# Characterization of Anxiety, Depression, and their Comorbidity from Texts of Social Networks

**Vanessa B. Souza**[1], **Jeferson Nobre**[1], **Karin Becker**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{vbsouza, jcnobre, karin.becker}@inf.ufrgs.br

***Abstract.*** *Depression has become a public health issue, and the high comorbidity rate with anxiety worsens the clinical picture. Early identification is crucial for decisions on the proper line of treatment. The use of social networks to expose personal difficulties has enabled works on the automatic identification of specific mental conditions, particularly depression. This paper explores deep learning techniques to develop an ensemble stacking classifier for the automatic identification of depression, anxiety, and their comorbidity, using a self-diagnosed dataset extracted from Reddit. At the lowest level, binary classifiers make predictions about specific disorders, outperforming all baseline models. A meta-learner explores these weak classifiers as a context for reaching a multi-label decision, achieving a Hamming Loss of 0.29 and Exact Match Ratio of 0.47. We performed a qualitative analysis using SHAP, which confirmed the relationship between the influential features and symptoms of these disorders.*

## 1. Introduction

Depression is a mood state that 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]. The World Health Organization[1] estimates depression affects near 322 million worldwide. Another prevalent disorder worldwide is anxiety, an emotion characterized by feelings of tension, excessive fear, recurring intrusive thoughts, and physiological changes [American Psychiatric Association 2013]. It includes disorders that share characteristics of excessive fear and anxiety for several domains (e.g. violence, profession). Physical symptoms include restlessness or tension, difficulty in concentrating, irritability, muscle tension, and sleep disorders.

Studies provide evidence of the close relationship between anxiety and depression. The comorbidity rate of anxiety and depression is high since 85% of patients with depression also experience significant symptoms of anxiety [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 in society requires prevention and intervention strategies, particularly concerning screening and early diagnosis [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].

The extensive use of social networks have promoted opportunities to deploy computational solutions to support the studies of mental disorders. To avoid prejudice, or

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

merely as a means to seek for help, people have used social networks to expose their difficulties anonymously. This results in an increasing volume of high-value data that can be explored to automatically recognize mental disorders and its diagnostic criteria and the discovery of interactions among such disorders. 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 related work focuses on the automatic identification of specific disorders using supervised learning techniques.

The automatic classification of depression is the focus of most works using supervised learning algorithms on textual, social, and sentiment features extracted from data using extensive feature engineering (e.g., [Tsugawa et al. 2015, De Choudhury et al. 2017, Park et al. 2015]). More recently, deep learning techniques have been explored for the classification of depression [Yates et al. 2017, Mann et al. 2020]. Deep learning has the benefit of including the extraction of data representations from input data as part of the learning process [Murphy 2012]. Few works address anxiety [Dutta et al. 2018] and its comorbidity with depression [Cohan et al. 2018]. Our research addresses the automatic classification of depression, anxiety, and their comorbidity, with the aim of contributing with insights about the common and differentiating patterns that can be derived from textual social interaction.

In this paper, we take initial steps towards this goal by proposing an ensemble classifier for the automatic identification of depression, anxiety, and their comorbidity, using a self-diagnosed dataset extracted from Reddit [Cohan et al. 2018]. The use of a stacking ensemble 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, which predict class probabilities related to diagnosed/control users of a specific target condition. 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. The weak classifiers are variations of a base Long Short-Term Memory (LSTM) deep learning architecture, such that the meta-learner level can combine their strengths. The paper details these architectural choices.

We developed experiments to assess quantitatively and qualitatively the proposed solution. The single-label, binary LSTM classifiers outperform existing solutions on Reddit data [Yates et al. 2017, Cohan et al. 2018], achieving F-measures of 0.77 for depression, 0.71 for anxiety, and 0.72 for their comorbidity. Our ensemble solution achieved encouraging results (Hamming Loss of 0.29 and Exact Match Ratio of 0.47), outperforming a multi-class, multi-label baseline. To qualitatively assess the models, we adopted Shapley Additive Explanation (SHAP) [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 we could relate many influential features to symptoms described in Diagnostic and Statistical Manual of Mental Disorders (DSM-5) psychology manual [American Psychiatric Association 2013], which describes disorders and their symptoms.

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

## 2. 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. These works applied shallow learning algorithms such as Support Vector Machine (SVM) or regression on data resulting from extensive features engineering. These 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 posts, and additionally consider other features such as the frequency and morphological structure of words, sentiment, and social features [De Choudhury et al. 2017, Dutta et al. 2018, Park et al. 2015, Sharma and De Choudhury 2018, Tsugawa et al. 2015].

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 Convolutional Neural Network (CNN) architecture that summarizes users' posting activities as vectors, followed by dense layers that perform user classification. A multimodal depression classifier that takes Instagram data as input is reported in [Mann et al. 2020]. It combines an ELMo model to process the textual content, and a ResNet model for image processing. The reported F-measure performance of these classifiers are 0.65 and 0.75, respectively, but the results are not comparable due to the distinct characteristics of Reddit and Instagram posts.

Very few works address the comorbidity of mental disorders. Using Reddit data, [Cohan et al. 2018] developed classifiers for nine (9) mental conditions, 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 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. Only the posts written by the user (i.e. submission and comments) are considered.

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

## 3. 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[2]. We used only data related

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

**Table 1. Datasets derived from SMHD for the experiments.**

| 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 | Anxiety,Depression | 1,320 | 191,056 |
| | | Control | 1,320 | 390,892 |

| Dataset | Type dataset | Classes | Total Users | Total Posts |
|---|---|---|---|---|
| **SMHD A-D-AD** | multi-label | Anxiety | 1,320 | 202,370 |
| | | Depression | 1,320 | 195,711 |
| | | Anxiety,Depression | 1,320 | 191,056 |
| | | Control | 2,640 | 764,174 |

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 1. 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.

## 4. An Ensemble Architecture for the Identification of Depression, Anxiety and Comorbidity

In this section, we describe a stacking ensemble classifier targeted at identifying depression, anxiety, and comorbidity. The ensemble is a means to leverage and combine the generalization power of distinct models targeted at specific disorders. The individual weak models predict class probabilities related to diagnosed/control users of a specific target condition. In this way, we can combine the strengths of distinct single-label binary models, by exploring variations of a base architectural choice. To consolidate all these individual predictions into a final multi-class, multi-label prediction, the meta-learner level is represented by a dense-neural network. Figure 1 outlines the architecture of the stacking ensemble, where $\text{AC}_i$, $\text{DC}_i$, and $\text{ADC}_i$ ($i > 0$) are binary classifiers for Anxiety, Depression and comorbidity, respectively.

The design choices involved the two levels of the ensemble. The decisions regarding *Level 0* involved the weak classifiers. We explored the LSTM architecture under the assumption that the posting temporal sequence of SMHD users could be leveraged for the discovery of patterns. We developed experiments to define the representation of the input data, hyperparameters, and word embeddings. To the best of our knowledge, the resulting LSTM-based models outperform the state-of-the-art models reported in the literature for specific disorders [Yates et al. 2017, Cohan et al. 2018]. The meta-learner level (*Level 1*) involved experiments to define the best topology and hyperparameters for the dense neural network. In particular, we investigated the contribution of the comorbidity classifiers to this multi-label, multi-class classification task.

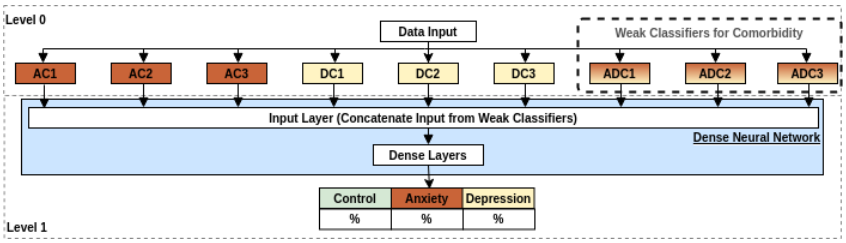


**Figure 1. Architecture of the Stacking Ensemble Classifier**

The remaining of this section describes these design choices and the results of the experiments. The implementation was developed using the Python 3.6 scientific package and the libraries for the machine learning project (Keras 2.2.5[3] with TensorFlow 1.14.0[4] backend). The implementation of the ensemble and its components, as well as the Jupyter experiments notebooks, are all available as supplementary material in a public repository[5].

### 4.1. *Level 0*: Weak Classifiers

#### 4.1.1. Base LSTM Architecture

We adopted an LSTM architecture to identify patterns that allow us to distinguish between control and self-diagnosed users, using single-label, binary classifiers for each specific situation: anxiety, depression, and comorbidity of these conditions. For this purpose, we used the A, D, and AD datasets listed in Table 1.

The base LSTM architecture is composed of one embedding layer (pre-trained Glove 6B[6] - 300 dimensions - with static learning model), three LSTM layers with 16 units each, and *tanh* activation function. We use a Dense output layer with three units, *sigmod* activation function and *binary crossentropy* loss function. For optimization function, we use *Adam* with learning rate of 0.001. The *recurrent dropout* hyperparameter is activated on all LSTM layers, while the *return sequence* is activated on the first two layers. 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. We experimented many other variations of this architecture, including: input as a sequence of posts; variations for training (batch size, training epochs, repetitions by training); hyperparameters (return sequences and stateful, number of neurons per layer, hidden layers, activation and loss functions); other pre-trained embeddings (Glove Twitte[6], Google News[7]); and embedding learning models (random, static and non-static).

The variation that yielded differences in the results was achieved using a distinct kernel function for network initialization, which is responsible for initializing the "kernel" weight matrix, used for the linear transformation of the inputs. A proper kernel function can reduce/avoid problems related to network convergence during learning (e.g. the explosion/decay of the gradients). Glorot and LeCun initialization functions differ in terms of the distribution function used to generate the set of values. Although the F-measures are relatively similar, Glorot affects more the precision, whereas LeCun influences the recall. We hypothesized that these functions are complementary solutions to establish a trade-off between recall and precision when considered by the meta-learner level.

Table 2 shows the performance of the base LSTM architecture (models AC1, DC1, and ADC1) and its variation using the Lecun kernel function (models AC2, DC2, and ADC2). 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

[3]https://keras.io/
[4]https://www.tensorflow.org/
[5]https://github.com/borbavanessa/deep-learning-for-mental-health
[6]https://nlp.stanford.edu/projects/glove/
[7]https://code.google.com/archive/p/word2vec/

**Table 2. Performance of the *Level 0* LSTM topologies**

| Disorder | Weak Classifier LSTM Model | Word Embedding Type: Domain Souce | Word Embedding Type: Algoritm | Word Embedding Type: Learning | Kernel Initializer Function | P | R | F1 |
|---|---|---|---|---|---|---|---|---|
| **Anxiety** | AC1 (base) | General Purpose (6B) | Glove | Static | glorot uniform | 0.73 | 0.62 | 0.67 |
| | AC2 | | | | lecun uniform | 0.73 | 0.69 | 0.71 |
| | AC3 | *Targed Diagnosed Users* | | | glorot uniform | 0.62 | 0.79 | 0.70 |
| **Depression** | DC1 (base) | General Purpose (6B) | Glove | Static | glorot uniform | 0.75 | 0.77 | 0.76 |
| | DC2 | | | | lecun uniform | 0.74 | 0.79 | 0.77 |
| | DC3 | *Targed Diagnosed Users* | Word2Vec CBOW | Non-static | glorot uniform | 0.67 | 0.81 | 0.73 |
| **Anxiety, Depression** | ADC1 (base) | General Purpose (6B) | Glove | Static | glorot uniform | 0.72 | 0.58 | 0.65 |
| | ADC2 | | | | lecun uniform | 0.67 | 0.64 | 0.66 |
| | ADC3 | *All Diagnosed Users* | Word2Vec CBOW | Non-static | glorot uniform | 0.77 | 0.67 | 0.72 |

**Table 3. Performance of domain-related word embeddings.**

| Classifier | Domain Source | General Purpose | | | All Diagnosed Users | | | | | | | | | Target Diagnosed Users | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WE Algorithm | Glove 6B | | | Word2Vec Skip-gram | | | Word2Vec CBOW | | | Glove | | | Word2Vec Skip-gram | | | Word2Vec CBOW | | | Glove | | |
| **Anxiety** | **Learning** | Static | | | Non-Static | | | Non-Static | | | Non-Static | | | Static | | | Static | | | Static | | |
| | **Metrics** | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| | | 0.73 | 0.62 | 0.67 | -0.10 | -0.03 | -0.06 | -0.03 | 0.02 | 0.00 | -0.12 | -0.02 | -0.06 | 0.02 | 0.01 | 0.02 | -0.10 | -0.02 | -0.06 | -0.11 | 0.17 | 0.03 |
| **Depression** | **Learning** | Static | | | Static | | | Static | | | Static | | | Static | | | Non-Static | | | Non-Static | | |
| | **Metrics** | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| | | 0.75 | 0.77 | 0.76 | 0.05 | -0.14 | -0.05 | -0.17 | -0.10 | -0.13 | -0.13 | -0.11 | -0.12 | 0.00 | -0.09 | -0.05 | -0.08 | 0.04 | -0.03 | 0.00 | -0.13 | -0.07 |
| **Anxiety. Depression** | **Learning** | Static | | | Non-Static | | | Non-Static | | | Non-Static | | | Static | | | Static | | | Non-Static | | |
| | **Metrics** | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| | | 0.72 | 0.58 | 0.65 | 0.01 | 0.08 | 0.04 | 0.05 | 0.09 | 0.07 | 0.03 | -0.07 | -0.04 | -0.02 | 0.05 | 0.02 | -0.15 | 0.12 | -0.02 | -0.13 | 0.02 | -0.05 |

test set. We outperformed the results reported in [Cohan et al. 2018] for anxiety and depression, and the one reported in [Yates et al. 2017] for depression. The proposed LSTM architecture performed consistently for all self-diagnosed scenarios, with F-measure ranging from 0.65 to 0.77, and a good balance between precision and recall.

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). Another explanation may be that depression is characterized by more clear behavioral patterns, compared to anxiety.

### 4.1.2. Domain-related Word Embeddings

We also investigated whether the use of domain-specific word embeddings would improve the classifiers. To generate them, we deployed both Glove and Word2Vec (Skip-gram and CBOW) algorithms on the SMHD dataset data. The embeddings dictionaries were generated considering two approaches: (1) posts from all diagnosed users in the datasets SMHD (*All Diagnosed Users*) and (2) posts from diagnosed users in the datasets of Table 1 (*Target Diagnosed Users*). Table 3 shows the results in terms of differences regarding the base LSTM model, represented in the column *General Purpose*. The table displays the best results obtained for each embedding extraction algorithm in terms of Precision (P), Recall (R), and F-measure (F) for each dataset and classifier. The *Learning* rows indicate learning model (static or non-static). All other hyperparameters are the same.

The highlighted cells indicate the best two results for each disorder classifier, using the F-measure as the ranking metric. The comorbidity classifier improved the most: up to 7 pp (percentage points) in the F-measure, with an equal balance between recall and precision. However, the performance of the Depression classifiers was inferior in all but one case, when a slight increase in the recall was observed (4 pp). Finally, the classifier for Anxiety presented some improvements. In the most significant case (3 pp in the F-measure), we observed a significant increase in the recall at the expense of precision.

We observed distinct outcomes in the experimented variations, from which very few patterns could be derived. In general, Word2Vec presented a better result in the *All Diagnosed Users* dataset, and Glove in the *Target Diagnosed Users* dataset. Anxiety and comorbidity classifiers performed better using a non-static model when using the *All Diagnosed Users* dataset, and non-static when using the other one. This behavior was inverse in the case of the Depression classifiers. Regarding the domain data source, the best results for the classifiers of isolated disorders were obtained using the *Target Diagnosed Users* dataset, while comorbidity showed gain using the dictionary composed of *All Diagnosed Users*. 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.

### 4.2. Stacked Ensemble Architecture

The ensemble proposed 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) the *Level 0* classifiers that generate these probabilities; and (b) the meta-learner architecture for consolidating the predictions in terms of one or more labels.

The choices for the *Level 0* include two orthogonal decisions regarding the weak classifiers: which diagnosed users should be represented and which binary classifier topologies should be used. Regarding the first aspect, we generated two variations of the ensemble topology: one that encompasses only Depression and Anxiety classifiers, and another that also includes comorbidity classifiers (dotted square in Figure 1). This choice influences on the contexts that are provided for the upper level to learn the weights for a multi-class, multi-label decision. The second aspect concerns the architectural choices for the binary weak classifiers. We chose the best three solutions derived from our extensive set of experiments described in Section 4.1, and which are summarized in Table 2.

We implemented the meta-learner (*Level 1*) using a dense neural network, composed of different fully-connected perceptron layers. The number of hidden layers and units per layer, batch size for training, and number of training seasons, and other hyperparameters, 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 weak classifier present at *Level 0* ensemble. For the output layer, we maintain the same parameters defined in the output layer of the weak classifiers.

An important issue is how to train the ensemble. We assumed that each weak 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 in the results. Therefore, 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. We set the *k-fold* = 5, given that each experiment could take as much as hundreds of hours.

## 5. Experiments

### 5.1. Performance Evaluation

**a) Method.** We aim to assess (1) the performance of the proposed stacking ensemble model for this multi-label multi-classification task, and (2) the impact of including classi-

**Table 4. Ensemble results: *Ensemble 1* (A-D-AD), *Ensemble 2* (A-D), Baseline (multi-label, multi-class base LSTM)**

| Model | Correct Prediction per Label | | | EMR | Hamming Loss | Control | | | Anxiety | | | Depression | | | Anx.,Depr. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Control | Anx. | Depr. | | | P | R | F | P | R | F | P | R | F | P | R | F |
| **Ensemble 1** | 0.75 | 0.70 | 0.66 | 0.45 | 0.29 | 0.67 | 0.77 | 0.71 | 0.61 | 0.71 | 0.66 | 0.57 | 0.65 | 0.61 | 0.33 | 0.73 | 0.44 |
| **Ensemble 2** | 0.77 | 0.70 | 0.67 | 0.47 | 0.29 | 0.69 | 0.77 | 0.73 | 0.59 | 0.75 | 0.66 | 0.57 | 0.66 | 0.61 | 0.33 | 0.75 | 0.45 |
| **Baseline** | 0.63 | 0.63 | 0.62 | 0.31 | 0.39 | 0.53 | 0.61 | 0.57 | 0.57 | 0.29 | 0.38 | 0.55 | 0.24 | 0.33 | 0.30 | 0.27 | 0.29 |

fiers targeted at the comorbidity in the ensemble. We adopted as baseline the base LSTM architecture adapted for a multi-class, multi-label, since we lack a baseline in the literature. Recall the one in [Cohan et al. 2018] encompasses the 9 conditions, and thus it cannot be compared. We produced two variations of the ensemble depicted in Figure 1, referred to as *Ensemble 1* (weak classifiers for specific conditions and their comorbidity) and *Ensemble 2* (weak classifiers for depression and anxiety only). All models were trained using the SMHD A-D-AD dataset. In the case of the baseline, we used the original training, validation, and testing sets. For the ensembles, each weak classifier was trained using the respective SMHD A, SMHD D, and SMHD AD datasets, and the meta-learner neural network was trained using part of the SMHD A-D-AD test dataset, as described in Section 4.2.

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. As an auxiliary partial measure, we estimated the correctness rate for each label (*Correct Prediction per Label*). 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.

**b) Results.** Table 4 summarizes the results. We can observe that the ensemble models outperformed the baseline, both in terms of HL (10 pp lower) and EMR (14 to 16 pp higher). We observe gains in both precision and recall for all disorders.

*Ensemble 1* and *Ensemble 2* models are equivalent regarding HL, but *Ensemble 2* yields a better EMR. The analysis of metrics by class reveals that *Ensemble 2* is superior in identifying users with disorders, with higher recall for anxiety (4 pp), depression (1 pp) and comorbidity (2 pp). Thus, *Ensemble 2* is more appropriate for our purposes.

We analyzed the types of errors performed by *Ensemble 2* in relation to the total number of samples in the test set in terms of (1) errors in distinguishing between healthy and diagnosed users and (2) errors involving only diagnosed users. For the first analysis, the most common error (14% of test users) is related to predicting a diagnosed user as a control one. These errors are distributed as follows: anxiety 4.5%, depression 6.1%, and comorbidity 3.4%. The prediction of a control user as a diagnosed one is less frequent (7.7%). These errors are concentrated in the wrong prediction of a control user as a user diagnosed with comorbidity 6.6% or anxiety 1.1%. For the second analysis, the most frequent error was to miss-classify a user with a single condition as a user presenting comorbidity (13.2% of the test users diagnosed with anxiety only, 11.4% of the test users diagnosed with depression only). Among the users diagnosed with comorbidity, only two prediction errors were observed involving anxiety. Thus, *Ensemble 2* is able to identify most users with comorbidity (high recall), but with limitations in precision. On the other hand, the precision involving the specific disorders is encouraging, as few errors were detected (1.4% of the depressed predicted as anxious).

**Table 5. List of relevant SHAP terms for correctly classified samples**

| Disorder | Model | The 20 most relevant SHAP terms | Relevant SHAP terms found in disorder dictionaries according to word embedding used to each model | | |
|---|---|---|---|---|---|
| | | | Common Terms Anxiety and Depression Dictionaries | Only in Anxiety Dictionary | Only in Depression Dictionary |
| Anxiety | CA1 | though, time, would, told, love, shirts, looked, me, long | my, think, very, wish | really, know, tried, going, crazy, if, me | actually, love |
| | CA2 | probably, going, have, think, really, cool, crazy, know | think, very, wish | really, know, tried, going, crazy, if, me | love, way |
| | CA3 | tell, people, wish | | haha, crazy | either |
| Depression | CD1 | me, cause, motivation, feel, lot, impact, things, ask, man, | because, cause, failed, feel, something, things, think, too | attempts, help, hope, do, you, going, hey, if, so, me | this, anything, always, bad |
| | CD2 | similar, just, favor, going, dont, loose, hope, wrong, | because, cause, failed, feel, something, things, think, too | attempts, help, hope, do, you, going, hey, if, so, me | anything, this, always, even, bad |
| | CD3 | grow, think, discovered | mindset | keep, help | motivation |
| Comorbidity | CA1,2,3 | fiction, leaves, route, my, starting, consider, trying, | ideas, mind, my, night, too, what | trying, really, do, having, so, cold, afraid, weird, to, will | obviously, experience, terrible, much |
| | CD1,2,2 | definitely, titles, suggest, comes, mind, night, history, really, wont, periods, remind, post, ideas | ideas, mind, my, night, too, what | trying, really, do, having, so, cold, afraid, weird, to, will | obviously, experience, much |
| Control | CA1,2,3 | strange, jack, connection, sorry, guys, today, writing, calm, working, word, think, interesting, game, know, | something, think, what | could, down, going, guys, if, instead, me, know, mind, started, strange, thing, why | calm, sorry, found |
| | CD1,2,3 | man, mistake, maybe, confirmed, secret, thing | maybe, something, think | could, down, going, guys, if, instead, know, me, strange, think, why, do | calm, sorry |

## 5.2. Qualitative Assessment

**a) Method.** In addition to the quantitative assessment, it is important to understand the most influential features used by the classifiers to make decisions, and how they are related to the disorders. To this purpose, we calculated the 100 highest SHAP values for a sample of test users using KernelExplainer[8]. The SHAP values were calculated for each weak classifier, but this library does not provide support for custom ensemble models.

Recall that all terms used to identify self-diagnosed users were removed from the *corpora* (Section 2), and thus there are no obvious words among the influential features. Thus, we created a Domain Dictionary (DD) with terms representing the symptoms of each disorder and assessed all SHAP influential features present in the DD. To create the DD, we extracted the most frequent terms used in the DSM-5 manual [American Psychiatric Association 2013] to define the symptoms of each disorder and validated them with the help of two psychologists. Each disorder was then related to 59 terms, with 7 common terms. Then, we expanded these lists by including for each term the 20 closest words in each pre-trained embeddings set (Glove 6B, *All Diagnosed Users*, *Target Diagnosed Users*). We used these close terms according to the respective pre-trained embeddings used to train each weak model.

**b) Results.** Table 5 presents the top-20 SHAP influential features for correctly classified instances of the sample, as well as the influential features that were found in the DD. These features are listed for each weak classifier used in *Ensemble 2*. Thus, anxiety and depression are detailed by each weak classifier, and comorbidity and control users by the union of all weak classifiers. We can see that the terms of the DD could be related to many influential features in all class labels.

Users correctly classified as anxious are related to more DD Anxiety terms. For example, the terms ”crazy” and ”really” are close to ”weird” (Glove6B), ”tiring” and ”upsetting” (*Target 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.

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

Likewise, users correctly classified with Depression are related to Depression terms of the DD, or terms common to both disorders. For instance, ”bad” is close to ”situation”, ”terrible” and ”think”, which could indicate a concern about being negatively evaluated by other individuals; ”motivation” is related to ”energy” and ”willpower”, which could indicate lack of energy and difficulty to perform tasks. Among the common terms, we have “failed”, which is associated with “problem”, “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 DD 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 ”experience”, which are close to terms ”doubts”, ”insecurities”, ”frustrations” and ”urges”, 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: (a) anxiety DD terms, such as ”afraid”, close to ”fear”, ”worry” and ”danger”, feelings strongly present for Anxious Disorders (e.g., Generalized Anxiety Disorder); and (b) depression DD terms such as ”terrible”, which are associated with ”sad” and ”melancholy”, feelings typical of depressed users [American Psychiatric Association 2013].

Users correctly classified as Control are evenly related to terms in all DD lists, although in smaller quantities when compared to samples of diagnosed users. Nevertheless, the models seem to have learned differentiation 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”, ”hostility”.

Although each weak classifier is associated with the respective list of DD terms, we noticed that the terms related to Anxiety are influential features in all classifiers (diagnosed and control users). According [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 weak classifiers.

On the other hand, we noticed that some users, although correctly classified by the respective set of weak classifiers, are miss-classified by *Ensemble 2* as comorbidity, being assigned a second (wrong) disorder label. This explains the high recall and low precision of comorbidity displayed in Table 4. To understand this behavior, we examined how users correctly classified as anxious by the set of $AC_i$ classifiers are handled by the $DC_i$ classifiers, and vice-versa. For instance, for one anxious user also classified as depressed, we identified the term ”next” (close to ”future” and ”prospects”). This could represent both concerns and explanations about future issues, 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]. Conversely, considering a user correctly classified as depressed by the weak classifiers, but also as anxious by *Ensemble 2*, we noticed that the term ”bad” was considered influential by the $AC_i$ classifiers. The concern with being negatively assessed by other individuals is a behavior that can differ between anxiety and major depression disorders. While for depressed individuals this

concern arises from the feeling of considering themselves as bad people or not worthy of being appreciated, for anxious people this concern is based on specific social behaviors or physical symptoms [American Psychiatric Association 2013]. The above analysis reveals that for some manifestations of anxiety and depression disorders, the symptoms can be very similar, but motivated for different reasons. This characteristic requires the identification of more subtle differentiating patterns between these disorders, according to the context in which the symptom is expressed.

Finally, this analysis also revealed that despite the difference in the performance metrics, the change in the kernel initialization function did not suit the purpose to handle the data variability, as we can note that AC1 and AC2, as well as DC1 and DC2 rely basically on the same set of influential features. Thus this behavior is reinforced in the ensemble. On the other hand, the use of a different set of pre-trained embeddings for AC3 and DC3 did result in a complementary set of influential features.

## 6. 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 weak classifiers that distinguish between control and diagnosed users, and the *Level 1* explores these probabilities for a multi-class, multi-label prediction. We performed many experiments to define the weak classifiers, varying in the representations of posting behavior input, LSTM topology and hyperparameters, and word embeddings. In the ensemble, we investigated the influence of comorbidity classifiers to distinguish 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 qualitative assessment revealed strong points of our solution and issues that need to be improved. First, we confirmed that meaningful features do influence the weak classifiers' predictions, which are related to these disorders. Second, it pointed out the importance of varying the pre-trained embeddings, since it results in a broader range of contexts to be considered by the ensemble. Third, we could identify that the kernel initialization function variation strategy, despite the trade-off between recall and precision, did not guarantee variability between the weak classifiers. 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 of identifying more subtle patterns that allow differentiating the presence of anxiety signs, according to their intensity, helping to differ between a specific disorder or their comorbidity.

Future work will focus on identifying patterns of differentiation between anxiety disorders, depression and comorbidity. To this end, we will include weak binary classifiers to distinguish between anxiety and depression. We will also explore alternative architectures to increase the variability in the set solution and for the generation of embedding (e.g. ELMo). We will also experiment with the fine-tuning of advanced language representation models, such as Bidirectional Encoder Representations from Transformers (BERT), which became the state of art for a wide range of natural language processing tasks. Finally, we can evaluate the performance by fine-tuning hyperparameters, possibly using automatic solutions (e.g. Ray framework).

**Acknowledgments:** This research was partially funded by FAPERGS - Brazil (grant 19/2551-0001862-2).

## References


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

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

De Choudhury, M., Sharma, S. S., Logar, T., Eekhout, W., and Nielsen, R. C. (2017). Gender and cross-cultural differences in social media disclosures of mental illness. *CSCW'17*.

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

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

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

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

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

Park, S., Kim, I., Lee, S. W., Yoo, J., Jeong, B., and Cha, M. (2015). Manifestation of depression and loneliness on social networks: A case study of young adults on facebook. *CSCW'15*, pages 557–570.

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

Sharma, E. and De Choudhury, M. (2018). Mental health support and its relationship to linguistic accommodation in online communities. *CHI'18*, pages 1–13.

Tiller, J. (2013). Depression and anxiety. *The Medical journal of Australia*, 199:S28–31.

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H. (2015). Recognizing depression from twitter activity. *CHI'15*, pages 3187–3196.

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

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