Journal of Namibian Studies, 32 (2022): 267-275 ISSN: 2197-5523 (online)

Application Of Text Mining Techniques For Pattern Recognition In Distress Analysis Interview Corpus DAIC-WOZ

Sonia Jaramillo-Valbuena[1] , Cristian-Giovanny Sánchez-Pineda[2] , Sergio-Augusto Cardona-Torres[3]

[1] PhD in Engineering. Computer Engineering Dept., Universidad del Quindío, UQ. Armenia, (Colombia)
[2] Master's student in engineering. Universidad del Quindío, UQ. Armenia, (Colombia),
[3] PhD in Engineering. Computer Engineering Dept., Universidad del Quindío, UQ. Armenia, (Colombia)

Abstract

The WHO defines depression as a frequent mental disorder, in which the individual experiences guilt, loss of self-esteem, sleep problems, poor concentration, in consolation, permanent melancholy, lack of interest, changes in appetite and fatigue (WHO, 2022).

In this paper, we generate predictive models by making use of data mining and machine learning techniques.

The data used for this research corresponds to DAIC-WOZ (University of Southern California, 2019), a real world data set provided by the University of Southern California, which has clinical interviews in different formats: audio, video and questionnaire responses. We use different vectorization techniques (TF/IDF and BERT Vectorizer) and apply different supervised learning techniques and Deep learning, namely: BERT, Decision Tree, Logistic Regression for global features, and also the combined techniques, BERT/Logistic Regression and Decision Tree/Logistic Regression. We use accuracy metric to assess the quality of the models obtained.

We identify that the BERT approach has a good performance over Logistic Regression model. Deep learning opens the doors to work with new deep learning and PLN techniques to analyze structured and unstructured information.

Introduction

The WHO defines depression as a frequent mental disorder, in which the individual experiences guilt, loss of self-esteem, sleep problems, poor concentration, inconsolation, permanent melancholy, disinterest, changes in appetite and tiredness, depression can come to generate difficulties in relationships at work, school and family level, affect the ability to cope daily living and aggravating pre-existing medical conditions (WHO, 2022). Likewise, it is considered as the disorder that generates the greatest disability worldwide, far exceeding that generated by physical problems (Wagner, 2012).

Depression is the consequence of interactions complex relationships between different elements at a psychological, social and biological level, and the individuals who go through adverse situations are more likely to suffer from it. According to (Sartorato, 2018) the psychological pressure on people to obtain greater productivity and addiction to technology, are factors that affect the presence of this disorder. Depressive disorder can be mild or complicated to become chronic. The Depression affects human beings regardless of their age or social status. It is considered the second cause of death in the world, in the population between 15 and 29 years.

In this paper we apply different supervised learning techniques and Deep learning to generate predictive models and get depression patterns in an audio and text dataset. We use the Crisp-DM methodology.

The paper is organized as follows. Section 2 presents related works. Section 3 describes the corpus and techniques we use. Section 4 shows the experiments carried out. The last section we show a summary of the findings in this work and lines of future work.

II. RELATED WORK

There are several approaches for Automatic Depression Detection. The approach of (Sau & Bhakta, 2017) makes use of artificial neural networks (Multilayer Perceptron) to predict depression from the geriatric population. They get dataset by interviewing the 105 selected elderly people in India. For model training, they analyze sociodemographic variables, morbidity conditions, and sleep problems. To assess the quality of the model, they use accuracy index.

In (Burdisso, Errecalde, & y-Gómez, 2020) the authors present SS3, a framework designed for the detection of early traces of ERD risks, which uses incremental learning, vector generation, and the use of policies to select higher values and give the classification of texts of social networks, with the objective of a early detection of indications of depression. They use (CLEF, 2017) as dataset. For training, the system uses in total of 295023 submissions, respectively 30,851 (depression) and 264,172 (control).

In (Li, Yang, Li, Chen, & Du, November 2020) the authors carry out a study for the recognition of mild depression using electroencephalography (EEG). Through graph theory they explore the abnormal organization in the Network of functional connectivity and generate functional connectivity matrices of five EEG bands (delta, theta, alpha, beta and gamma). They then use convolutional neural networks to recognize mild depression over the matrices. The results show that the functional network of the group of mild depression shows a larger characteristic path length and coefficient of clustering lower than the healthy control group.

Geraci et al. (Geraci, et al., 2017) apply deep neural networks to texts from 861 medical records to phenotype depression in youth. The process followed considers: the use of encryption to eliminate the personal identification, the addition of 2 psychiatrists to label the EMR documents, the use of a brute-force search and finally, the training of a deep neural network. They implement the system in R, and use a dataset contains 861 documents, labeled in Adequate (depressed) and Not-Adequate (control). They uses sensitivity and specificity to assess the models, and Cross validation (Hea & Cao). They perform an automated analysis of depression using convolutional neural networks speech. The dataset used is AVEC2013 which has 150 videos of 82 people.

In (Yang, y otros, 2017) they present a multi modal depression analysis in order to target the Depression Sub-Challenge (DSC) task of the AVEC2017. In here the data contains audio, video and text, and the proposed system target a hybrid depression classification framework. For Audio/Video model, they implemented a unimodal DCNN-DNN model as depression recognition where each segment of audio and video pass through the DCNN, followed by one ReLU, Pooling and Dropout layers, and other two connected layers at the end. The second final fully connected layer will be the input of the

DNN. Now for text-based classification, they created some global features among the interviews transcriptions, i.e., number of filler words, number of laughs, number of sighs, among others. This new features were model using a RandomForest Classifier. For each interviews answers transcriptions, they extract the Paragraph Vector descriptors, which fed a layer of Support Vector Machines, which is used as an input for a Random Forest Classifier. Each of this outputs are multiplied using the AND operator to get the final result. They performed automatic depression analysis and created a multi-modal framework that can be used with different source of data.

Finally, (Orabi & Orabi, 2018) use the CNNWithMax, MultiChannelPoolingCNN, MultiChannelCNN and BiLSTM techniques to recognize depression in Twitter users. They use a data set containing the age, gender, and text of 1,145 Twitter users whose messages were tagged Control, Depressed, and PTSD (Coppersmith et al., 2015b). They use Accuracy, F1, AUC, Precision and Recall to assess the quality of the models.

III. DATASET AND METHODS

Dataset and implementation details

In this study, we use the data from DAIC-WOZ Database (University of Southern California, 2019). It is a corpus contains one hundred eighty nine clinical records of patients and its is used to help the diagnostic of Psychological distress conditions. It contains audio, text and transcriptions. The dataset is partitioned into 3 data sets, namely training, development and testing sets. Training set has 107 instances. The second dataset, contains 35 instances and the last one is 47. Each record is associated with its corresponding class label (non-depressed or depressed).

Ellie conducts interviews, each of them lasts between 12 and 20 minutes. Ellie es a virtual pollster. The corpus contains levels of depression according to PHQ-8 (Wu, et al., 2021). The values to take range from 0 to 24.

Regarding data scrubbing, we delete the Ellie's sentences. Also, remove stop words, and then, when the PHQ-8 score is greater than 10, we tag the record as 1 (depressed). During experimentation, we apply different vectorizers, namely: TF-IDF

and BERT (Scikit-learn development, 2022). We also extract some additional global text features for each interview after the cleaning process. These include the number of stop words, laughters, the number of words, and the number of filler words.

Interview classification with BERT

BERT is an ANNs-based technique that builds contextual representations. BERT learns relationships between words and identify their context. (Jacob Devlin and Ming-Wei Chang, Research Scientists, Google AI Language, 2018). We present in Figure 1 BERT's neural network Architecture. E1 is the vector representation of a word, T1 is the final output and the middle representations of the same word are called Trm. Diverse middle representations of a word have the same size (Jaramillo, Sánchez, & Cardona, 2022) (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014).

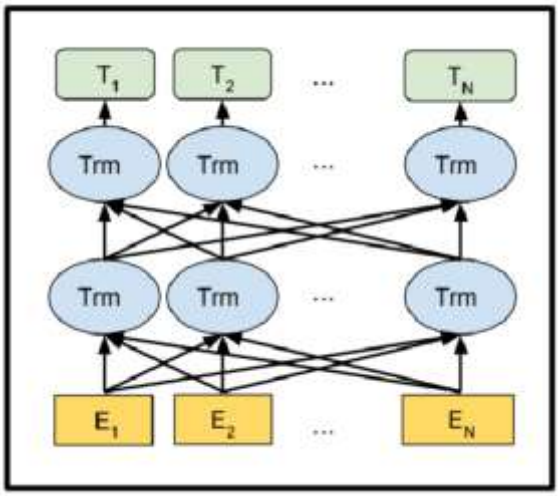

Figure 1. BERT's Neural Network Architecture, Source: (Jacob Devlin and Ming-Wei Chang, Research Scientists, Google AI Language, 2018) (Devlin, Chang, Lee, & Toutanova, 2018)

We use BERT tokenizer and set batch size =16, convolutional_neural_network layers =3 with the kernel of two, three and four, respectively. The first densely connected neural network is powered by the output obtained by concatenating the 3 layers of the convolutional neural network. We use a second densely connected neural network to predict the class label. The hyper parameters are NB_FILTERS = 100, FFN_UNITS = 256,

NB_CLASSES = 2, DROPUT_RATE = 0.2 and NB_EPOCHS = 5, EMB_DIM = 200.

Interview classification with Logistic Regression and Decision Tree

Logistic Regression is a statistical model which estimates de probability of occurrence based on a dataset of independent variables. This algorithm is most recommended for binary classification as the output is bounded between 0 and 1 (IBM, 2023). In this work we implement a Logistic Regression from Sklearn for the global text features, using the next hyperparameters, fit_intercepts=True, penalty=l1 and finally a solver=liblinear.

The Decision Tree model is supervised learning technique for classifying and also regression models (Scikit-learn development, 2022). For the interviews, we implemented a Decision Tree using a TF/IDF vectorizer, both models were used to get the final output by multiplying using the AND operator since both outputs will be binary.

Finally, we did use the Logistic Regression Model in order to use the global features with the BERT tokenizer model. The same behaviour is used to compare both techniques.

IV. RESULTS OF EXPERIMENTATION

In our evaluation, we get the quality of predictive models. For this, we use the accuracy metric. Table 1 shows a comparative analysis of BERT, Decision Tree, Logistic Regression for global features, and also the combined techniques, BERT/Logistic Regression and Decision Tree/Logistic Regression, for Training and test and full datasets. The result of the experiment shows that BERT used as a tokenizer shows better feature extractions that models such as TF-IDF combined. The model that used Decision Tree/Logistic Regression, even though reach a 77% of accuracy, BERT/Logistic Regression hits almost 80% using the full dataset.

	BERT	**Decision Tree (Interviews)**	**Logistic Regression (Global text features)**	**BERT/Logistic Regression**	**Decision Tree/ Logistic Regression**
Training set	0.7500	1.00	NA	NA	NA
Test set	0.8929	0.7300	NA	NA	NA

Cross validation	NA	0.6500	0.6300	0.7700	0.7692

Table 1. Model Accuracy

I. CONCLUSIONS

In this paper we apply different supervised learning techniques and Deep learning, namely: BERT, Decision Tree, Logistic Regression for global features, and also the combined techniques, BERT/Logistic Regression and Decision Tree/Logistic Regression to identify depression patterns in Distress Analysis Interview Corpus DAIC-WOZ.

For the analysis, we apply data scrubbing techniques. The results of the experimentation show that Bert outperforms to the other techniques. BERT lets to get contextual representations of words and considers directionality.

As future work, we propose apply LDA to get topics from DAIC-WOZ database.

REFERENCES

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., . . . Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. ECML PKDD Workshop: Languages for Data Mining and Machine Learning, (págs. 108–122).

Burdisso, S. G., Errecalde, M., & y-Gómez, M. M. (2020). τ-SS3: A text classifier with dynamic n-grams for early risk detection over text streams. Pattern Recognition Letters, 138, 130-137. doi:https://doi.org/10.1016/j.patrec.2020.07.001

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805. Obtenido de http://arxiv.org/abs/1810.04805

Geraci, J., Wilansky, P., De-Luca, V., Roy, A., Kennedy, J., & Strauss, J. (2017). Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. Evid-Based Ment Health, 20(3):83–7.

IBM. (2023). What is logistic regression?Learn how logistic regression can help make predictions to enhance decision-making. IBM. Obtenido de https://www.ibm.com/topics/logistic-regression#:

:text=Resources-,What is logistic regression?,given dataset of independent variables.

Jacob Devlin and Ming-Wei Chang, Research Scientists, Google AI Language. (2018). Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. Recuperado el 11 de 12 de 2022, de https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Jaramillo, S., Sánchez, C., & Cardona, S. (2022). Benchmarking Different Classification Techniques To Identify Depression Patterns In An Audio And Text Dataset. Webology.

Li, Y., Yang, H., Li, J., Chen, D., & Du, M. (November 2020). EEG-based intention recognition with deep recurrent-convolution neural network: Performance and channel selection by Grad-CAM. Elsevier B.V., Volume 415Pages 225-233.

Orabi, & Orabi. (2018). Deep Learning for Depression Detection of Twitter Users. Proceedings of the Fifth Workshop on Computational Linguistics and Clinical .

Sartorato, V. (2018). Depresión. Recuperado el 2022, de https://www.notimerica.com/sociedad/noticia-brasil-pais-mas-depresion-iberoamerica-20170224163216.html

Sau, A., & Bhakta, I. (2017). Artificial Neural Network (ANN) Model to Predict Depression among Geriatric Population at a Slum in Kolkata, India. J Clin Diagn Res, v.11(5).

Scikit-learn development. (2022). scikit-learn. Recuperado el 12 de 12 de 2022, de https://scikit-learn.org/stable/

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15, 1929–1958. Obtenido de http://jmlr.org/papers/v15/srivastava14a.html

University of Southern California . (s.f.). Extended DAIC Database. (2019) Recuperado el 27 de 11 de 2022, de https://dcapswoz.ict.usc.edu/

University of Southern California. (2019). Extended DAIC Database. (2019) Recuperado el 27 de 11 de 2022, de https://dcapswoz.ict.usc.edu/

Wagner, F. G.-F.-G.-F. (2012). Enfocando la depresión como problema de salud pública en México. Salud Mental, 35, 3-11.

WHO. (2022). Depression. Recuperado el 27 de 11 de 2022, de https://www.who.int/news-room/fact-sheets/detail/depression

Wu, Y., Levis, B., Riehm, K., Saadat, N., Levis, A., Azar, M., . . . Ayalon, L. (2021). Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: A systematic review and individual participant data meta-analysis. Psychol Med, 50(8), 1368–1380.

Yang, L., Sahli, H., Xia, X., Pei, E., Oveneke, M. C., & Jiang, D. (2017). Hybrid Depression Classification and Estimation from Audio

Video and Text Information. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (págs. 45–51). New York, NY, USA: Association for Computing Machinery. doi:10.1145/3133944.3133950

Yao, L., Mao, C., & Luo, Y. (2019). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC Med Inform Decis Mak 19 , (Suppl 3), 71.