Identifying Depression in Tweets Using CNN-deep and BILSTM with Attention Model

Fatima Boumahdi, Amina Madani, Ibrahim Cheurfa
Saad Dahlab University
Algeria
{f_boumahdi@esi.dz}, {a_madani@esi.dz}, {ibrahim.cheurfa@gmail.com}

Hamza Hentabli
UTM University
Malaysia
{hentabli_hamza@yahoo.fr}

ABSTRACT: *Mental health is considered as one of today's world's most prominent plagues. Therefore, our work aims to use the potential of social media platforms to solve one of mental health's biggest issues, which is depression identification. We propose a new deep learning model that we train on a depression-dedicated dataset in order to detect such mental illness from an individual's tweets. Our main contributions lie in the three following points: (1) We trained our own word embeddings using a depression-dedicated dataset. (2) We combined a Convolutional Neural Networks model with the Message-level Sentiment Analysis model in order to improve the feature extraction process and enhance the model's performance. (3) We analyzed through different experiments the performance of three deep learning models in order to provide more perspectives and insights for depression researches. Our model achieved a 99 % accuracy, outperforming any statistical or deep learning models found in literature currently.*

Keywords: Depression Identification, Mental Health, Deep Learning, Sentiment Analysis, Message-level Sentiment Analysis

Received: 12 October 2019, Revised 3 February 2020, Accepted 24 February 2020

DOI: 10.6025/ijwa/2020/12/2/47-61

1. Introduction

Nowadays, people are using social media to express their feelings and share their innermost thoughts and desires, most importantly, all of that is done in a naturalistic way, giving us an opportunity to overcome the manipulation issue addressed in self-reported depression questionnaires. Thus, it allows us to capture these thoughts in their rawest form and use them to identify the publisher's present state of mind, which can be used, using sentiment analysis techniques, to detect clinical depression.

Major Depressive Disorder identification has been the subject of research of many fields, psychiatry, psychology, medicine and even sociolinguistics fields. Depression comes in different degrees and the examinations are usually done through one of the popular questionnaires used by psychologists, such as the Center of Epidemiologic Studies Depression Scale (CES-D) [29], Beck's Depression Scale (BDI) [4] and Zung's Self-Rating Depression Scale (SDS) [46]. But, these examinations lack empirical data as they use the patient's observations or a thirdparty's ones which puts the results under the risk of flawed subjective human testing that can be manipulated easily, often with the purpose of gaining antidepressants or just to hide one's own depression from peers [27].

The aim of this study is to use artificial intelligence's deep learning techniques to overcome one of mental health's most prominent challenges, which is identifying depression automatically from an individual's behavior. We use tweets as a medium to track such behavior due to the thought expression culture of the platform and the availability of the data. Despite the efforts invested in it, mental health still remains one of the most life-threatening health issues in the world. What makes such matter worth more attention is the threats that comes from neglecting it. Research has shown that individuals suffering from one or more mental illnesses will likely experience a snowball effect towards other disorders, leading to life-degrading consequences, and in some cases, to fatal ones.

One of the most popular mental disorders amongst the world's populations is Major Depressive Disorder, commonly known as clinical depression, with nearly 300 million individuals suffering from it globally [1]. Studies have shown that 3 - 5% of males and 8 - 10% of female from the total world's population are likely to experience a major depressive episode within a period of one year [1]. What makes depression the most known disorder in the world, is its likelihood of being triggered by other health issues as it often co-occurs with other illnesses and mental conditions. With that being said, it has been reported that it's one of the major causes of suicide, something that shows why it deserves more attention.

The reason why depression is considered as life threatening, is because of its methods of identification. Diagnosis is extracted from the patient's self-reported experiences, behavior questionnaires, and surveys, which makes it prone to manipulation. Moreover, individuals suffering from depression tend to hide what they are going through and never seek out for help in most of the cases, something that can cause their state to worsen, and sometimes lead to suicide.

In our daily life, our behaviors and decisions are highly influenced by other people's opinions. In Social Psychology, this phenomenon is called social compliance. Whether consciously or not, every choice we make is based upon a previous opinion. Our beliefs and perceptions of reality are shaped according to what the world has influenced us, hence, everything we come across has a direct or an indirect influence on our actions.

With that being said, and with the birth of web 2.0, people began to express their opinions freely, publicly, and in different forms, something that increased the level of interest of organizations and companies in such data because of their precious value. Such insights can be very profitable to businesses and have a great political and economical impact on society. From that, a new research field birthed and became the center of attention of computer science's research community, and it is called sentiment analysis.

One of mental health's most important challenges is identifying depression automatically from an ind vidual's behavior. Once identified and treated, depression has been proven to be cured. Therefore, our main objective with this study is to propose a new way for identifying depression, a way that is based on concrete data and tracked natural behavior. For that, we strive to use one of today's most advanced technologies, which is deep learning.

Since people are using social media more and more to express their feelings and share their innermost thoughts and desires, we take Twitter as a source of data, as it records people's self expressions in their most naturalistic way.

Our main contributions with this work lie in the three following points:

• We trained our own word embeddings using a depression-dedicated dataset.

• We combined a CNN (convolutional neural networks) layer [23] [5] with the MSA(Message-level Sentiment Analysis) model [3] in order to improve the feature extraction process and enhance the model's performance.

• We analyzed through different experiments the performance of three deep learning models in order to provide more perspectives and insights for depression esearches.

This paper is composed of 4 sections:

• The first section is dedicated to depression detection related works that tackled the same or a similar problematic as ours.

• In Section 3 we present our proposed model and we explore each of its layers while explaining its origins and the different

models it is built upon.

• In Section 4, we go through the step-by-step process that we followed to build, train and evaluate our model. By the end we share the results we got and we compare it to other works that used the same dataset.

• Finally, we will end our paper with a conclusion, in which we will summarize the key points of this research and future perspectives for our proposed model and how can it be improved and extended.

2. Related Works

Nowadays, people are using social media to express their feelings and share their innermost thoughts and desires, most importantly. Twitter is one of the richest source of data in terms of quantity, diversity, and rich content (sentiment-wise). It has been proven that such data can be used to study clinical matters, especially when it comes to mental illnesses.

Several researchers focus on detecting mental illness and depression on Twitter. But almost all of them used traditional statistical models as classifiers.[14] propose a crowdsourcing method to build a data on depression from Twitter. They develop an SVM classifier trained on these data to verify if tweets could indicate depression. The classifier is based on language, emotion, style and user engagement. Their method can predict if a tweet is depression-indicative, with a high value of accuracy. They propose a new metric called SMDI (social media depression index) that help to characterize the levels of depression in populations. Their results correlate highly with depression statistics defined by the Centers for Disease Control and Prevention (CDC).

In [11], a dataset was built via automatically derived samples from a large amount of Twitter data. They examine four mental health disorders in particular: depression, bipolar disorder, post traumatic stress disorder and seasonal affective disorder. A statistical classifier was used while building the dataset to differentiate between the users with these disorders. A LIWC (Linguistic Inquiry and Word Count)[37] analysis of each disorder was conducted after that in order to measure the deviations in each illness from a control group. Different experiments were done during the process and taken together. The results indicate that there are diverse sets of quantifiable signals relevant to mental health observable in Twitter.

[38] propose to use features from the activity histories of Twitter users to estimate degree of depression. They also propose a questionnaire that was completed by users who agreed to participate. Several machine learning classifiers were used by SVM to estimate the presence of active depression. Their results bring that longer observation periods (more than two months) do not improve the accuracy of evaluation for depression in progress.

Another study [32] uses machine learning to automatically categorize anxiety patient's internal sentiment and emotions using classifiers based on n-grams syntactic patterns, sentiment lexicon features, and distributed word embeddings. The dataset used was annotated by psychology experts, specifically targeted towards sentiment analysis for mental health. In this work, four-class polarity prediction were used for experiments: positive class, negative class, neutral polarity class (both positive and negative) and a neither positive nor negative sentiment. The results showed that the latter brought better results, while the former fell short.

In [27], a dataset created by [11]. for the Computational Linguistics and Clinical Psychology (CLPsych) 2015 Shared Task [13] was used to study the potential of using Twitter as a tool for measuring and predicting Major Depressive Disorder. They used a Bag of Words approach to quantify each tweet. Several statistical classifiers were used: Decision Trees, Linear Support Vector Classifier, Naive Bayes 1-gram, Naive Bayes 2-gram, Logistic Regression, and Ridge Classifier. The study shows that if prioritization has to be made, recall is more important than precision, because identifying a few false positives is better than strictly identifying the most depressed individuals and missing potentially affected ones. Accuracy was prioritized as well over F1-score because a model which identifies depression well is more important than one which becomes unreliable through a myriad of false positives.

Another study [16] reviews recent studies that aimed to predict mental illness, including but not limited to depression only, using social media. Mentally ill users have been identified using screening surveys, their public sharing of a diagnosis on Twitter, or by their membership in an online forum, and they were distinguishable from control users by patterns in their language and online activity.

The study emphasizes on the potential of using social media for mental illness detection and how it can fill the gaps existing in other sources of data used in the experiments. However, it highlights the major obstacle in such platforms, which is the difficulty to detect the illness in people who are unaware of their mental health status. It addresses as well one of the most important points that pops up in such studies, which is how ethical and legal is it to use public data for such purposes.

In the work of [19], a user-level and tweet-level classifiers were developed and compared to discover which is best to detect at-risk individuals. The data used was collected from the #BellLetsTalk Twitter campaign, which is a wide-reaching initiative that aims to break the silence around mental illness and support mental health across Canada. Both classifiers were tested through different experiments and the results showed that the user-level models perform much better even with a small number of features.

The method of [31] aims to make timely depression detection via harvesting social media data. Benchmark datasets were constructed specifically for online depression detection that are: a well-labeled depression dataset, a non-depression one and a large-scale depressioncandidate one. After that, six depression-related feature groups were extracted, covering not only the clinical depression criteria, but online behaviors as well. A multimodal depressive dictionary learning model was proposed and validated through a series of experiments, which showed an outstanding performance comparing to other related works.

In depression detection from Twitter, studies are focused on the analysis of textual contents of tweets. However, emotions from tweets over time are not very investigated. [6] propose to identify users with or at risk of depression using eight emotions as eatures from tweets over time (Anger, Disgust, Fear, Happiness, Sadness, Surprise, shame and Confusion). They apply a temporal analysis on these emotions to produce a set of temporal features using machine learning classifiers: LR, SVM, NB, DT and RF. Using emotional expressions, their results outperform other models [9], [8] and [12].

In depression detection from Twitter, studies are focused on the analysis of textual contents of tweets. However, emotions from tweets over time are not very investigated. [6] propose to identify users with or at risk of depression using eight emotions as features from tweets over time (Anger, Disgust, Fear, Happiness, Sadness, Surprise, shame and Confusion). They apply a temporal analysis on these emotions to produce a set of temporal features using machine learning classifiers: LR, SVM, NB, DT and RF. Using emotional expressions, their results outperform other models [13], [12] and [8].

The study of [41] analyzed the word usage of messages on Reddit for bipolar and depressive disorder to understand how people shared their feelings and described their symptoms. Semantic network analysis was applied and significant topics related to these mental disorders were identified to understand more details about them.

Many works addressed the deep learning usage in sentiment analysis such as [45], [2], [36], [34], [30], [24], [39] and [33]. Our focus is only on mental health using deep learning. In this area, few works exist. Thus, deep learning techniques were not taken advantage of to solve mental health purposes.

In [20], authors use deep learning to investigate the relationship between computational models and psychological states. A CNN, an LSTM (Long Short-Term Memory)[17], and a GRU (Gated Recurrent Units)[7] were used for experiments and the results showed that even though CNN outperformed the other two models in terms of accuracy. Its output is unreasonable comparing to human's sentiments. Thus, it was concluded that the accuracy of the model can not reflect the psychological state of a person. On the other hand, GRU showed more reasonable results.

In [40] the authors use deep learning to solve the same problematic as ours. The dataset used was generated by scraping tweets of various Twitter pages and labelling them with the aid of a polarity score generated by Textblob's python package. Different deep learning models were experimented with in this study: CNN, RNN (Recurrent Neural Networks) [7] and GRU. Other criteria were taken in consideration to compare the performance. They examined character-based against word-based models and pretrained embeddings against learned embeddings. The results showed that word-based GRU outperformed all the other models. A word-based CNN was considered as one of the most effective models.

For the problem of depression, [25] propose a new design based on deep learning model called DK-LSTM. They propose to incorporate semantic and domain knowledge into LSTM. The authors did not implement their proposed model and then no experiments were presented in the paper.

As a result to our literature review, the final decisions were made about what our study should use and how are we going to continue the rest of the process:

• We will use Twitter as a data source. More specifically the benchmark datasets developed in [31].

• We will use a deep learning model.

• We will focus on binary-class polarity but instead of the common positive-negative model, ours will be depressed, not depressed.

• We will use a tweet-level model instead of a userlevel one because according to our research it is the area which needs more enhancements.

3. Our Approach CNN-MSA Depression

After going through the literature review and seeing some of the works that have been done in this field for similar purposes, we will introduce in this section the architecture of our proposed model.

The architecture for our deep learning model consists of five layers: Preprocessing layer, Word Embeddings layer, CNN layer, Bi-LSTM (Bidirectional LSTM) [3] with Attention layer, and finally a Softmax layer.

Different works have used CNN or LSTM separately for similar research studies but according to our research, none have used these two deep learning methods simultaneously for the same purpose as ours.

Now, we dive deeper into each one of our architecture components by explaining its origins and how has it been adapted for our goal and problematic.

3.1. Preprocessing

As in any social media platform, users tend to express themselves in everyday's slang language, which makes it rare to find well-formed sentences that respect grammatical and linguistic rules.

Furthermore, abbreviations, emojis and smileys are widely used, especially when expressing feelings, opinions or any form of self-expression, in other words, sentiment, which is the subject of our study. These factors have been known to pose a major challenge in NLP (Natural Language Processing) and as much as this field is reaching its most advanced levels. They still are considered as the most important bottlenecks when dealing with raw text. In addition to that, dealing with tweets gives us other factors to take in consideration, such as URLs, hashtags, mentions, and reserved words (RT, FAV).

Even though, some of this data can be useful to the sentiment expressed in the tweet, keeping them requires a very complex model that is able to handle every possibility, changing our focus from depression detection to solving natural language processing issues. Thus, the use of a text preprocessor becomes a necessity rather than an optional step.

Our preprocessing phase is divided in two major steps: cleaning and tokenization.

3.1.1. Cleaning

As a first step, we treat the tweets as one big text corpus. We parse through it and delete all of the following elements: Punctuation.

• URLs.

• Hashtags.

• Mentions.

• Reserved words (RT, FAV).

• Emojis.

• Smileys.

During this phase, we treat all tweets equally without taking in consideration the language of the tweet or any syntactic features of the sentences and words. Tweets are stored then in one text file which will be used as the corpus fed to our neural network.

3.1.2 Tokenization

Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens. These tokens are often loosely referred to as terms or words, but it is sometimes important to make a type/token distinction. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. A type is the class of all tokens containing the same character sequence.

3.2 Embedding Layer

In traditional NLP systems and techniques, words are treated as atomic units, no notion of similarity is present because these words are represented as indices in a vocabulary. [26] explains that using distributed representations of words in a vector space help learning algorithms to achieve better performance in NLP tasks, such as sentiment analysis, by grouping similar words.

The word representations computed using neural networks are very interesting because the learned vectors explicitly encode many linguistic regularities and patterns.Some what surprisingly, many of these patterns can be represented as linear translations.

The Skip-gram model, introduced by [26], is an efficient method for learning high quality vector representations of words from large amounts of unstructured text data. Unlike former neural network architectures used for learning word vectors, training the Skipgram model doesn't involve dense matrix multiplications, which makes it outrageously efficient by giving the possibility of training the model on more than 100 billions words in one day, on an optimized single-machine implementation [26].

$$\frac{1}{t}\sum_{t=1}^{T}\sum_{-c\preceq j\preceq c, j\neq 0} logp(w_{t+j} \mid w_t) \qquad (1)$$

Where c is the size of the training context, and the more we increase it the more training examples we will have and thus a possibility of having a higher accuracy, at the expense of the training time.

3.3 CNN Layer

Famously known in the Computer Vision field, Convolutional Neural Networks (CNNs) provoked major breakthroughs in Image Classification and are the core of most Computer Vision systems today. CNNs are a special type of feedforward neural networks, originally inspired from the human visual cortex. They consist of multiple convolutional layers, each of which performs the function latter's cells [42].

In recent years, CNNs started being used even in Natural Language Processing tasks and the results were surprisingly impressive. By using word vectors to build the input matrix of the model, text was treated in the same way as images, both for feature-extraction and classification, and ever since it became one of the most used neural networks in NLP.

We start by defining Convolution to understand how features get extracted from images and text.

3.3.1 Convolution

Convolution can be thought of as a sliding window function that we apply to a matrix in order to extract features. This window is known as a kernel, filter, or feature detector. The matrix represents an image, and to make it simple we say that it's a black and white image. Each entry in the matrix represents one pixel, 0 for black and 1 for white. We are using a 3x3 filter that multiplies its values element-wise with the original matrix then sums up. We get the full convolution by sliding the filter over the whole matrix and doing the same procedure for each element.

3.3.2 Convolutional Neural Networks Model

Inspired from [10], [21] created the model shown in Figure 1 for sentence classification purposes and it's from this model that we are going to build the convolutional part of our model.

Let $x_i \in R^k$ be the k-dimensional word vector corresponding to the i-th word in the sentence. A sentence of length n (padded

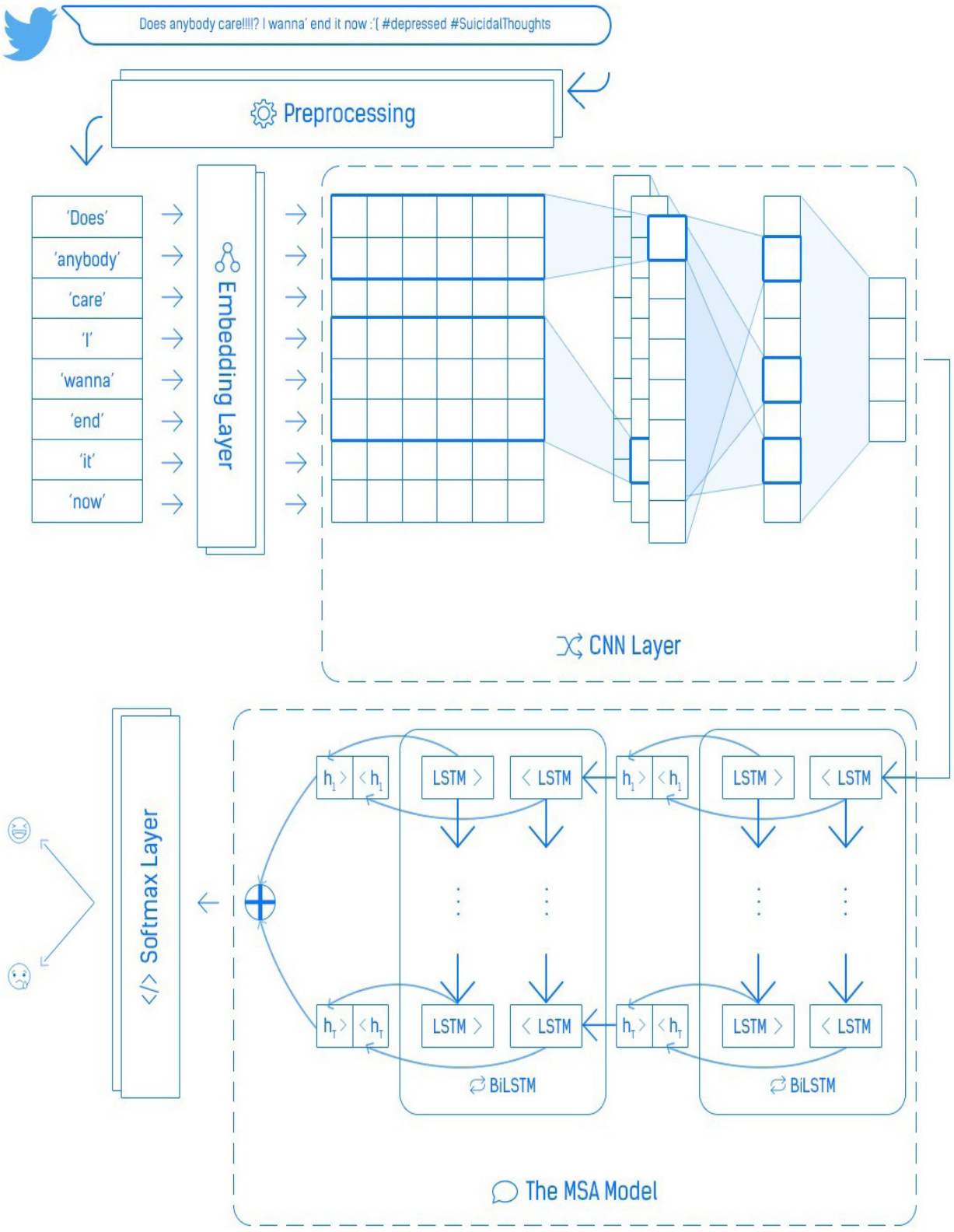

Figure 1. Our depression detection model architecture

where necessary) is represented as:

$$x_{1:n} = x_1 \oplus x_2 \oplus x_3 ... \oplus x_n \quad (2)$$

Where $\oplus$ is the concatenation operator.

In general, let $x_{i:i+j}$ refer to the concatenation of words x_i, x_{i+1}, x_{i+2}, .. , $x_{i:i+j}$.

A convolution operation involves a filter $w \in R^{hk}$, which is applied to a window of words to produce a new feature. For example, a feature c_i is generated from a window of words $x_{i:i+h-1}$ by:

$$c_i = f(w \times x_{i:i+h-1} + b) \quad (3)$$

Here $b \in R$ is a bias term and f is a non-linear function such as the hyperbolic tangent. This filter is applied to each possible window of words in the sentence $x_{1:h}, x_{2:h+1}, \ldots, x_{n-h+1:n}$ to produce a feature map.

$$c = [c_1, c_2, ..., c_{n-h+1}] \quad (4)$$

With $c \in R^{n-h+1}$.

We then apply a maxovertime pooling operation [21] over the feature map and take the maximum value $\Theta = max\{c\}$ as the feature corresponding to this particular filter. The idea is to capture the most important feature one with the highest value for each feature map. This pooling scheme naturally deals with variable sentence lengths.

Here we went through the step-by-step process through which one feature is extracted from one filter, but the CNN model uses multiple filters (with varying window sizes) to obtain multiple features. The layer formed by these features is called penultimate layer and is connected to a final softmax layer that provides the probability distribution over labels [21].

Same as it was described in the previous section, in our model, we are going to start with a tokenized tweet which we then convert to a tweet matrix, the rows of which are word vector representations of each token. These are the outputs of the embedding layer previously defined. According to [9], we can then effectively treat the tweet matrix as an image, and perform convolution on it using linear filters.

In text applications there is inherent sequential structure to the data. Because rows represent discrete symbols (namely words), it is reasonable to use filters with widths equal to the dimensionality of the word vectors (i.e., d). Thus, we can simply vary the height of the filter, i.e., the number of adjacent rows considered jointly. We will refer to the height of the filter as the region size of the filter.

The difference between the standard CNN model and our model is that we don't have a softmax layer but we have instead a BiLSTM with Attention layer through which the features will be inputted, and we are going to cover the process of the latter in the next section.

3.4. BiLSTM with Attention Layer

As mentioned previously, our model is uses the MSA model ([3], [44] and [43]) as a part of it and its sequential layer consists of 2-layer bidirectional LSTM (BiLSTM) with an attention mechanism, to improve the feature extraction process and enhance the model's performance.

We can not start talking about LSTMs before addressing the limitations of Recurrent Neural Networks (RNNs) that triggered the invention of the former in the first place. (RNNs) are a class of neural networks whose connections between neurons form a directed cycle. Comparing it with feed forward neural networks, RNNs are distinguished by having a "memory" that is used for processing sequential information. Memory here is defined by performing the same task for every element of a sequence with

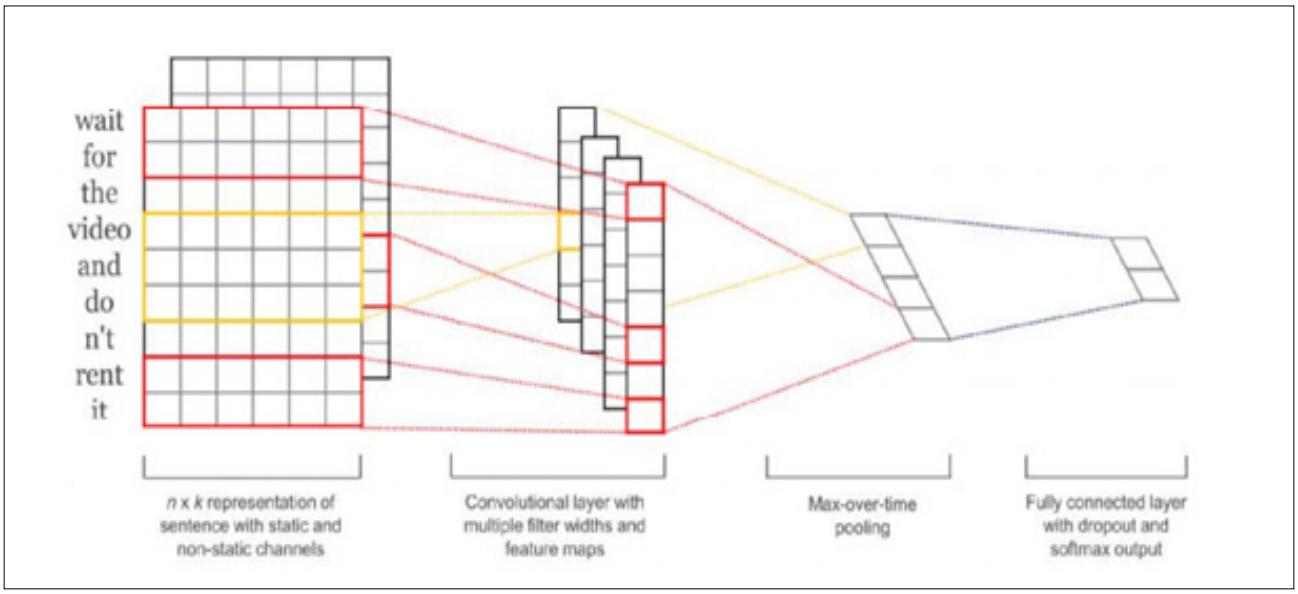

Figure 2. CNN model architecture [10]

each output being dependent on all previous computations. It's remembering information about what has been processed so far, which makes them very efficient and more human brain- like [42].

RNNs success and popularity is due to its ability to connect previous information to the present task, something we called previously as "memory". When the present task only needs recent information, this doesn't cause any obstacles but the older the information needed is the more this "memory" forgets about it. Once the task requires older information, something we call "context", RNNs fail to identify it, in other words, to "remember" it.

To make it clearer, let's use a word prediction model as an example. If we take the sentence "the fish in the" independently and we try to predict the next word after it, it's obvious that it will be "sea". This doesn't need any further context, the gap between the relevant information and where it's needed is small and due to the "memory" of the RNN we are able to use past information "fish" to predict "sea". If we take the same sentence in a paragraph that starts with "I visited AquaDom..." and ends with "... the fish in the", recent information suggests that the next word is probably "sea" but if we take in consideration the context of AquaDom, it's probably "aquarium", and this is where RNNs fail because as the gap grows between the needed information and the current task, the connection between them fades away.

Long Short Term Memory (LSTM) networks on the-other hand are a special kind of RNN that is designed specifically to solve RNNs' long-term dependency problem. LSTMs are being used in a wide variety of problems due to their ability to remember information over long periods of time ([22], [15] and [35]). Similarly to standard RNNs, LSTMs have the same structure, the difference is in the repeating modules, which contain a four-layer neural network instead of a single one.

In general, an LSTM takes the words of a tweet as an input and produces annotations $H = (h_1, h_2, .., h_T)$, where h_i is the hidden state of the LSTM at time-step i, summarizing all the information of the sentence up to x_i. For this case, the use of bidirectional LSTM (BiLSTM) brings the advantage of getting word annotations that summarize the information from both directions, first, going forward from x_1 to x_T, then backward from x_T to x_1. The final annotation of a given word is the concatenation of the annotations from both directions. Finally, the reason why two layers of BiLSTMs are used is to make the model learn more abstract features.

The uniqueness of this BiLSTM model is lies in the attention layer. Knowing that not all words contribute equally to the expressions of a sentiment in a message, the attention mechanism allows the model to find the relative importance of each word

to the expression by assigning a weight ai to each word annotation then computing the fixed representation r of the whole message as the weighted sum of all the word annotations.

3.5 Softmax Layer

In the final layer, the representation *r* is fed to the final fully-connected softmax layer as a feature vector used for classification. The result will be a probability distribution over all classes.

We have gone through the different phases and layers of our proposed model. Most importantly, we dug deep in each part to give a better understanding of the basics of all the models used, this way everything is justified.

In the next section, we're going to describe the process we went through in order to implement this model and the results that we got at the end.

4. Experiments and Results

Now that we have our model designed, it is time to turn theory into practice and evaluate our proposed architecture. To implement the model designed and run all the programs and tests, we used JetBrains PyCharm Community Edition 2018.1.2 development environment.

We share more details about the data used in our experiments and the process followed, and by the end, we present the results we got and we compare them with other results from a similar study.

4.1 Data

Due to the limitations of access and to the narrow category of people that we are targeting with our work, before we do anything we needed to check the available data and build our architecture based upon it instead of proposing an architecture then collecting the data for it.

For that, we used the dataset of [31], which is constructed of three 3 parts (Figure 4): Depression dataset, non-Depression dataset, and Depressioncandidate dataset.

1. Depression Dataset : the first dataset, named D1, is based on tweets between 2009 and 2016 and only contains tweets of people who used the anchor tweet "(I'm/I was/I am/I've been) diagnosed with depression" and that was inspired from the work of [11], where the authors suggested that a tweet is enough to categorizing someone as depressed.With that being said, 1402 depressed users have been obtained making the dataset as 292564 within one month.

2. Non-Depression Dataset : the second dataset, D2, is a kind of the opposite of the first one, as it contains tweets of users that never posted any tweet containing the character string "depress". The tweets that were selected were of December 2016 and it resulted in more than 10 billion tweets of more than 300 millions users.

3. Depression-candidate Dataset D3 : the final dataset, contains people who might potentially have depression, and that is characterized by people whose anchor tweet contained the character string "depress" and they ended up with 36993 depression candidates with over 35 millions tweets. To sum it all up, this dataset is exactly what we have been looking for. Given the variety of sub-datasets and their preciseness, we were able to propose an architecture that is more focused on depression instead of sentiment polarity in general.

4.2 Performance Measures

We need to define the metrics through which we will measure the results of our work and compare it to other works. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is: How many of the users we identified as depressed are actually depressed? High precision relates to the low false positive rate.

$$Precision = \frac{TruePositive}{TruePositive+FalsePositive} \quad (5)$$

Recall is the ratio of correctly predicted positive observations to the all observations in actual class. The question recall answers is: out of all of the depressed users, how many did we properly detect?

$$Recall = \frac{TruePositive}{TruePositive+FalseNegative} \tag{6}$$

F1 Score is the weighted average of Precision and Recall.

$$F1 - Score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{7}$$

Accuracy is the most execution measure and it is essentially a proportion of accurately anticipated perception to the aggregate perceptions. One may imagine that, in the event that we have high precision then our model is ideal. Indeed, exactness is an extraordinary measure yet just when you have symmetric datasets where estimations of false positive and false negatives are relatively same. Thusly, you need to take a gander at different parameters to assess the execution of your model.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

4.3 Experimental Setting

In this section, we will go through the detailed process that we went through to train our model. Since our model includes different types of neural network, we will go through each of them separately.

We experimented with different approaches so that we get rich insights by the end of the study. Since we have distinct datasets that are depression oriented, using a supervised learning approach was the best option we could take. Therefore, instead of using a pre-trained word embedding, we trained our own word embeddings. We also trained a CNN and the MSA model separately using our depression-trained word embeddings to see how efficient they are and what improvements does our model bring.

4.3.1 Word2Vec Neural Network

Although there is a wide variety of pre-trained sentiment analysis word embeddings, we wanted to make our deep neural network more depression oriented. Therefore, we decided to train our own word embeddings using the datasets D1 and D2. As it was mentioned above, D1 contains tweets of depressed people, therefore, it allows us to detect the similarities between the words and expressions they use by generating more relevant word vectors to our study. Same goes for D2, but we did not mix them, we generated the vectors for each dataset separately so that we used them later depending on our needs.

4.3.2 CNN

Even though we met a study that tested a CNN-only model for depression detection, we wanted to experiment with the our own data and word embeddings. Therefore, our first experiment was to train a CNN-only model with the data we got from the previous step to see how efficient CNNs are, when used individually.

4.3.3 MSA Model

Our second experiment was to train the MSA model that we mentioned in the previous section, individually as well, using our depression-trained word embeddings to see if the BiLSTM deep neural network is able to detect depression with high efficiency.

4.3.4 Our model

The most important experiment, and the one all of this study was done for, we used the word embeddings we trained to feed the first part of our model, which is the CNN, and then we took the vectors output from it and fed it to the MSA model.

4.4 Results

Figure (4) shows the results we got from each experiment in addition to the results that were shared from [31], which used the same datasets as us, but with statistical models instead.

As it is shown in the figure 4, we notice that deep learning methods are way more effective than any traditional methods, no matter

how advanced and tailored it is. Most importantly, we see that the deeper the neural network is the better the learning is, the better results we get. Our model achieved a 99%accuracy, outperforming other models.

5. Conclusion

The objective of our study was to solve one of mental health's biggest issues, which is the inefficiency of the traditional identification methods of mental illnesses. More specifically, we chose to focus on major depressive disorder, also known as clinical depression, because of its major popularity and likelihood of spreading, and most importantly, to prevent the tragedies that might occur from such disorder. For that, we wanted to design and propose a new way of identifying depression using an advanced artificial intelligence technology known as deep learning.

In order to achieve that, we reviewed as many studies about the matter as we could. First, we reviewed the different projects that tackled the same problematic as ours. Such research introduced us to different approaches and allowed us to get a clear perspective about where does the current solutions stand. It allowed us as well to notice that technologies such as deep learning weren't taken advantage of for the sake of solving these problems and tackling such life-saving matters.

Our next step then was to propose our own model and use the knowledge we acquired during the literature review to predict the most suitable approach to solve our problematic and achieve our goal. For that, we took advantage of an existing model that was used previously strictly for sentiment analysis, and we tried to include it in our depression detection model with our unique enhancements. Our model mainly consists of two types of deep neural networks: a convolutional neural network and a recurrent one. This combination we believed, in addition to a depression-trained word embeddings layer in entrance, could predict better results than any of the works found in the literature thus far.

Finally, we implemented our model using a humble experimental setup. We used a dataset developed specifically for depression studies, which allowed us to take a supervised learning approach. This helped us conduct different experiments and compare their results in order to discover which technique is the most efficient. We compare our experiments by the end with the results of another study that used the same data, but different methods, more specifically, traditional ones. We proved through that that deep learning is much more efficient than the older statistical (or any other) classification techniques, which makes it more qualified to tackle neglected world problems, such as the one we chose to pursue, not only for its current outstanding results, but for its promising future as well.

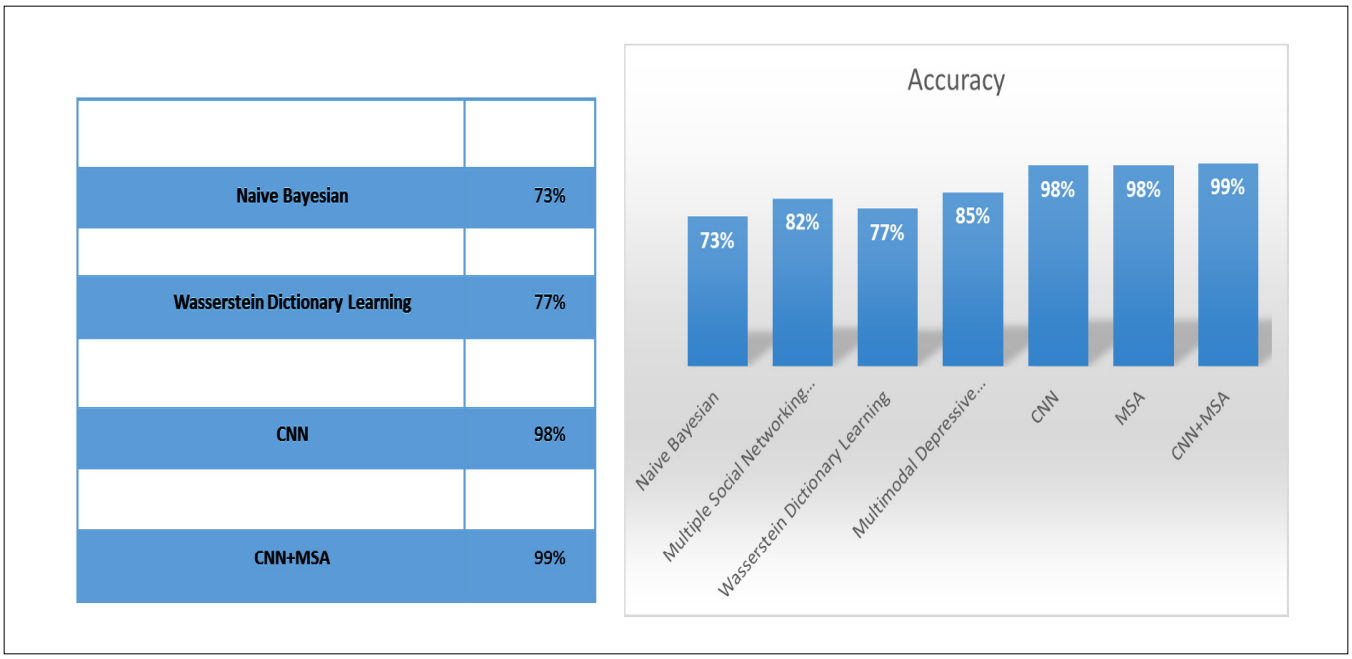

Figure 4. The results of the different experiments conducted on the dataset used

As future perspectives, we see this study as a first step towards unveiling new ways of identifying depression, and other mental illnesses, not only in Twitter or social media, but also in different type of platforms and applications that can be used in day-to-day basis.

First, we suggest that a more complex should be developed from ours, this new one should be able to take in consideration not only the tweets's text but also their timing and other metadata as well. Furthermore, a user-based, instead of a tweet-based model will probably get better results, in terms of diagnosis. Such model should be able to treat more than the tweets of a given user, but all his details as well, this includes: country, profile picture, number of followers, color of the profile, age... etc and all possible details that can be retrieved from a Twitter account.

Finally, our most hopeful and futuristic perspective is for the model to be enhanced to be able to detect the degree of depression an individual is suffering from. If such possibility becomes a reality, then the uses of the model will grow exponentially. One proposition we offer future researchers interested by the subject, is to create a journaling platform for depression patients in order to follow up and report their state of progress to their doctors using that last suggested model.

References

[1] Andrade, Laura., Caraveo-Anduaga, Jorge J., Berglund, Patricia., Bijl, Rob V., De Graaf, Ron., Vollebergh, Wilma., Dragomirecka, Eva Kohn, Robert., Keller, Martin., Kessler, Ronald C. (2003). The epidemiology of major depressive episodes: results from the international consortium of psychiatric epidemiology (icpe) surveys. *International Journal of Methods in Psychiatric Research,* 12(1) 3–21.

[2] Araque, Oscar., Corcuera-Platas, Ignacio., Sanchez-Rada, J Fernando., A Iglesias, Carlos. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77. 236–246.

[3] Christos Baziotis., Nikos Pelekis., Christos Doulkeridis. (2017). Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. *In*: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), p. 747–754, 2017.

[4] Aaron T Beck., Calvin H Ward., Mock Mendelson., Jeremiah Mock., John Erbaugh. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, 4 (6) 561–571, 1961.

[5] Bengio, Yoshua., Aaron Courville., Pascal Vincent. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8) 1798–1828, 2013.

[6] Chen, Xuetong., Sykora, Martin D., Jackson, Thomas W, Elayan, Suzanne. (2018). What about mood swings: Identifying depression on twitter with temporal measures of emotions. *In:* Companion Proceedings of the The Web Conference 2018, WWW '18, p. 1653– 1660. International World Wide Web Conferences Steering Committee, 2018.

[7] Cho, Kyunghyun., Merriënboer, Bart Van., Gulcehre, Caglar., Bahdanau, Dzmitry., Fethi Bougares., Holger Schwenk., Yoshua Bengio. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

[8] Munmun De Choudhury., Michael Gamon., Scott Counts., Eric Horvitz. (2013). Predicting depression via social media. *In*: Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013.

[9] Collobert, Ronan., Weston, Jason. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *In*: Proceedings of the 25th International Conference on Machine Learning, ICML '08, p. 160–167, New York, NY, USA, 2008. ACM.

[10] Ronan Collobert., Jason Weston., Léon Bottou., Michael Karlen., Koray Kavukcuoglu., Pavel Kuksa. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12. 2493–2537.

[11] Coppersmith, Glen., Dredze, Mark., Harman, Craig. (2014). Quantifying mental health signals in twitter. *In* Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, p. 51–60. Association for Computational Linguistics.

[12] Coppersmith, Glen., Dredze, Mark., Harman, Craig., Hollingshead, Kristy. (2015). From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. *In:* Proceedings of the 2nd Workshop on Computational Linguistics

and Clinical Psychology: From Linguistic Signal to Clinical Reality, p. 1–10. *Association for Computational Linguistics*, 2015.

[13] Coppersmith, Glen., Dredze, Mark., Harman, Craig., Hollingshead, Kristy., Mitchell, Margaret. (2015). Clpsych 2015 shared task: Depression and ptsd on twitter. *In* Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, p. 31–39. *Association for Computational Linguistics*.

[14] De Choudhury, Munmun., Counts, Scott., Horvitz, Eric. (2013). Social media as a measurement tool of depression in populations.

In: Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13, p. 47–56. ACM, 2013.

[15] Fischer, Thomas., Krauss, Christopher. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270 (2) 654 – 669.

[16] Guntuku, Sharath Chandra., Yaden, David B., Kern, Margaret., Ungar, Lyle H., Eichstaedt, Johannes C. (2017). *Detecting depression and mental illness on social media: an integrative review*. 18, p. 43–49. Elsevier, 2017.

[17] Hochreiter, Sepp., Schmidhuber, Jürgen. (1997). Long short-term memory. *Neural computation*, 9 (8)1735–1780.

[18] Orabi, Ahmed Husseini., Buddhitha, Prasadith., Orabi, Mahmoud Husseini., Inkpen, Diana. (2018). Deep learning for depression detection of twitter users. *In*: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, p. 88–97. *Association for Computational Linguistics*, 2018.

[19] Zunaira Jamil., Diana Inkpen., Prasadith Buddhitha., Kenton White. (2017). Monitoring tweets for depression to detect atrisk users. *In*: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — *From Linguistic Signal to Clinical Reality*, p. 32–40. *Association for Computational Linguistics*, 2017.

[20] Jo, Hwiyeol., Kim, Soo-Min., Ryu, Jeong. (2017). What we really want to find by sentiment analysis: The relationship between computational models and psychological state. arXiv preprint arXiv:1704.03407, 2017.

[21] Kim, Yoon. (2014). Convolutional neural networks for sentence classification. *In*: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, p. 1746–1751, 2014.

[22] Kumar, Jitendra., Goomer, Rimsha., Singh, Ashutosh Kumar. (2018). Long short term memory recurrent neural network (lstmrnn) based workload forecasting model for cloud datacenters. *Procedia Computer Science*, 125:676 – 682, 2018. The 6th International Conference on Smart Computing and Communications.

[23] LeCun, Yann., Bengio, Yoshua., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10).

[24] Li, Changliang., Xu, Bo., Wu, Gaowei., He, Saike., Tian, Guanhua., Zhou , Yujun. (2015). Parallel recursive deep model for sentiment analysis. *In:* Pacific-Asia Conference on Knowledge Discovery and Data Mining, p. 15–26. Springer, 2015.

[25] Li, Wenwen., Chau, Michael. (2018). Applying deep learning in depression detection. *In*: 22nd Pacific Asia Conference on Information Systems, PACIS 2018, Yokohama, Japan, June 26-30, p. 333, 2018.

[26] Mikolov, Tomas., Sutskever, Ilya., Chen, Kai., Corrado, Greg., Dean, Jeffrey. (2013). Distributed representations of words and phrases and their compositionality. *In:* Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, p. 3111–3119, USA, 2013. Curran Associates Inc.

[27] Nadeem, Moin. (2016). Identifying depression on twitter. arXiv preprint arXiv:1607.07384, 2016.

[28] Plank, Barbara., Søgaard, Anders., Goldberg, Yoav. (2016). Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. *In*: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), p. 412–418. *Association for Computational Linguistics*, 2016.

[29] Sawyer Radloff, Lenore. (1977). The ces-d scale: A self-report depression scale for research in the general population. Applied psychological measurement, 1 (3) 385–401.

[30] Severyn, Aliaksei., Moschitti, Alessandro. (2015). Twitter sentiment analysis with deep convolutional neural networks. *In:* Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 959–962. ACM, 2015.

[31] Shen, Guangyao., Jia, Jiang., Nie, Liqiang., Feng, Fuli., Zhang, Cunjun., Hu, Tianrui., Chua, Tat-Seng., Zhu, Wenwu. (2017).

Depression detection via harvesting social media: A multimodal dictionary learning solution. *In* Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI' 17, p. 3838–3844. AAAI Press.

[32] Shickel, Benjamin., Heesacker, Martin., Benton, Sherry., Ebadi, Ashkan., Nickerso, Paul., Rashidi, Parisa. (2016). *Self-reflective sentiment analysis. In:* Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology, p. 23–32, 2016.

[33] Richard Socher., Alex Perelygin., Jean Wu., Jason Chuang., Christopher D Manning., Andrew Ng., Christopher Potts. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *In*: Proceedings of the 2013 conference on empirical methods in natural language processing, p. 1631–1642, 2013.

[34] Su, Zengcai., Xu, Hua., Zhang, Dongwen., Xu, Yunfeng. (2014). Chinese sentiment classification using a neural network tool—word2vec. *In:* Multisensor Fusion and Information Integration for Intelligent Systems (MFI), 2014 International Conference on, p. 1– 6. IEEE, 2014.

[35] Suhara, Yoshihiko., Xu, Yinzhan. (2017). Alex 'Sandy' Pentland. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. *In:* Proceedings of the 26th International Conference on World Wide Web, 72 WWW '17, p.715–724. International World Wide Web Conferences Steering Committee, 2017.

[36] Tang, Duyu., Wei, Furu., Qin, Bing., Liu, Ting., Zhou, Ming. (2014). Coooolll: A deep learning system for twitter sentiment classification. *In*: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014) p. 208–212.

[37] Yla, R., Tausczik., James., Pennebaker, W. (2010), The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29 (1) 24–5.

[38] Tsugawa, Sho., Kikuchi, Yusuke., Kishino, Fumio., Nakajima, Kosuke., Itoh, Yuichi., Ohsaki, Hiroyuki. (2015). Recognizing depression from twitter activity. *In*: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, p. 3187–3196. ACM, 2015.

[39] Vo, Duy-Tin., Zhang, Yue. (2015). Target-dependent twitter sentiment classification with rich automatic features. *In*: Proceedings of the 24th International Conference on Artificial Intelligence, IJCAI'15, p. 1347–1353. AAAI Press, 2015.

[40] Singh, Diveesh., Wang, Aileen. Detecting depression through tweets.

[41] Yoo, Minjoo., Lee, Sangwon., Ha, Taehyun. (2018). Semantic network analysis for understanding user experiences of bipolar and depressive disorders on reddit. *Information Processing and Management*, Xiaolong Press, 2018.

[42] Lei Zhang., Shuai Wang., Bing Liu. (2018). Deep learning for sentiment analysis: A survey. *In* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. John Wiley and Sons Inc., 1 2018 Xiaolong press.

[43] Zhang, You., Wang, Jin., Zhang, Xuejie. (2018). Ynuhpcc at semeval-2018 task 1: Bilstm with attention based sentiment analysis for affect in tweets. *In:* Proceedings of The 12th International Workshop on Semantic Evaluation, p. 273–278. Association for Computational Linguistics, 2018.

[44] Zhou, Peng., Shi, Wei., Jun Tian., Qi, Zhenyu., Li, Bingchen., Hao , Hongwei., Bo Xu. (2016). Attention-based bidirectional long short-term memory networks for relation classification. *In*: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics Volume 2: Short Papers, p. 207–212. *Association for Computational Linguistics*, 2016.

[45] Zhou, Shusen., Chen, Qingcai., Wang, Xiaolong. (2013). Active deep learning method for semi-supervised sentiment classification. *Neurocomputing*, 120. 536–546.

[46] Zung, William WK., Richards, Carolyn B., Short, Marvin J. (1965). Self-rating depression scale in an outpatient clinic: further validation of the sds. *Archives of General Psychiatry,* 13(6) 508–515.