ISSN: 0974-5823 Vol. 7 No. 1 January, 2022
International Journal of Mechanical Engineering

Detection of depression in text posts on Facebook profiles of Spanish-speaking users

Eduardo Paucarpura Sánchez and Juan Gonzales Suarez
San Marcos National University

Abstract - The main aim of the paper is to investigate and design a web system like help detect the level of depression in Facebook users who communicate in Spanish. For the construction of this web application, natural language processing techniques were used, proceeding to reduce and encode the information, performing a data preprocessing that included a selective analysis and normalization of the Facebook information that makes it accurately described. The random forest algorithm was used in the evaluation original data set, since through the literature review and previous research it was demonstrated to be one of the most effective algorithms for the detection of depression, being able to serve as a basis for specialists in treating this evil.

Index Terms - Depression, machine learning, natural language processing, random forests.

INTRODUCTION

The World Health Organization (WHO, 2020) defines depression as a mental disorder which generates sadness, loss of interest or pleasure, lack of self-esteem, feelings of guilt, sleep or appetite disorders, tiredness and lack of concentration to those who suffer from this change or mental disorder [1]. According to the WHO (2020), the pandemic caused by COVID-19 has caused an increase in mental problems, depression being one of the most significant [2].

The first step to be able to combat depression is an adequate diagnosis in the possible patient, which is made through an evaluation test based on a questionnaire, where the individual must indicate their moods in the last days, the perception of some aspects of his life, among other psychological issues; However, this type of evaluation has certain shortcomings, which could lead to a wrong test result.

One of the main drawbacks with conventional depression screening methods is that it relies on the individual's memory, since the individual is expected to remember their moods of the last days, as well as other details regarding his perception of himself. The second problem with conventional depression screening methods is that they resort to the honesty of the individual, as many of the questions are direct to some degree and the individual may be aware of this and may cause the diagnostic result to be wrong. The third problem with conventional depression detection methods is that the individual can be influenced by their current state of mind when responding with the questions of the psychological evaluation, which can vary the result of the review. Therefore, these observations can cause an individual who does not really have depression to be diagnosed as having it; as well as, that an individual who has depression and is reviewed, a result is obtained, reporting that the person is in good condition.[3]

Various researchers around the world have proposed various solutions that aim to correctly and optimally predict depression using machine learning technique [5] [7] [9]. The proposals developed by these researchers were based on data collected through devices such as mobile phones, public repositories such as national health surveys of a given country and also with data from people on social networks.

The best performing machine learning algorithm used for each of the proposals made by these researchers had several factors that differentiate it, such as the dataset used, the country where each one of them was carried out, these proposals, the amount of data used, among others.

Despite the existence of proposals for the detection of depression in social networks, mainly for Asian countries, no investigations have been found made from data collected in the Spanish language and that is why this proposal proposes the development of a web application that allow the evaluation of publications in text format extracted from Spanish-speaking Facebook users in order to detect whether the user suffers from depression or not, making use of the technical principles of machine learning and natural language processing [4] [8].

PREPARATION OF THE INFORMATION

The initial methodology that is applied for the development of the application for the detection of depression consists of three parts that are summarized below.

A.- Construction of the dataset

For any project based on machine learning, the availability of a dataset with the necessary information is necessary. Datasets can be downloaded from repositories on the Internet or built. In this case, it was decided to build the dataset, since it is necessary with the text of the publications to be used as input data to be in the Spanish language.

For the construction of the dataset, the text of publications of Spanish-speaking Facebook users was extracted using web scraping techniques. The web scraping algorithm used was made in Python, since this language presents tools and libraries that facilitate this type of task. On the other hand, because the present proposal was made using supervised machine learning techniques, that is, the publications of users with depression were labeled as publications that showed signs of depression, a psychologist was hired to

label the publications that had presence of depression. To speed up this task, the posts that present depression were drawn from privated Facebook groups whose members are people who suffer from depression or by people who once suffered from this disorder and want to help those who currently suffer from it.

Figure 1. Web scraping (Source: Own elaboration)

B.- Data pre-processing

Before the design of the depression detection model, the data collected with web scraping was first preprocessed. This task was done in order to have clean data that may be more optimal when training the model and increase the effectiveness of the model's depression prediction. Data preprocessing consists of several sequences, but the main ones will be described.

i.- Data Cleaning: The dataset is made up of text strings; however, there are strings that are not very significant for the analysis, such as URLs and punctuation marks.

This type of strings or characters are completely eliminated from the text. Other types of chains that do not have much value for the extraction of characteristics are articles such as "the", "the", etc. And therefore they are eliminated.

Figure 2 . Data cleaning

ii. Tokenization: After cleaning the data of each publication, the text is tokenized, in this way, the text strings are decomposed and divided into words or tokens. In this way, dealing with these data becomes easier (Figure 3).

Figure 3. Tokenization.

iii. Normalization: The next step is normalization, which is comprised of various tasks that aim to put the text on an equal footing and allow processing to proceed in a uniform manner. For example, make all text uppercase or lowercase, convert numbers to their equivalent to text, etc. For example, after tokenizing we can have the words sadness, Sadness and SADNESS, they are essentially the same word, but our tokenizer recognizes them as different words. For the development of this proposal it is necessary that only the conventional forms appear in our list (for example, sadness, only in lower case) and therefore, it is necessary that the text be normalized.

It can be normalized in two ways, the first is through derivation. Derivation is the process of removing affixes (suffixes, prefixes, infixes, circumfixes) from a word to obtain the root of a word. For instance of Figure 4.

In this way, not only is the number of different words in our text reduced, but in this way we can relate words in our analysis. It may be that, when making the derivation, you cut the root too much and this causes words that really have no relationship to be related and it is for this reason that for this proposal the second way of normalizing was used, the lemmatization.

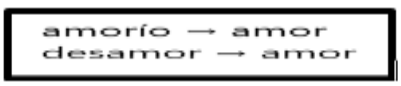

Figure 4. Bypass

Figura 5. Lemmatization

In Spanish, for example, we know that I sing, sing, sing, sing, sing, sing are different forms (conjugations) of the same verb (sing). In order to obviate the differences and put all these variants together in the same term, we use the lemmatization.

Lemmatization is related to derivation, and differs in that lemmatization relates an inflected or derived word to its canonical form or lemma. A motto is nothing more than the form that words have when you look them up in the dictionary.
The stemming takes into consideration the probable word class (adjective, verb, noun, etc.), so we can, if necessary, use this information to filter the generated list of slogans.

DEVELOPMENT OF THE DEPRESSION DETECTION MODEL

Once the dataset had passed the data preprocessing it was used for the construction of the depression detection model. The first thing was to select the appropriate machine learning algorithm for this proposal. After having reviewed different proposals made by other researchers, it was found that the random forest algorithm was the second machine learning algorithm with the best results for the detection of depression in the persons. The reason why the second best algorithm was chosen instead of the first is that the random forest algorithm was used in a greater number of proposals compared to the multilayer perceptron classifier, which was the algorithm with the best results [6] [10].
Before making use of the selected algorithm, natural language processing techniques are used, since it is not possible to work directly with the text of the publications when using machine learning, it is necessary to convert the text of these publications to numbers, since those classification algorithms take numbers as input.
To this process of reduction and coding, where an initial set of raw variables, such as the text of a Facebook post after data pre-processing, is reduced to more manageable characteristics for processing such as numbers; and that the original data set is accurately described.
For this proposal, the extraction of features was done with the TfidfVectorizer function from the scikit-learn Python library. TfidfVectorizer works as follows:
It proceeds to count the number of times each word is used.
A matrix is created, which contains the frequency of each term in each publication.

After the extraction of characteristics, each publication was classified using the random forest algorithm. This algorithm has two stages, one is the creation of the random forest and the other, it is to make a prediction from the random forest classifier that it was created in the first stage.
The algorithm first chooses k characteristics from the total entities m, where k >> m.
Then, between the characteristics k, calculate the node d using the best division point.
Divide the node into child nodes using the best division.
The first two steps are repeated until the number 1 of nodes has been reached.
Build a forest by repeating the previous steps for n number of times to create n number of trees.

Once with the created random forest classifier, the depression prediction is made.

Prediction of depression is done as follows:

The characteristics of the test are taken and based on the rules of each decision tree created at random, the result is predicted and subsequently, the predicted result is saved.
Votes are calculated for each planned goal.
The most voted predicted target is considered as the final prediction of the random forest algorithm.

The output of the random forest classifier algorithm is a prediction of whether depression is present in a user's post or not.

Finally, the dataset is divided into a training set and a test set. The training set is based on the experience that the algorithm uses to learn. While the test set is a set of data used to evaluate the performance of the model using, in this case, the random forest algorithm.

DEVELOPMENT OF THE APPLICATION INTERFACE

Once the depression detection model was built, we proceeded to build the web application for user use.

This web application was planned for the use of two types of users. The first is the standard user, who usually only wants to assess whether a post has signs of depression or not. The second type of user is the user who specializes in mental health, such as a psychologist or psychiatrist, who needs to evaluate a set of publications from a Facebook profile of a specific person.

Figure 7. Architecture of the Web Application

The user who enters the web system will be able to choose whether he wants to evaluate a set of applications or just one application. If you choose to evaluate a single Facebook post, the text of the post will need to be typed or pasted directly into the text box (Figure 8).

Figure 8. Interface to evaluate

On the other hand, if the user wishes to evaluate a set of publications, all the publications must be in a column of a .csv file which can be searched on the computer with the button implemented in this interface.

In both cases, after entering the publications to be evaluated, the "Analyze" button must be pressed, which will show the results of the prediction.

RESULTS

The dataset consisted of a set of 5,000 posts, of which 1,000 of those posts were depressed, while the other 4,000 were standard posts. 80% of the dataset was used for training and the remaining 20% for the test set. The proposed depression detection model obtained a precision of 0.7085; which, although it is not superior to the best proposals made by other researchers in the past, this proposal differs because it evaluates text from Facebook posts in Spanish.

CONCLUSIONS

Depression is a disorder that affects many people worldwide and its correct diagnosis is the first step to make an adequate treatment. In this article, a proposal for the detection of depression in the social network Facebook was presented, evaluating publications of Spanish-speaking users.

The random forest algorithm obtained an acceptable precision, but as future projects it is planned to evaluate the performance of other classification algorithms when detecting depression in Spanish-speaking users through Facebook posts as input data; in order to check if there are other classification algorithms that have a better result for this case under review.

In addition, it is necessary to automate various tasks of the web application to improve the user experience and facilitate their work.

REFERENCES

1. Organización Mundial de la salud. (OMS, 2020). Trastornos mentales: Depresión. https://www.who.int/mental_health/management/depression/es/
2. El Comercio. (2020). Depresión: uno de los efectos secundarios de la COVID-19 al que hay que poner especial atención. https://elcomercio.pe/viu/estar-bien/depresion-uno-de-los-efectos-secundarios-de-la-covid-19-al-que-hay-que-poner-especial-atencion-salud-mental-depresion-coronavirus-pandemia-bienestar-noticia/?ref=ecr
3. Tadesse, M., Hongfei, L., Bo Xu y Yang, L. (2019). Detection of Depression-Related Posts in Reddit Social Media Forum. IEEE Access, 7, 44883-44893. 10.1109/ACCESS.2019.2909180

4. Yan D., Xuemei C., Qiming F. y Shan Z. (2020). A Depression Recognition Method for College Students Using Deep Integrated Support Vector Algorithm. IEEE Access, 8, 75616-75629. 10.1109/ACCESS.2020.2987523
5. Yanga X., McEwenb R., Robee Ongc L. y Zihayatc M. (2020). A big data analytics framework for detecting user-level depression from social networks. Elsevier, 54. https://doi.org/10.1016/j.ijinfomgt.2020.102141
6. Baek J. y Chung K. (2020). Context Deep Neural Network Model for Predicting Depression Risk Using Multiple Regression. IEEE Access, 8, 18171-18181. 10.1109/ACCESS.2020.2968393
7. Yueh Chiu C., Yuan Lane H., Ling Koh J. y Arbee L. P. (2021). Multimodal depression detection on Instagram considering time interval of posts. Springer, 56, 25-47. https://doi.org/10.1007/s10844-020-00599-5
8. Saleh Alghamdi N., Hosni Mahmoud H., Ajith A., Awadh Alanazi S. y García Hernández L. (2020). Predicting Depression Symptoms in an Arabic Psychological Forum. IEEE Access, 8, 57317-57334. 10.1109/ACCESS.2020.2981834
9. Burdisso S., Errecalde M., Montes M. y Gómez. (2019). A text classification framework for simple and effective early depression detection over social media streams. Elseiver, 133, 182-197. https://doi.org/10.1016/j.eswa.2019.05.023
10. Kumar P., Garg S. y Garg A. (2020). Assessment of Anxiety, Depression and Stress using Machine Learning Models. Elseiver, 171, 1989-1998. https://doi.org/10.1016/j.procs.2020.04.213