DLRG@LT-EDI-ACL2022:Detecting signs of Depression from Social Media using XGBoost Method

Herbert Goldwin Sharen, Ratnavel Rajalakshmi*
School of Computer Science and Engineering
Vellore Institue of Technology
Chennai
rajalakshmi.r@vit.ac.in

Abstract

Depression is linked to the development of dementia. Cognitive functions such as thinking and remembering generally deteriorate in dementia patients. Social media usage has been increased among the people in recent days. The technology advancements help the community to express their views publicly. Analysing the signs of depression from texts has become an important area of research now, as it helps to identify this kind of mental disorders among the people from their social media posts. As part of the shared task on detecting signs of depression from social media text, a dataset has been provided by the organizers (Sampath et al.). We applied different machine learning techniques such as Support Vector Machine, Random Forest and XGBoost classifier to classify the signs of depression. Experimental results revealed that, the XGBoost model outperformed other models with the highest classification accuracy of 0.61% and an Macro F_1 score of 0.54.

1 Introduction

Depression is a risk factor for Dementia. Dementia patients often experience a deterioration in cognitive abilities such as thinking and remembering (Dong and Yang, 2021). Early detection and treatment of depressive symptoms can greatly improve the chances of controlling depression and reducing the harmful effects of depression on a person's well-being, health, and social–economic life. The task of distinguishing between depressed and non-depressed people using online social media is critical. Information, communication, and posts on social media describe a user's emotional state (Al-adb et al., 2018). Their sentimental state, on the other hand, will be strong, which could lead to a misdiagnosis of depression. Clinical interviews and questionnaire surveys conducted by hospitals or organizations, where psychiatric assessment tables are used to determine mental disorder prognosis, are currently the most common procedures used. Depression affects more than 300 million people worldwide, according to the World Health Organization. Depression can have a negative impact on one's personal well-being, family life, and educational institutions at work, as well as contribute to physical damage.

Recently, the task of detecting depression in an earlier stage is attempted by researchers in alternative ways too. One such attempt is to mine the social media posts of people, from which the signs of depression can be detected. To this end, various machine learning techniques could be applied to diagnose depression from feelings or emotions expressed in social media texts, by using Artificial Intelligence (AI) and Natural Language Processing (NLP)-based approaches. We have extracted features from the text and BOW method is appliled. For building the model, we have used SVM and ensmeble based methods Random Forest classifier and XGB classifier. From the experimental results, we found that, XGB outperforms other methods with an accuarcy of 0.60 and an F_1 score of 0.54.

2 Literature Review

Ibitoye A.O (Ibitoye et al., 2021) looked at two research that looked at how supervised machine-learning classifiers could predict the interaction of emotions. They used classification methods to classify depression-related messages on social media. Mathur (Mathur et al., 2020) suggested a strategy based on a Bidirectional LSTM (BLSTM) + Attention model for detecting depression early in Twitter users' messages. To detect depression from the Twitter dataset, Orabi (Husseini Orabi et al., 2018) suggested using the Continuous Bag of Words (CBOW) embedding approach. Kim Jin et al. (Kim et al., 2021) researched about how a supervised machine learning algorithm can help detect post-traumatic stress disorder by measuring predicting parameters.

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 346 - 349
May 27, 2022

3 Methodology

The overall workflow of the proposed system is depicted in Figure 1 The flow depicts data preprocessing, lemmatization followed by training and testing using machine learning model. Support Vector Machine, Random Forest, and XGBoost are utilized to classify depression from text data into no depression, severe and moderate depression.

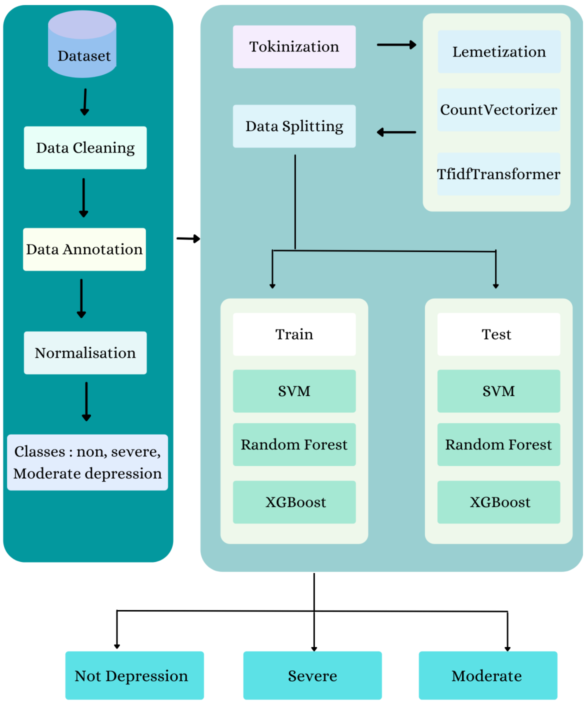

Figure 1: Overall workflow of the proposed system

3.1 Data Processing

Applying pre-processing is essential for classification in the experiment since the data are not organised in the same format or structure. To minimise redundancies and make sure the data is computer usable, data cleaning and transformation were utilised. The data types are justified so that the dataset can be compared and compared. The scale condition in data had to be uniformed, hence normalisation was also required. The psychological domain knowledge in functional diagnostic criteria is employed in the rebuilding of data structure while data preprocessing is being implemented. All data must be reconstructed into only three labels, which correspond to three types of depression diagnostic criteria: not depression, severe depression, and mild depression.

3.2 Lemmatization

In Natural Language Processing (NLP) and machine learning in general, lemmatization is one of the most used text pre-processing techniques. A given word is reduced to its base word in both stemming and lemmatization. In the lemmatization process, the root word is called lemma. As a result, a lemmatization algorithm would recognise that better is derived from good, and hence the lemmatizer is good. Because lemmatization entails determining a word's meaning from a source such as a dictionary, it takes a long time. As a result, most lemmatization methods are slower than stemming techniques. Although there is a processing expense for lemmatization, computational resources are rarely a consideration in an ML challenge. WordNet Lemmatizer is used where NLTK is used to convert Special character, dot remove, multiple space converted to single space, conversion from upper case to lower case.

3.3 TFIDF transform

The TF-IDF is a subtask of information retrieval and extraction that seeks to represent the value of a word in a document that is part of a corpus of documents. Some search engines utilise it to assist them get better results that are more relevant to a particular query. TF-IDF stands for Term Frequency — Inverse Text Frequency and is a statistic that attempts to better describe how essential a term is for a document while also considering its relationship to other papers in the same corpus. This is done by counting the number of times a term appears in a document as well as the number of times the same word appears in other documents in the corpus

3.4 Support Vector Machine (SVM)

Support Vector Machines are typically thought of as a classification method, however they can be used to solve both classification and regression problems. It can handle both continuous and categorical variables with ease. To differentiate various classes, SVM creates a hyperplane in multidimensional space. SVM iteratively generates the best hyperplane, which is then utilised to minimise an error. The goal of SVM is to find a maximum marginal hyperplane (MMH) that splits a dataset into classes as evenly as possible. A hyperplane is a decision plane that divides a group of items that belong to various classes. A margin is the distance between the two lines on the class points that are closest to each other.

3.5 Random Forest (RF)

Random Forest is a classifier that combines a number of decision trees on different subsets of a dataset and averages the results to increase the dataset's predicted accuracy. Instead than relying on a single decision tree, the random forest collects the forecasts from each tree and predicts the final output based on the majority votes of predictions. The bigger the number of trees in the forest, the more accurate it is and the problem of overfitting is avoided. The random forest is formed in two phases: the first is to combine N decision trees to build the random forest, and the second is to make predictions for each tree created in the first phase.

3.6 XGBoost Classifier

The eXtreme Gradient Boosting (XGBoost) technique is a more advanced version of the gradient boosting algorithm. The eXtreme Gradient Boosting (XGBoost) technique is a more advanced version of the gradient boosting algorithm. XGBoost is a sophisticated machine learning algorithm that excels in terms of speed and accuracy. While implementing an XGBoost model, we must take into account many parameters and their values. To increase and completely use the advantages of the XGBoost model over competing methods, parameter adjustment is required.

4 Experimental Study and Results Discussion

The implementation work for the depression detection challenge is described in this section. The dataset and data acquisition procedure are explained in Section 4.1. The division of data for training and testing purposes is briefly explained in Section 4.2. Section 4.3 describes the results attained from the models.

4.1 Dataset

The CodaLa dataset contains social media postings in English, the system is required classify the signs of depression into three labels namely "not depressed", "moderately depressed", and "severely depressed". The dataset collected from codaLabs consists of 8891 texts which were also included labels. Figure 2 shows the sample data set with class moderate, severe and no depression.

PID	Text_data	Label
train_pid_1	Waiting for my mind to have a breakdown once the "New Year" feelin	moderate
train_pid_626	Goodbye : [removed]	not depression
train_pid_889	With each passing day my depression is getting worse and worse, I d	severe

Figure 2: Sample Dataset with labels and Text Value

4.2 Data Splitting For training and testing

The complete dataset is split into two sets: a training set and a test set, which are used to train and evaluate the model. This method can also be used to assess the overall performance of the model during training and validation. The shape of training set was (8891,1500) and the shape of testing set was (3245,1500).

4.3 Discussion

XGBoost outperforms the other methods used with an accuracy of 64.3%, F1 score 0.54 , recall of 0.64 and precision of 0.52. Random forest attained an accuracy of 56.4%, F1 score 0.55 , recall of 0.56 and precision of 0.56 which performs less when compared with XGBoost with respect to accuracy. SVM which was also implemented attained an accuracy of 56.7%, F1 score 0.55 , recall of 0.56 and precision of 0.55. Based on the results obtained the depression prediction of the three class data of SVM, Random Forest and XGBoost is shown in Table 1. Learning accuracy like Recall, Precision, F1-Score of the three class classification using Random Forest, SVM and XGBoost is given in table 1.

5 Conclusion

Early detection and treatment of depressive symptom improves a person's chances of controlling depression and reducing its harmful effects on their well-being, health, and social–economic life. The dataset comprises a text data set for categorising depression into three categories: moderate, severe, and not depressed. Machine learning algorithms are used to classify the text data for identifying the depression data. Using SVM and Random Forest classifiers resulted in an accuracy of 0.55 and 0.43. The highest classification accuracy of 0.60 was achieved XGBoost classifier.

References

Ahmet Emre Aladb, Serra Muderrisoglu, Naz Berfu Akbas, Oguzhan Zahmacioglu, and H. Bingol. 2018. Detecting suicidal ideation on forums: Proof-of-

	Accuracy	Macro F1-score	Macro Recall	Macro Precision	Weighted F1-score	Weighted Recall	Weighted Precision
XGBoost	0.6434	0.29954	0.337	0.3664	0.5457	0.643	0.525
Random Forest	0.564	0.36	0.37	0.40717	0.556	0.564	0.562
SVM	0.567	0.357	0.358	0.4136	0.5516	0.567	0.5528

Table 1: Learning accuracy of the three class classification using Random Forest, SVM and XGBoost

concept study. *Journal of Medical Internet Research*, 20.

Yizhuo Dong and Xinyu Yang. 2021. A hierarchical depression detection model based on vocal and emotional cues. *Neurocomputing*, 441:279–290.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 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*, pages 88–97, New Orleans, LA. Association for Computational Linguistics.

Ayodeji Ibitoye, Ranti Famutimi, Odunayo Dauda, and Akioyamen Ehisuoria. 2021. User centric social opinion and clinical behavioural model for depression detection. *International Journal of Intelligent Information Systems*, 10:69–73.

Jina Kim, Daeun Lee, and Eunil Park. 2021. Machine learning for mental health in social media: Bibliometric study. *J Med Internet Res*, 23(3):e24870.

Puneet Mathur, Ramit Sawhney, Shivang Chopra, Maitree Leekha, and Rajiv Ratn Shah. 2020. Utilizing temporal psycholinguistic cues for suicidal intent estimation. In *Advances in Information Retrieval*, pages 265–271, Cham. Springer International Publishing.

Kayalvizhi S and Thenmozhi D. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. *CoRR*, abs/2202.03047.

Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and booktitle = Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion month = May year = Mahibha C, Jerin ". Findings of the shared task on Detecting Signs of Depression from Social Media.