representative. Typically, the architecture is organized in terms of layers of LSTM units, where each unit is a cell composed of an input gate, an output gate, and a forget gate. The cell remembers values over arbitrary time intervals, and the three gates regulate the flow of information into and out of the cell. Design decisions involve the number of layers and units within layers, as well as many hyper-parameters for regularization (e.g., dropout, recurrent dropout), maintenance of states within the batch (e.g., return sequence), or between batches (e.g., stateful).

Convolutional neural networks (CNNs) [Chollet 2017] utilize layers with convolving filters. In a convolution, a filter slides (convolves) over the input space to find local patterns, and generate feature maps. Typically, a convolutional layer applies different filters. Following a convolutional layer, a pooling layer is used to reduce the spatial size of the extracted representation progressively, and thus to reduce the number of features. Convolutional layers may be organized in a hierarchy, and finally, serve as input to a dense layer. CNNs were initially proposed for computer image problems, but have achieved particularly good results in traditional Natural Language Processing (NLP) tasks. CNNs have a particular spatially local correlation by enforcing a local connectivity pattern between neurons of adjacent layers. Such a characteristic is useful for classification in NLP, in which we expect to find meaningful local clues regarding class membership, such as the combination of specific terms/phrases, regardless where they appear in a document. Textual data are typically processed by one-dimensional CNN, which considers the input a sequence of words [Kim 2014]. The design choices of a CNN network typically involve the number of convolutional layers, the size of the sliding window (i.e. kernel size), the total of filters, pooling decisions, among others.

Model explainability is fundamental for the broader adoption of deep learning. SHAP (SHapley Additive exPlanation) [Lundberg and Lee 2017] explains the prediction of a given instance by computing the individual contribution of each feature according to coalitional game theory. Unlike other methods focused on individual explanations (e.g., LIME), SHAP provides global interpretation methods based

[2]https://nlp.stanford.edu/projects/glove/
[3]https://code.google.com/archive/p/word2vec/

on aggregations of Shapley values. In this work, we adopt SHAP values to identify the influential features in recognizing a specific target condition, such as depression, anxiety or comorbidity.

2.2 Ensemble Classifiers

An ensemble consists in combining different base (or weak) models with the purpose of a collective decision. It is a technique used to improve the performance of different machine learning tasks. The premise is that each individual base model contributes with a different hypothesis space, and their aggregation in a final model reduces the computational cost of training a single model for a complex task [Gama et al. 2011]. When designing an ensemble, it is necessary to ensure that the base models meet the following criteria [Zhang and Ma 2012]: (1) diversity, as they must be able to carry out analyzes independently one of another; and (2) accuracy, the error rate of the base models must be smaller than 50%.

Ensemble stacking [Zhang and Ma 2012] is a popular technique to organize ensembles. The *level 0* is composed of different base classifiers, all of them trained on the same data, but by different algorithms (and/or parameterizations), as a means to generate variability. Each additional level receives the predictions of the previous layer, and propagates their predictions to the next layer. The final level is the Meta-Learner, responsible for consolidating the individual results for a final prediction. The meta-learning function needs to manage and consolidate the individual probabilities according to some method (e.g., voting, meta-classifier). In this work, we adopt two-levels stacking ensembles to consolidate the prediction of base classifiers as a multi-label, multi-class problem.

3. RELATED WORK

Several works have contributed to the characterization of mental disorders from texts and interactions available on social media. A systematic review [Wongkoblap et al. 2017] reveals that most works focus on developing a predictive model for a single, specific disorder, where depression is the most studied one. Another systematic review [Giuntini et al. 2020] details the properties of studies specifically focused on depressive mood identification. Most of these works applied shallow learning algorithms such as SVM or regression on data resulting from extensive features engineering. These works vary on the information extracted from social media, their representation, as well as on the techniques to handle high dimensionality. All these works extract information from the text of posts, and additionally consider other features such as the frequency and morphological structure of words, sentiment, and social features [Lin et al. 2014; De Choudhury et al. 2014; Park et al. 2015; Tsugawa et al. 2015; De Choudhury et al. 2016; De Choudhury et al. 2017; Sharma and De Choudhury 2018; Dutta et al. 2018; Tadesse et al. 2019].

A pioneer work in the use of deep learning to identify depressed users is presented in [Yates et al. 2017], using a large dataset extracted from Reddit. It proposes a CNN architecture that summarizes users' posting activities as vectors, followed by dense layers that perform user classification. To the best of our knowledge, this work represents the state-of-the-art in depression classification (F1 = 0.65). [Lin et al. 2014] used a CNN to reduce the dimensionality of manually engineered features to identify stress. A more recent study [Mann et al. 2020] presents a multi-model approach for detecting depression in Instragam users, combining ELMo for text processing and ResNet for picture processing. An approach for classifying posts in Reddit expressing anxiety is described in [Shen and Rudzicz 2017], but the problem of consolidating posting behavior at user level was not addressed. The classification of Reddit posts is also addressed in [Gkotsis et al. 2017; Ive et al. 2018] using CNN and bi-directional GRU, for the single-label problem of selecting one out of eleven mental conditions.

Very few works address the comorbidity of mental disorders. Using Reddit data, classifiers for nine (9) mental conditions were developed in [Cohan et al. 2018], including anxiety and depression. The authors experimented with both shallow and deep learning techniques to develop classifiers for

each individual condition (single-label, binary classification), as well as their comorbidity (multi-label multi-class classification). The results were unsatisfactory, where the highest F-measure were achieved using FastText (0.54 for anxiety and depression binary classifiers, 0.27 for the multi-label, multi-class classifier). A multi-label model is described in [Benton et al. 2017] to detect Twitter users with suicidal risks, which considers seven other conditions (in isolation or comorbidity), including depression and anxiety.

A critical factor for mental disorder classification is the availability of large, non-biased training datasets. A technique for automatically labeling Reddit social network users was proposed in [Yates et al. 2017] for depression, and later extended to nine other mental conditions [Cohan et al. 2018]. The authors propose the use of high precision patterns to identify users who claimed to have been diagnosed with a mental health condition (*diagnosed users*) and use exclusion criteria to match them with *control users*. The method is designed to prevent biases between the control and diagnosed groups, such that the classification task is not artificially easy due to the presence of obvious expressions.

The present work contributes to the field by addressing the automatic classification of depression, anxiety, and their comorbidity. This is done to gain insights about the common and differentiating patterns that can be derived from textual social interaction.

4. DATA

This work uses the Self-reported Mental Health Diagnoses (SMHD) dataset [Cohan et al. 2018], which contains public Reddit posts from users with one or multiple mental health conditions along with matched control users[4]. It encompasses 9 mental disorders for the posting period from January/2006 to December/2017. We used only data related to depression and anxiety, together with the respective control groups. To investigate the best way to recognize each condition individually, and their comorbidity, we derived four datasets from SMHD, listed in Table I. The first three datasets contain users diagnosed with Anxiety only (A), Depression only (D), and comorbidity (AD), together with the respective control users. They were all prepared as single-label datasets. The last dataset (A-D-AD) is multi-label, and contains all possible combinations of these disorders, together with control users. For reproducibility purposes, the SMHD dataset organizes users into three subsets (training, validation, and test), equality and randomly distributed, and we maintained the original division of instances within these subsets.

Dataset	Type dataset	Classes	Total Users	Total Posts
SMHD A	single-label	Anxiety	1,560	240,330
		Control	1,560	458,364
SMHD D	single-label	Depression	3,230	474,271
		Control	3,240	932,259
SMHD AD	single-label	Comorbidity	1,320	191,056
		Control	1,320	390,892
SMHD A-D-AD	multi-label	Anxiety	1,320	202,370
		Depression	1,320	195,711
		Comorbidity	1,320	191,056
		Control	2,640	764,174

Table I. Datasets derived from SMHD for the experiments.

5. AN ENSEMBLE ARCHITECTURE FOR THE IDENTIFICATION OF DEPRESSION, ANXIETY AND COMORBIDITY

Studies reveal that anxiety and depression are mental conditions widely observed in the population word-wide, and their commorbidity is also frequent [Tiller 2013; Hirschfeld 2001]. As discussed in the

[4]The SMHD dataset was made available to this work by Georgetown University under a data usage agreement.

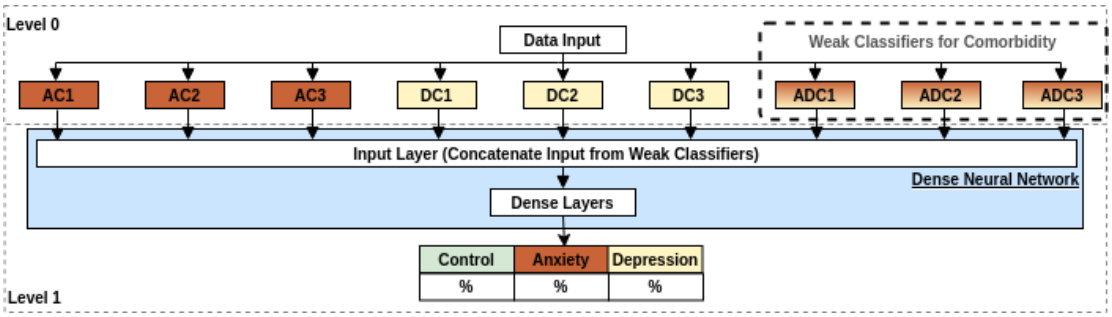


Fig. 1. Architecture of the Stacking Ensemble Classifier

previous section, while there are many solutions proposed for the automatic recognition of depression, fewer exist for anxiety and its comorbidity with depression. In this work, the identification of anxiety and depression conditions are considered as a multi-label classification task, such that a user can be identified with one or more disorders, thus encompassing a comorbidity condition. The multi-label classification was adopted due to the advantage of making the proposed solution extensible to other mental conditions and possible comorbidities. The proposed model is trained to identify the traits of mental disorders based on the message history of the Reddit social network user.

We adopted an ensemble approach in order to address the difficulties of dealing with the problem of multi-label classification involved in a comorbidity scenario. Thus, combining specialized classifiers for each disorder can be more effective in identifying the condition of comorbidity, compared to the development of a single model. The ensemble architecture is outlined in Figure 1. At the lowest level (*Level 0*), single-label binary base classifiers predict class probabilities related to diagnosed/control users of a specific target condition: Anxiety (AC_i), Depression (DC_i) and Comorbidity (ADC_i). In *Level 1*, a meta-learner is a dense-neural network that consolidates all these probabilities into a final multi-class, multi-label prediction.

To develop the base classifiers, we explored three deep learning architectures: (1) LSTM, based on the assumption that the posting temporal sequence of Reddit users could be leveraged for the discovery of patterns; (2) CNN, considering that the patterns of different complexity levels could help to distinguish between the disorders and (3) a hybrid CNN-LSTM architecture, in an attempt to combine their strengths. The adopted classifiers are the result of extensive experiments to define the representation of the input data, hyperparameters, and word embeddings. To the best of our knowledge, we outperform the state-of-the-art classifiers for specific disorders [Yates et al. 2017; Cohan et al. 2018].

To define the ensemble, we experimented to combine the base classifiers of *Level 0* according to two topologies (1) *homogeneous*, composed of models of the same architecture (e.g., CNN models only), dedicated to each target condition, and (2) *heterogeneous*, where we explored the strength of combining different architectures (LSTM, CNN, and Hybrid) in an ensemble. We also investigated the isolated contribution of comorbidity classifiers to the final prediction.

The implementation was developed using the Python 3.6 scientific package, Keras 2.2.5[5] and TensorFlow 1.14.0[6] backend, and its components as well as experiments are all available as supplementary material in a public repository[7].

[5]`https://keras.io/`
[6]`https://www.tensorflow.org/`
[7]`https://github.com/borbavanessa/stacking_ensemble`

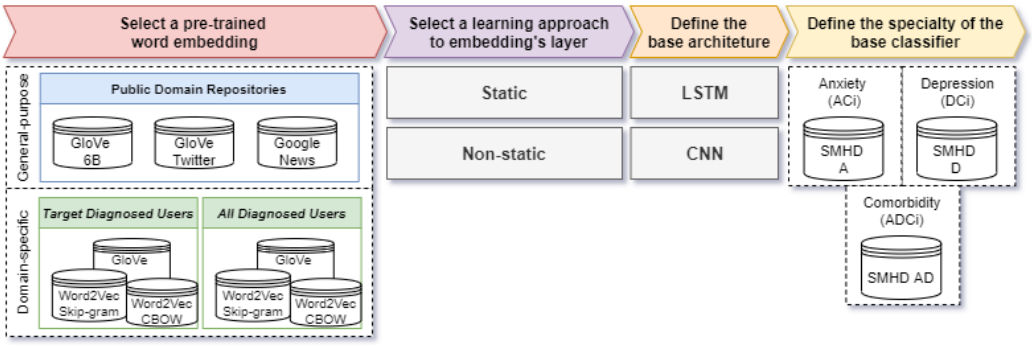


Fig. 2. The flow of choices for the formation of base classifiers, according to each word embedding.

5.1 *Level 0*: Base Classifiers.

The single-label binary classifiers targeted at each target condition were trained using the data sets listed in Table I: A (Anxiety), D (Depression), and AD (Comorbidity). We adopted three architectures (LSTM, CNN, and hybrid), and performed an extensive set of experiments that included many variations: different formats for input data; training parameters (batch size, training epochs, repetitions by training); general hyperparameters (number of neurons per layer, hidden layers, activation and loss functions) and specific ones (*return sequences*, *stateful* for LSTM; *filters*, *kernel size* for CNN); embeddings and learning models (random, static and non-static). Figure 2 shows the flow used to define the base models, exploring different pre-trained embeddings and deep learning architectures, which are further described in the remaining of this section.

5.1.1 *Word Embeddings.* We experimented with different pre-trained embeddings for the formation of base classifiers to generate variability between these models. This strategy was motivated by the premise that the use of embeddings, generated from different data sets, enables the recovery of complementary contexts for the same term, adding variability to the final decision of the ensemble. We adopted two types of embeddings, namely general-purpose and domain-specific, the later extracted from the SMHD corpus [Cohan et al. 2018].

As *general-purpose*, we adopted widely known word embeddings that were trained on huge volumes of datasets, namely GloVe 6B, GloVe Twitter, and Google News. We also generated two types of domain-specific embeddings from the SMHD corpus, referred to as *All Diagnosed Users* and *Target Diagnosed Users*. For the *All Diagnosed Users*, we considered the posts of all users diagnosed in the SMHD data set, so as to capture the vocabulary used by users diagnosed with many mental disorders, including anxiety and depression. For the *Target Diagnosed Users*, the embeddings generation is restricted to posts from users diagnosed only with the disorders targeted in this study. To generate the domain specific embeddings, we adopted both GloVe and Word2Vec algorithms. Table II shows the configuration used for the formation of the domain-specific embeddings.

5.1.2 *Architectures, Parameters and Hyperparameters.* For all architectures, the dense output layer was defined with three units, activation function *sigmod* and the loss function *binary crossentropy*. We maintained the default configuration for Kera's training algorithm, which manages the internal state and gradient estimate for updating the weights in the backward step. This model takes as input the concatenation of all users' posts, forming a single entry per user. This input is tokenized and typical normalization actions are applied (lower case conversion, removal of punctuation and special characters), resulting in a sequence of tokens. Only the most frequent 5000 tokens are considered. Regarding the pre-trained embeddings, we experimented both the static and non-static learning models.

Parameters specific to each architecture are described as follows:

Algorithm	Data Source (SMHD)	Parameter	Value
Word2Vec	*All Diagnosed Users* and *Target Diagnosed Users*	Embedding Size	300
		Window Size	5
		Minimun Count Frequence	5
		Number worders	6
		Algorithm	Skip-gram, CBOW
GloVe	*All Diagnosed Users* and *Target Diagnosed Users*	Window Size	5
		Number of components	300
		Learning rate	0,25
		Epochs	30
		Number of Threads	6

Table II. Domain-specific embeddings: Parameterization of the algorithms.

a) LSTM: All models that follow this architecture have an embedding layer, three LSTM layers with 16 units each, and *tanh* activation function. The common hyperparameters are: *recurrent dropout* activated on all LSTM layers (20%); *return sequence* activated on the first two layers; *Adam* optimization function (learning rate=0.001). We noticed a trade-off in terms of recall/precision when using different initialization functions (*Glorot Uniform* and *Lecun Uniform*). Table III details the specific embeddings used and respective learning model. As a convention, all base classifiers following the LSTM architecture have the prefix L, and the anxiety, depression and comorbidity target conditions are represented by the suffixes AC, DC and ADC, respectively.

b) CNN: It is composed of a pre-trained embedding layer, followed by a 1D convolution layer and an average pooling layer (each followed by a corresponding dropout layer). The main hyperparameters are: *relu* activation function, *Glorot Uniform* kernel initializer function; and *Ada Delta* optimization function (learning rate=0.001). The parameters that define the convolutional layer and the dropout rate are specific to each trainning dataset. The best performance results were achieved using: 250 *filters* for anxiety and depression, and 100 for comorbidity; *kernel size* equal to 4 for anxiety, 3 for depression, and 5 for comorbidity; and 50% *dropout* for depression and comorbidity, and 20% for anxiety. Table IV shows the performance of the models with the best results, with the striking parameters and hyperparameters. All base classifiers following the CNN architecture are depicted with the prefix C, and the suffixes AC, DC and ADC represent the target conditions.

c) Hybrid: This model is composed of a CNN network connected to another LSTM network. The number of layers and the type of activation function in each specific architecture was maintained, and the best results achieved using pre-trained embeddings; *Glorot Uniform* kernel initializer function; *Ada Delta* optimization function (learning rate=0.001). It was not possible to reach a single configuration for the classification of all disorders. Table V shows the best performance settings for each disorder. Thus, in the hybrid model, the CNN configuration that performed better contained 128 *filters* for anxiety and depression, and 64 for comorbidity; *kernel size* equal to 5 for all disorders; dropout of 20% anxiety, and 50% for depression and comorbidity. Likewise, the LSTM was tuned for each specific disorder: the number of units per layer equal to 64 for anxiety, 128 for depression, and 256 for comorbidity; and dropout set equal to 20% for all disorders. All base classifiers following the hybrid architecture are depicted with the prefix H, and the suffixes AC, DC and ADC represent the target conditions.

Tables III, IV, and V summarizes the best base classifiers to LSTM, CNN, and hybrid architectures, respectively, in terms of Precision (P), Recall (R), and F-measure (F). These are the average results obtained for diagnosed users, where each model was trained 5 times using the training/validation sets and assessed using the respective test set. Results depicted in bold highlight the best performance for each architecture.

When comparing the performance between all the explored architectures, we notice that the CNN models obtained the best performance for all disorders. In general, all models outperformed the results reported in [Cohan et al. 2018] for anxiety and depression, and the one reported in [Yates et al.

Disorder	Base Classifier Model	Word Embedding Type		Learning	Kernel Initializer	Performance		
		Domain Source	Algorithm			P	R	F1
Anxiety	L-AC$_1$	General Purpose (6B)	GloVe	Static	Glorot Uniform	0.73	0.62	0.67
	L-AC$_2$				**Lecun Uniform**	**0.73**	**0.69**	**0.71**
	L-AC$_3$	Targed Diagnosed Users			Glorot Uniform	0.62	0.79	0.70
Depression	L-DC$_1$	General Purpose (6B)	GloVe	Static	Glorot Uniform	0.75	0.77	0.76
	L-DC$_2$				**Lecun Uniform**	**0.74**	**0.79**	**0.77**
	L-DC$_3$	Targed Diagnosed Users	Word2Vec CBOW	Non-static	Glorot Uniform	0.67	0.81	0.73
Comorbidity	L-ADC$_1$	General Purpose (6B)	GloVe	Static	Glorot Uniform	**0.72**	**0.58**	**0.65**
	L-ADC$_2$				Lecun Uniform	0.67	0.64	0.66
	L-ADC$_3$	**All Diagnosed Users**	**Word2Vec CBOW**	**Non-static**	**Glorot Uniform**	**0.77**	**0.67**	**0.72**

Table III. LSTM Architecture: Final Performance of Base Classifiers

Disorder	Base Classifier Model	Word Embedding Type*		Configuration Variations			Performance		
		Domain Source	Algorithm	Filters	Kernel Size	Dropout	P	R	F1
Anxiety	C-AC$_1$	General Purpose (6B)	GloVe	250	5	0.5	0.77	0.75	0.76
	C-AC$_2$	General Purpose (Twitter)	GloVe	100	5	0.5	0.78	0.77	0.78
	C-AC$_3$	**All Diagnosed Users**	**Word2Vec****	**250**	**4**	**0.2**	**0.72**	**0.85**	**0.78**
Depression	C-DC$_1$	General Purpose (6B)	GloVe	250	4	0.2	0.76	0.81	0.79
	C-DC$_2$	**General Purpose (Twitter)**	**GloVe**	**250**	**3**	**0.5**	**0.72**	**0.87**	**0.79**
	C-DC$_3$	Targed Diagnosed Users	Word2Vec**	100	3	0.2	0.76	0.82	0.79
Comorbidity	**C-ADC$_1$**	**General Purpose (Google News)**	**Word2Vec**	**100**	**5**	**0.5**	**0.74**	**0.84**	**0.79**
	C-ADC$_2$	General Purpose (Twitter)	GloVe	250	3	0.5	0.74	0.84	0.78
	C-ADC$_3$	Targed Diagnosed Users	Word2Vec**	250	5	0.2	0.77	0.79	0.78

* All models presented better performance with static learning for the embedding layer.
** Word2Vec with Skip-gram algorithm.

Table IV. CNN Architecture: Final Performance of Base Classifiers

Disorder	Base Classifier Model*	CNN Cells			LSTM Cells		Performance		
		Filters	Kernel Size	Dropout	Units	Dropout	P	R	F1
Anxiety	**H-AC$_1$**	**128**	**5**	**0.2**	**64**	**0.2**	**0.71**	**0.78**	**0.74**
	H-AC$_2$	128	4	0.2	64	0.2	0.77	0.71	0.74
	H-AC$_3$	32	5	0.2	256	0.2	0.67	0.71	0.69
Depression	H-DC$_1$	128	5	0.5	128	0.2	0.73	0.80	0.76
	H-DC$_2$	**128**	**5**	**0.5**	**128**	**0.2**	**0.81**	**0.73**	**0.77**
	H-DC$_3$	128	5	0.5	128	0.2	0.70	0.73	0.71
Comorbidity	H-ADC$_1$	32	5	0.2	256	0.2	0.59	0.58	0.58
	H-ADC$_2$	**64**	**5**	**0.5**	**256**	**0.5**	**0.70**	**0.74**	**0.72**
	H-ADC$_3$	64	5	0.2	256	0.2	0.53	0.80	0.64

* All models were generated using the GloVe 6B general purpose word embedding type with static learning.

Table V. Hybrid Architecture: Final Performance of Base Classifiers

2017] for depression. Note that the results for depression are consistently higher, compared to the ones for the other conditions. This can be explained by two factors. First, data may be biased, as the automatic labeling technique displayed lower accuracy concerning anxiety (approximately 6% lower, considering the other conditions). Other possible explanations are that depression have more clear patterns, when compared to the other target conditions, or that anxiety traits are, in different levels, present in all these conditions, as defined by the DSM-5 manual.

5.2 Stacked Ensemble Architecture.

The proposed stacked ensemble architecture aims at benefiting from the strengths of each individual classifier to make decisions about specific disorders, producing class probabilities to be consolidated by the meta-learner as a multi-label, multi-class problem. The key design decisions were:

a) Topology: We experimented with different combinations for the best performing base classifiers presented in Tables III, IV, and V to compose the ensemble, according to two dimensions: model architecture (homogeneous vs heterogeneous), and inclusion of classifiers for comorbidity (dotted square in Figure 1). Thus, each topology includes 3 classifiers for each disorder.
Table VI specifies the combination of base classifiers used in each ensemble model topology explored. We use prefixes a convention to identify the topology regarding the architectures. For homogeneous

topologies, the prefixes are L, C and H, corresponding to the LSTM, CNN and Hybrid architectures respectively. The prefix LCH represents an heterogenous topology. When the topology includes comorbidity base classifiers, the ensemble includes the acronym CC (Comorbidity classifier). Therefore, an ensemble that contains only with CNN base classifiers, and which includes classifiers for detecting comorbidity.

Topology	Architecture	Ensemble Model	Base Classifiers by Disorder		
			Anxiety	Depression	Comorbidity
Homogeneous	LSTM	L-CC	L-AC_1, L-AC_2, L-AC_3	L-DC_1, L-DC_2, L-DC_3	L-ADC_1, L-ADC_2, L-ADC_3
		L	L-AC_1, L-AC_2, L-AC_3	L-DC_1, L-DC_2, L-DC_3	-
	CNN	C-CC	C-AC_1, C-AC_2, C-AC_3	C-DC_1, C-DC_2, C-DC_3	C-ADC_1, C-ADC_2, C-ADC_3
		C	C-AC_1, C-AC_2, C-AC_3	C-DC_1, C-DC_2, C-DC_3	-
	Hybrid	H-CC	H-AC_1, H-AC_2, H-AC_3	H-DC_1, H-DC_2, H-DC_3	H-ADC_1, H-ADC_2, H-ADC_3
		H	H-AC_1, H-AC_2, H-AC_3	H-DC_1, H-DC_2, H-DC_3	-
Heterogeneous	LSTM, CNN, Hybrid	LCH-CC	L-AC_2, C-AC_3, H-AC_1	L-DC_2, C-DC_2, H-DC_2	L-ADC_3, C-ADC_1, H-ADC_2
		LCH	L-AC_2, C-AC_3, H-AC_1	L-DC_2, C-DC_2, H-DC_2	-

Table VI. Ensemble Model: Topologies of the base classifiers for *Level 0*.

b) Meta-learner: The role of the meta-learner is to consolidate the predictions of the lowest level in terms of one or more labels. We implemented this level using a dense neural network, composed of different fully-connected perceptron layers. Hyperparameters (e.g., number of hidden layers, units per layer, batch size) were defined experimentally. The final configuration of the stacked ensemble is composed of three Dense layers. In each dense layer, activation function *tanh* and 12 units are configured for each base classifier present at *Level 0* ensemble. For the output layer, we maintain the same parameters defined in the output layer of the base classifiers.

c) Training: We assumed that each base classifier should not be influenced by the results generated by other individual classifiers, and thus should be trained independently one of another. This means that the *Level 1* dense network should be trained using a set of instances previously unknown, in order not to introduce bias. We trained it using the original test set, which was split into a proportion of 80% for training/validation, and 20% to test the ensemble results. To compensate for the smaller number of training instances, compared to the sets used to train the individual classifiers, we used cross-fold validation (*k-fold* = 5).

6. EXPERIMENTS

Our experiments aim to answer the following questions:

Q1: Does the choice of word-embeddings affect the performance of the base classifiers?

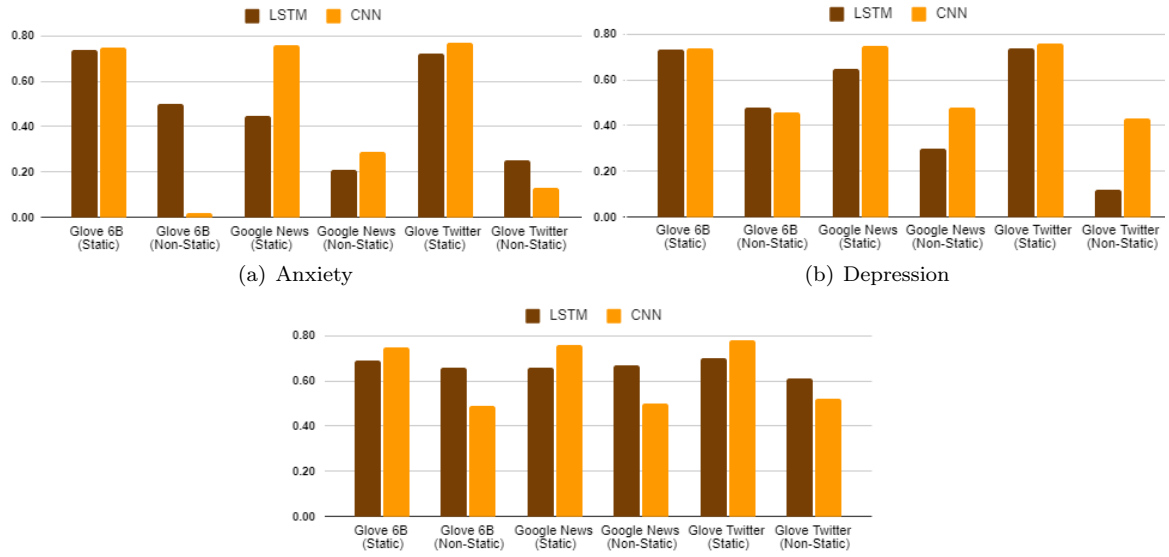
Q2: Which ensemble topology yields the best performance?

Q3: Are the influential features for the classification of the target conditions related to relevant symptoms?

Each one of these questions was addressed by a specific assessment, described in terms of method, results and discussions in the remaining of this section.

6.1 Word Embeddings Evaluation

a) Method. We aim to assess the impact of pre-trained embedding variation for the performance of the base classifiers, considering each architecture, and target condition. Our assessment included only the LSTM and CNN architectures, due to the training time required for the hybrid architecture. We developed 108 models, which result from the combination of distinct (1) pre-trained embeddings



(a) Anxiety

(b) Depression

(c) Comorbidity

Fig. 3. General-purpose pre-trained embeddings: Performance F1.

(general-purpose and domain-specific), (2) learning approach explored in the training (static and non-static), (3) architectures (LSTM and CNN), and (4) target condition of the base classifier (Anxiety, Depression and Comorbidity). Each model was trained 10 times, and the performance was measured in terms of the metrics Precision (P), Revocation (R), and measure F (F1). We considered the average results for our analysis. We established the following performance comparison: (1) general-purpose embeddings only; (2) domain-specific embeddings only, and (3) comparison between general-purpose and domain-specific using the best results. For performance comparison, we evaluated the statistical significance of pairs of models using the two-tailed Student T-Test ($alpha = 0.5$), where the null hypothesis formulated is that there is no difference between two models, concerning a specific parameter variation ($p - value >= 0.05$).

b) Results. Figure 3 shows the performance comparison between architectures in terms of average F1 for all general-purpose pre-trained embeddings considering each target condition. We observed that the best results were achieved with the static learning approach in both architectures in general. The LSTM classifiers achieved the best performance for all target conditions using GloVe 6B with static learning. Considering the F1 measure, although it presents results superior to all the tested embeddings, the differences were statistically significant concerning GloVe 6B with non-static learning for the Anxiety and Depression classifiers, being higher by 23 and 25 percentage points (pp), respectively. For the Comorbidity classifiers, it was also 3 pp higher, but this difference was not statistically significant. When compared to other pre-trained general-purpose embeddings using static learning, the models generated with GloVe 6B were statistically higher to Google News (3 pp for Comorbidity and 9 pp for Depression). For the CNN classifiers, the best result was achieved using GloVe Twitter (static) for all target conditions. This result was superior to each other general-purpose embeddings tested. The superiority of performance for the F1 measure was statistically significant when compared to the models generated with (1) GloVe Twitter (non-static) in all target conditions, where the biggest difference observed for the Anxiety classifiers (64 pp); (2) GloVe 6B static, where it was superior for the Anxiety and Comorbidity classifiers (both at 2 pp); and (3) Google News (static), where it was higher in 1 pp for all target conditions.

Figure 4 shows the performance comparison between architectures in terms of average F1 measurement for domain-specific pre-trained embeddings. Models using embeddings generated with Word2Vec performed better than those generated with GloVe in general. We observed a distinct behavior of Co-

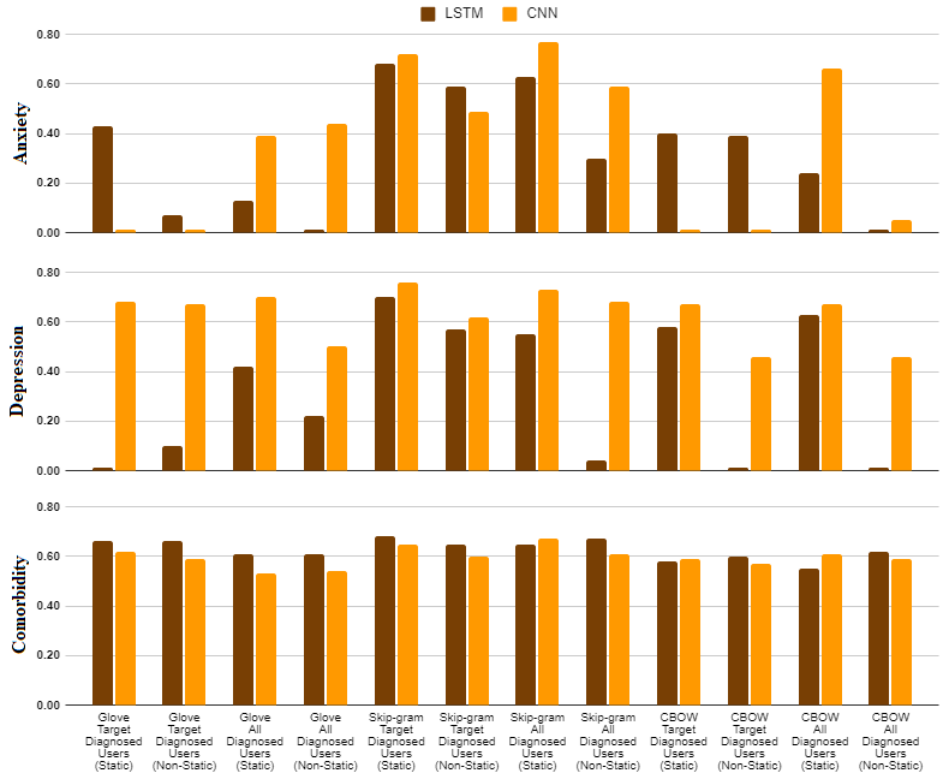


Fig. 4. Domain pre-trained embeddings: Performance F1.

morbidity classifiers, when compared to Anxiety/Depression classifiers. First, the performance is very similar, regardless the embeddings learning approach. Second, in general the LSTM architecture performed better, with a single exception (*All Diagnosed User*, generated with Word2Vec Skip-gram and CBOW). The opposite behavior was observed for the Depression classifiers, where the majority of domain-related embeddings presented better results for CNN architecture models.

Considering both learning approaches for all the domain-related embeddings tested, we noticed that the static learning showed statistically significant gains in different metrics for the Anxiety and Depression classifiers. For the Comorbidity classifiers, the difference between the learning approaches was not significant, except for the embedding *All Diagnosed User*, generated with the Word2Vec technique (CBOW algorithm), which showed better performance with non-static learning for F1 (7 pp) and precision (13 pp) metrics.

As for the technique used to generate the embeddings, we observed that Word2Vec with Skip-gram algorithm showed better performance for classifiers in both architectures. For the LSTM architecture, the embeddings generated from *Target Diagnosed Users* data source showed the best results for training the classifiers in all target conditions (static approach). Considering the F1 measure, the performance of these models was statistically significant, outperform the models formed with static approach and embeddings generated using the CBOW algorithm for both the *Target Diagnosed User* data source (in 28 pp for Anxiety, 12 pp for Depression and 10 pp for Comorbidity), and *All Diagnosed Users* (in 44 pp for Anxiety, 07 pp for Depression and 13 pp for Comorbidity). For the CNN architecture, the best result was observed with embedding generated from data source *All Diagnosed User* for Anxiety classifiers, whose performance for measure F1 was 5 pp higher that of models with *Target Diagnosed User* embedding. Inverse behavior was observed for Depression classifiers, whose performance for the F1 measure was 3 pp higher when using *Target Diagnosed User* embedding. For

the Comorbidity classifiers, the differences observed in favor of incorporating *All Diagnosed User* were not statistically significant.

We can observe distinct outcomes with the experimented variations, from which very few patterns can be derived. In general, the static learning approach showed the best results for all types of embeddings tested, considering the Anxiety and Depression classifiers. The Comorbidity classifiers, in general, presented similar results for both learning approaches, being in some cases superior to the non-static approach. In general, the best performance results were observed for the models using general-purpose embeddings generated with the GloVe technique in both architectures. Among domain-related embeddings, the Word2Vec technique (Skip-gram algorithm) showed the best results. Among the architectures, LSTM benefited the most by the variation of domain-related embeddings, especially for Comorbidity classifiers, which presented good results, both for the type of static and non-static learning. For the CNN architecture, few performance gains were observed, and it is not possible to derive a standard for the use of domain-related embeddings concerning each target condition.

We conclude that although the contribution of domain-related embeddings in terms of performance improvement is limited, they can contribute by highlighting different influential features for the classification. For this reason, models using domain-related embeddings that performed best in each target condition were selected for a training and fine-tuning step, where they achieved the performance shown in Tables III and IV.

6.2 Ensemble Performance Evaluation

a) Method. We aim to assess (1) the performance of a stacking ensemble for this multi-label classification task, (2) the impact of comorbidity classifiers in the ensemble, and (3) the contribution of using different architectures in the composition of the stacking ensemble model. Since we lack a baseline in the literature, we adapted the best configuration of each base architecture explored to a multi-class, multi-label prediction. All ensemble models were trained using the SMHD A-D-AD dataset. In the case of the baselines, we used the original training, validation, and testing sets. For the ensembles, each base classifier was trained using the respective SMHD A, SMHD D, and SMHD AD datasets, and the ensemble was trained using part of the SMHD A-D-AD test dataset, as described in Section 5.2.

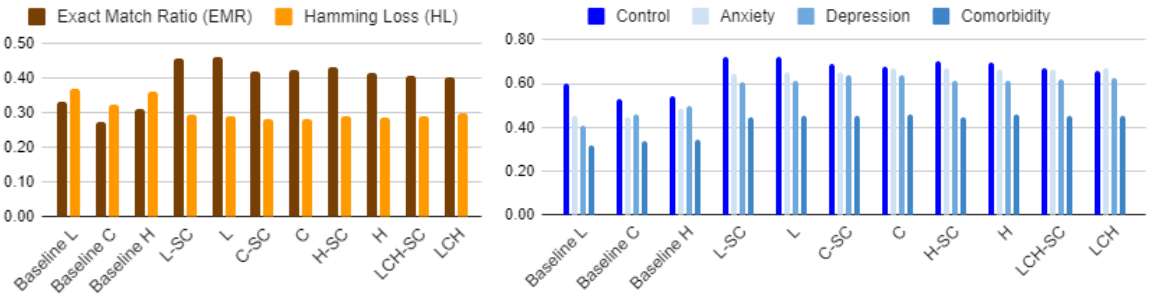
To assess the multi-label classification problem, we used both *Exact Match Ratio* (EMR), a harsh metric that measures the percentage of entirely correct labels assigned, and the *Hamming Loss* (HL), a soft metric that reports how many times, on average, a class label is incorrectly predicted. Finally, we calculated the F-measure (F), Recall (R), and Precision (P) for each class, in order to verify the ability to recognize characteristics of control and specific diagnosed users.

The reported results correspond to the average performance of each model over the SMHD A-D-AD test suite. This average was obtained by the number of repetitions resulting from the training process considering cross-validation (k-fold = 5). To check the statistical significance of the results obtained, we compared pairs of models using the two-tailed Student T-Test ($alpha = 0.5$). As a null hypothesis ($p-value >= 0.05$), it is adopted that there is no difference in performance between the two compared ensemble models, considering the variation proposed for the composition of *Level 0*.

Finally, for the best performing ensemble classifier, an analysis of the most frequent types of errors was performed, considering the performance of the classification over the entire set of test samples. To define the best performance ensemble model, we adopted the criterion best results for the EMR and HL metrics, associated with a balanced performance in terms of measure F for the classes representing the target conditions.

b) Results. Figure 5 summarizes the average results, using the same naming conventions described for Table VI. We also adopt the prefix L, C and H to denote the baselines creating using these LSTM,

CNN and Hybrid architectures, respectively. We observe that all ensemble models outperformed the three baselines, both in terms of HL (3 to 8 pp lower) and EMR (7 to 15 pp higher). We also observed gains in the F1 for all disorders, evenly distributed in terms of recall/precision.



(a) Multi-label Metrics (b) F-measure per Class

Fig. 5. Performance Ensemble Models

The ensemble model *C*, composed only by CNN base Anxiety and Depression classifiers achieved the best performance. This model is statistically superior to the *L* model by 2 pp for the Anxiety and Depression classes, and 1 pp for Comorbidity. Compared to the *H* model, *C* shows a loss of 2 pp for the Control class, but a performance gain of 1 pp for Anxiety and 2 pp for Depression. The performance of these models is comparable for Comorbidity. Compared to the *LCH* model, model *C* presented a superior performance of 2 pp for the Control class and 1 pp for the Anxiety class, and similar performance regarding the Comorbidity class. For these reasons, we regard the topology represented by *C* the best one for the ensemble *Level 0*.

The pair comparison of each combination of architectures with and without comorbidity classifiers reveals that, in general, the comorbidity base classifiers harm the performance, affecting mostly the classification of comorbidity (recall reduced in 1 to 4 pp). In general, the ensemble *C* is comparable to all *-CC* models in terms of EMR and HL, except for the *L-CC* model. This model was statistically superior to the ensemble *C* for the EMR and HL metrics, due to the performance presented for the Control class. However, when analyzing the performance for the target conditions between these models, we found that the ensemble *C* was significantly superior to the *L-CC* model in all classes.

The impact on the performance resulting from the combination of different architectures was analyzed by comparing the *LCH* ensemble with the other topologies without comorbidity (*L*, *C*, and *H*). In general, the *C* and *LCH* models achieved similar performances, where *C* yielded slightly higher results for the metrics related to the target conditions. We noticed an increase in the recall at the expense of precision when using the three different architectures, without a gain in terms of F1. The biggest differences were observed for Anxiety, where the *LCH* model promoted an increase in recall (8 pp), but a reduction in precision (4 pp).

Finally, we analyzed the types of errors performed by ensemble *C* in terms of (1) distinguishing between healthy and diagnosed users and (2) distinguishing among disorders. Regarding control/diagnosed users, the most common error (9% of test users) is related to predicting a diagnosed user as a control one. These errors are distributed as follows: anxiety 3%, depression 4%, and comorbidity 2%. The prediction of a control user as a diagnosed one is less frequent (7%). These errors are concentrated in the wrong prediction of a control user as a user diagnosed with comorbidity 5% or depression 2%. Regarding diagnosed users only, the most frequent error was to miss-classify a user with a single condition as a user presenting comorbidity (13% of the test users diagnosed with anxiety only, 12% of the test users diagnosed with depression only). Among the users diagnosed with comorbidity, 1% prediction errors were observed involving depression. Thus, *C* is able to identify most users with comorbidity (high recall), but with limitations in precision.

Disorder	Model	Relevant SHAP terms found in SD according to word embedding used to each model		
		Common Terms between Anxiety and Depression SD	Only in Anxiety SD	Only in Depression SD
Anxiety	C-AC$_1$, C-AC$_2$, C-AC$_3$	my, think, very, wish	really, know, tried, crazy, if, me, hell	love, probably, though
Depression	C-DC$_1$, C-DC$_2$, C-DC$_3$	because, cause, failed, feel, think	attempts, help, hope, you, if, me	anything, always, bad
Comorbidity	C-AC$_1$, C-AC$_2$, C-AC$_3$	ideas, mind, my, night, what	trying, really, cold, afraid, weird, me, will, beating	experience, terrible, much
	C-DC$_1$, C-DC$_2$, C-DC$_3$	ideas, mind, my, night, what	trying, really, cold, afraid, weird, crying, will	experience, much, obviously
Control	C-AC$_1$, C-AC$_2$, C-AC$_3$	something, think, what, mind	down, guys, if, instead, know, me, strange, why	calm, sorry, must
	C-DC$_1$, C-DC$_2$, C-DC$_3$	maybe, something, think, sure	strange, down, guys, instead, know, me, why	calm, sorry, if

Table VII. List of relevant SHAP terms for correctly classified samples

Considering homogeneous topologies, we conclude that there is a significant difference in performance between models composed only by the LSTM architecture when compared to those based on CNN or Hybrids. The LSTM architecture contributed more to the identification of the Control class, while the other architectures performed better to identify target conditions. Few significant differences were noted between CNN homogeneous and heterogeneous topologies. Considering the computational cost to train these models, the CNN homogeneous is more advantageous, since it achieves a similar performance at a lower cost. But it is worth investigating whether the restriction to a single deep learning architecture results in the necessary variability for a committee-type classifier. Finally, concerning the inclusion of Comorbidity classifiers proposed for the *-CC* variation, we are concluded that their presence did not promote performance gain. Furthermore, the presence of this classifier did not result in significant performance gains for the Comorbidity condition itself.

6.3 Qualitative Assessment

a) Method. This assessment aims at assessing if the ensemble makes classification decisions based on relevant features regarding the target conditions. We develop a method that employs SHAP to identify the influential features, and compare them to symptoms of the target conditions as defined by the DMS-5 manual [American Psychiatric Association 2013]. We calculated the 100 highest SHAP values for a sample of test users using KernelExplainer[8]. The SHAP values were calculated for each base classifier, as this library does not provide support for custom ensemble models.

Recall that due to the method used to compose the SMHD dataset without bias, all terms used to identify self-diagnosed users in SMHD dataset were removed from the *corpora* [Cohan et al. 2018], and thus there are no obvious words among the influential features. To relate these features to symptoms, we created a Symptom Dictionary (SD) with terms representing the symptoms of each disorder as described in the DMS-5 manual. To create the SD, we extracted the most frequent terms used in the DSM-5 manual use to define the symptoms of each disorder and validated them with the help of two psychologists. Each disorder was related to 59 terms, with 7 common terms between them. Then, we expanded these lists by including for each term the 20 closest words in all embeddingd set used (GloVe 6B, GloVe Twitter, *All Diagnosed Users*, *Target Diagnosed Users*). Finally, we created a function that retrieves terms according to embedding type used by the base classifier evaluated. The analysis is then performed considering the set of terms returned by these searches and their relationship with a description of symptoms reported in the DSM-5, so as to obtain insights on the patterns identified for each target condition.

b) Results. Table VII details the list of influential features found in each SD for a sample of users, and which are among the top-20 in the SHAP ranking. These features relate to the base classifiers used in ensemble C (Table VI). Thus, anxiety and depression are detailed by each base classifier, and comorbidity and control users by the union of all base classifiers. We can see that the terms of the SD could be related to many influential features in all class labels.

Users correctly classified as anxious are related to more Anxiety SD terms. For example, the terms "crazy" and "really" are close to "weird" (GloVe 6B), "scary", "sick" (GloVe Twitter), and "detached"

[8] `https://shap.readthedocs.io/en/latest/\#shap.KernelExplainer`

All Diagnosed Users), possibly indicating the fear of losing control; "tried" is close to "escape", "refuge" and "survived", which could indicate a state of extreme anxiety or panic attack.

Users correctly classified with depression are also related to Depression SD terms, or terms common to both disorders. For instance, "bad" is close to "situation", "worse", "terrible" and "suffer", which could indicate a concern about being negatively evaluated by other individuals. Among the common terms, we have "failed", which is associated with "problem", "insufficient", "inability" and "collapse" and could indicate excessive concerns about not being able to perform tasks, a symptom present in both disorders [American Psychiatric Association 2013].

Users who were correctly assigned the labels Anxiety and Depression are more related to SD terms that are common to both disorders. This number is higher if compared to anxious and healthy users, but smaller when compared to depressed users. For instance, "ideas" and "afraid", which are close to the terms "doubt" and "apprehensive" (Anxiety), as well as "avoid" and "frustrations", which would avoid doubts and frustrations (Depression), relate to symptoms observed in the Comorbidity [American Psychiatric Association 2013]. The influential features for these users are also related to terms specific to each list of the SD: (a) anxiety SD terms, such as the terms "weird" and "trying", close to "trouble", "worry" and "danger", and could indicate feelings of excessive concern about danger, feelings strongly present for Anxious Disorders (e.g., Generalized Anxiety Disorder); and (b) depression SD terms such as "terrible", which are associated with "sad", "melancholy" and "lonely", feelings typical of depressed users [American Psychiatric Association 2013].

Healthy users correctly classified as Control are evenly related to terms in all lists of the SD, although in smaller quantities when compared to samples of diagnosed users. Nevertheless, the models seem to have learned differentiating patterns between healthy users and those diagnosed with one of the disorders, according to the context in which these terms are presented. The analysis of the term "if", present in samples of anxiety and depression, revealed that its meaning changes according to the dictionary of the disorder. For anxiety, the term appears associated with "apprehensive", "leery", and "hesitant", whereas for depression the same term is associated with "change", "aggression" and "hostility".

Although each base classifier is associated with the respective list of SD terms, we noticed that the terms related to Anxiety are influential features in all classifiers (diagnosed and control users). According to [American Psychiatric Association 2013], anxiety contains signs that are present in different ways and various types of disorders, including healthy people at acceptable levels. This behavior is thus consistent with the influential features used by the base classifiers.

On the other hand, we noticed that some users, although correctly classified by the respective set of base classifiers, are miss-classified by C as comorbidity, being assigned a second (wrong) disorder label. This explains the high recall and low precision observed for comorbidity in C.

To understand this behavior, we examined how users correctly classified as anxious by the set of Anxiety classifiers are handled by the Depression classifiers, and vice-versa. For instance, for one anxious user also classified as depressed, we identified the term "worry" (close to "fear" and "avoid"). This term could represent excessive concerns about avoiding exposure situations, for fear of being evaluated negatively by other individuals. This behavior differs between anxiety and major depression disorders. While for anxious people this concern is based on specific social behaviors or physical symptoms, for depressed individuals this concern arises from the feeling of considering themselves as bad people or not worthy of being appreciated [American Psychiatric Association 2013].

Conversely, considering a user correctly classified as depressed by the base classifiers, but also as anxious by C, we noticed that the term "feeling" (close to "frustration" and "anxiety") was considered influential by the Anxiety classifiers. These feelings could represent excessive concerns about not being able to perform tasks, a common behavior in anxious individuals, or a specific type of depression, Persistent Depressive Disorder (Dysthymia), which presents a high risk of comorbidity with other

disorders, including anxiety [American Psychiatric Association 2013].

We conclude that the ensemble does use significant features to make decisions about the target conditions. The analysis reveals that for some manifestations of anxiety and depression disorders, the symptoms can be very similar, but motivated for different reasons. This is evidence that we need to find more subtle differentiating patterns between disorders, according to the context in which the symptom is expressed. These are valuable insights for improving the ensemble with other organization of classifiers, or different types of base classifiers, such as models targeted at distinguishing diagnosed users according to the disorders.

7. CONCLUSION

In this paper, we proposed a stacking ensemble targeted at the automatic identification of depression, anxiety and comorbidity. The *Level 0* is composed of base classifiers that distinguish between control and diagnosed users, and the *Level 1* explores these probabilities for a multi-class, multi-label prediction. To define the base classifiers, we experimented with alternative architectures (LSTM, CNN, and their combination) and word embeddings. For the ensemble, we assessed the topology of base classifiers and the influence of comorbidity classifiers to distinguish between the disorders or their association. Our work fills an important gap in the automatic classification of disorders by addressing another prevalent disorder, anxiety, and comorbidity with depression.

The main findings regarding our quantitative assessment is that both the combination of different pre-trained embeedings, as well as the combination of different deep learning architectures positively contributes to the ensemble performance; in general, the inclusion of comorbidity classifiers degraded the performance; and despite base classifiers based on pre-trained embeddings performed better than their counterpart trained using domain-specific embeddings, the consideration of both in the ensemble increase the variability of the solution.

The qualitative assessment revealed strong points of our solution and issues that need to be improved. First, we confirmed that meaningful features do influence the base classifiers' predictions related to these disorders. Second, it has re-enforced the importance of varying the embeddings, since it results in a broader range of contexts to be considered by the ensemble. It also confirmed that CNN architecture combined with different word embeddings, or combined with other architectures increased the performance of the proposed ensemble solution. Most importantly, it revealed the strong influence of anxiety in the decisions taken by the ensemble. The fact that characteristics of the anxiety disorder are present at some intensity level in all users, including depressed and healthy individuals, actively contributes to the difficulty of distinguishing between diagnosed users. This characteristic suggests the need to identify more subtle patterns that differentiate the presence of anxiety signs according to their intensity, helping to distinguish between a specific disorder or their comorbidity.

Future work will focus on the identification of the differentiating patterns between comorbidity and anxiety/depression disorders. It includes, among others, (a) formation to binary base classifiers to distinguish between anxiety and depression, (b) explore alternative architectures to increase the variability in the set solution, such as Bidirectional Encoder Representations from Transformers (BERT), which became the state of art for a wide range of natural language processing tasks, and (c) different topologies for the ensemble to combine the contexts of the base classifiers.

REFERENCES

American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. Autor, Washington, DC, 2013.

Amora, P. R. P., Teixeira, E. M., Lima, M. I. V., Amaral, G. M., Cardozo, J. R. A., and Machado, J. d. C. An analysis of machine learning techniques to prioritize customer service through social networks. *Journal of Information and Data Management* 9 (2): 135—-146, 2018.

Bagroy, S., Kumaraguru, P., and De Choudhury, M. A social media based index of mental well-being in college campuses. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Association for Computing Machinery, New York, NY, USA, pp. 1634–1646, 2017.

Becker, K., Harb, J. G., and Ebeling, R. Exploring deep learning for the analysis of emotional reactions to terrorist events on twitter. *Journal of Information and Data Management* 10 (2): 97–115, 2019.

Benton, A., Mitchell, M., and Hovy, D. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, pp. 152–162, 2017.

Chollet, F. What is deep learning? In *Deep Learning with Python*. Manning, 1, pp. 8–22;94–96,102–104,123;184–185,196–197,202–206,215–216;264–266, 2017.

Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., and Goharian, N. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, E. M. Bender, L. Derczynski, and P. Isabelle (Eds.). Association for Computational Linguistics, pp. 1485–1497, 2018.

De Choudhury, M., Counts, S., Horvitz, E. J., and Hoff, A. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work amp; Social Computing*. CSCW '14. Association for Computing Machinery, New York, NY, USA, pp. 626–638, 2014.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., and Kumar, M. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. Association for Computing Machinery, New York, NY, USA, pp. 2098–2110, 2016.

De Choudhury, M., Sharma, S. S., Logar, T., Eekhout, W., and Nielsen, R. C. Gender and cross-cultural differences in social media disclosures of mental illness. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. Association for Computing Machinery, New York, NY, USA, pp. 353–369, 2017.

Dutta, S., Ma, J., and De Choudhury, M. Measuring the impact of anxiety on online social interactions. In *ICWSM*. ICWSM'18. The AAAI Press, 2018.

Gama, J., Faceli, K., Lorena, A., and De Carvalho, A. *Inteligência artificial: uma abordagem de aprendizado de máquina*. Grupo Gen - LTC, 2011.

Giuntini, F. T., Cazzolato, M. T., dos Reis, M. d. J. D., Campbell, A. T., Traina, A. J. M., and Ueyama, J. A review on recognizing depression in social networks: challenges and opportunities. *Journal of Ambient Intelligence and Humanized Computing* (1868-5145, 2020.

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J. P., Dobson, R. J. B., and Dutta, R. Characterisation of mental health conditions in social media using informed deep learning. *Scientific Reports* vol. 7, pp. 2045–2322, 2017.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems* 28 (10): 2222–2232, Oct, 2017.

Gruda, D. and Hasan, S. Feeling anxious? perceiving anxiety in tweets using machine learning. *Computers in Human Behavior* vol. 98, pp. 245 – 255, 2019.

Hamilton, M. Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology* 6 (4): 278–296, 1967.

Harb, J. G., Ebeling, R., and Becker, K. Exploring deep learning for the analysis of emotional reactions to terrorist events on twitter. *J. Inf. Data Manag.* 10 (2): 97–115, 2019.

Hirschfeld, R. The comorbidity of major depression and anxiety disorders: Recognition and management in primary care. *Prim Care Companion J Clin Psychiatry* 3 (244-254), 12, 2001.

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

Ireland, M. and Iserman, M. Within and between-person differences in language used across anxiety support and neutral Reddit communities. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Association for Computational Linguistics, New Orleans, LA, pp. 182–193, 2018.

Ive, J., Gkotsis, G., Dutta, R., Stewart, R., and Velupillai, S. Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Association for Computational Linguistics, New Orleans, LA, pp. 69–77, 2018.

Kim, Y. Convolutional neural networks for sentence classification. In *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) - ACL*. pp. 1746–1751, 2014.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., and Brown, D. E. Text classification algorithms: A survey. *Inf.* 10 (4): 150, 2019.

Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., Cai, L., and Feng, L. User-level psychological stress detection from social media using deep neural network. In *Proceedings of the 22nd ACM International Conference on Multimedia*. MM '14. Association for Computing Machinery, New York, NY, USA, pp. 507–516, 2014.

Lundberg, S. M. and Lee, S. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems : Proc. of the 30th Annual Conf. on Neural Information Processing Systems (NIPS)*, I. Guyon, U. von Luxburg, and et alli (Eds.). pp. 4765–4774, 2017.

Mann, P., Paes, A., and Matsushima, E. H. See and read: Detecting depression symptoms in higher education students using multimodal social media data. In ICWSM. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (1): 440–451, May, 2020.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., pp. 3111–3119, 2013.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. Deep learning based text classification: A comprehensive review. *CoRR* vol. abs/2004.03705, 2020.

Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

Park, S., Kim, I., Lee, S. W., Yoo, J., Jeong, B., and Cha, M. Manifestation of depression and loneliness on social networks: A case study of young adults on facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work Social Computing*. CSCW '15. Association for Computing Machinery, New York, NY, USA, pp. 557–570, 2015.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1532–1543, 2014.

Radloff, L. S. The ces-d scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement* 1 (3): 385–401, 1977.

Sharma, E. and De Choudhury, M. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Association for Computing Machinery, New York, NY, USA, pp. 1–13, 2018.

Shen, J. H. and Rudzicz, F. Detecting anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Vancouver, BC, pp. 58–65, 2017.

Souza, V. B., Nobre, J. C., and Becker, K. Characterization of Anxiety, Depression, and their Comorbidity from Texts of Social Networks. In *Anais do XXXV Simpósio Brasileiro de Banco de Dados*. SBC, Porto Alegre, RS, Brasil, 2020.

Tadesse, M. M., Lin, H., Xu, B., and Yang, L. Detection of depression-related posts in reddit social media forum. *IEEE Access* vol. 7, pp. 44883–44893, 2019.

Tiller, J. Depression and anxiety. *The Medical journal of Australia* vol. 199, pp. S28–31, 09, 2013.

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

Wongkoblap, A., Vadillo, M., and Curcin, V. Researching mental health disorders in the era of social media: Systematic review. *Journal of Medical Internet Research* 19 (6), 2017.

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

Zhang, C. and Ma, Y. *Ensemble Machine Learning: Methods and Applications*. Springer Publishing Company, Incorporated, 2012.

Zhang, L., Wang, S., and Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 8 (4), 2018.