A Novel Text Mining Approach for Mental Health Prediction using XLNet and RNN Model

Dr. P. Maragathavalli1, Ms. Monica Emmanuel 2, Ms. Ann Sandra.j 3, and Ms. Divyabharathy.S 4
Assistant Professor[1], B.Tech Final Year Student[2]
Department of Information Technology, Puducherrry Techonological University, Puducherry, India
marapriya@ptuniv.edu.in

Abstract— **With the current advancement in the Internet, there has been a growing demand for building intelligent and smart systems that can efficiently address the detection of health-related problems on social media, such as the detection of depression and anxiety These types of systems, which are mainly dependent on machine learning techniques, must be able to deal with obtaining the semantic and syntactic meaning of texts posted by users on social media. The data generated by users on social media contains unstructured and unpredictable content. Several systems based on machine learning and social media platforms have recently been introduced to identify health-related problems. However, the text representation and deep learning techniques employed provide only limited information and knowledge about the different texts posted by users. This is owing to a lack of long-term dependencies between each word in the entire text and a lack of proper exploitation of recent deep learning schemes. A novel framework is proposed to efficiently and effectively identify depression and anxiety-related posts while maintaining the contextual and semantic meaning of the words used in the whole corpus when applying XLNet – generalized Auto Regressive model , RNN to effectively analyze and detect depression and anxiety signs from social media posts.**

Keywords — ***Twitter, XLNet, RNN, Word2Vec***

I. INTRODUCTION

With the current advancement in the Internet, there has been a growing demand for building intelligent and smart systems that can efficiently address the detection of health-related problems on social media, such as the detection of depression and anxiety.These type of systems must be able to deal with obtaining the semantic and syntactic meaning of texts on social media. However, the text representation and deep learning techniques employed provide only limited information and knowledge about the different texts posted by users.This is owing to a lack of long-term dependencies between each word in the entire text and a lack of proper exploitation of recent deep learning schemes.We employed word2vec and XLNet with RNN to effectively analyze and detect depression and anxiety signs from social media posts.

II ABBREVIATIONS AND ACRONYMS

RNN - Recurrent Neural Network
WHO - World Health Organization
LSTM - Long Short-Term Memory
Bi-LSTM - Bidirectional Long Short-Term Memory
BERT - Bidirectional Long Short-Term Memory
API - Application Programming Interface
PLM - Permutation Language Modeling
NLTK - Natural Language Toolkit
XGBoost - Extreme Gradient Boosting
CNN - Convolutional Neural Network
LIWC - Linguistic Inquiry and Word Count
SVM - Support Vector Machine
RF - Random forest
MLP - Multilayer Perceptron

III LITERATURE SURVEY

Koyel Chakraborty,S. Bhattacharyya,R. Bag et al [1] In the current era of automation, machines are constantly being channelized to provide accurate interpretations of what people express on social media. The human race nowadays is submerged in the idea of what and how people think and the decisions taken thereafter are mostly based on the drift of the masses on social platforms. This article provides a multifaceted insight into the evolution of sentiment analysis into the limelight through the sudden explosion of plethora of data on the internet. This article also addresses the process of capturing data from social media over the years along with the

similarity detection based on similar choices of the users in social networks. The techniques of communalizing user data have also been surveyed in this article. Data, in its different forms, have also been analyzed and presented as a part of a survey in this article. Other than this, the methods of evaluating sentiments have been studied, categorized, and compared, and the limitations exposed in the hope that this shall provide scope for better research in the future.

Xianghua Fu; Jingying Yang; Jianqiang Li; Min Fang; Huihui Wang et al Long short-term memory networks (LSTMs) have gained good performance in sentiment analysis tasks. The general method is to use LSTMs to combine word embeddings for text representation. However, word embeddings carry more semantic information rather than sentiment information. Only using word embeddings to represent words is inaccurate in sentiment analysis tasks. To solve the problem, we propose a lexicon-enhanced LSTM model. The model first uses sentiment lexicon as an extra information pre-training a word sentiment classifier and then get the sentiment embeddings of words including the words not in the lexicon. Combining the sentiment embedding and its word embedding can make word representation more accurate. Furthermore, we define a new method to find the attention vector in general sentiment analysis without a target that can improve the LSTM ability in capturing global sentiment information. The results of experiments on English and Chinese datasets show that our models have comparative or better results than the existing models.

Ananna Saha; Ahmed Al Marouf; Rafayet Hossain et al [3] The sentiment levels such as positive, negative and neutral sentiments from depression related posts and comments generated in social media platforms are detected. Social media platforms such as Facebook, Twitter are not only used for communication or building networks among connections, but also are getting useful for supporting needy peoples who are on special need or care in terms of mental support. In Facebook, there are several depression support groups, which are very much effective to provide mental support to the victims. In this paper, we try to formalize the depression-related posts and comments into a concise lexicon database and detect the sentiment levels form each instance. We have segmented the total work into two parts: sentiment detection and applying machine learning algorithms to analyze the ability to detect sentiment from such special category of texts. We have utilized python textblob package to detect the sentiment levels and applied traditional machine learning algorithms such as Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Sequential Minimal Optimization (SMO), Logistic Regression (LR), Adaboost (AB), Bagging (Bg), Stacking (St) and Multilayer Perceptron (MP) on the linguistic features. We have determined the precision, recall, F-measure, accuracy, ROC values for each of the classifiers. Among the classifiers Random Forest has outperformed others showing 60.54% correctly classified instance. We believe such sentiment analysis on special category of texts may lead to further investigation in natural language understanding.

Nirmal Varghese Babu , E.Grace Mary Kanaga et al [4] Sentiment analysis is an emerging trend nowadays to understand people's sentiments in multiple situations in their quotidian life. Social media data would be utilized for the entire process ie the analysis and classification processes and it consists of text data and emoticons, emojis, etc. Many experiments were conducted in the antecedent studies utilizing Binary and Ternary Classification whereas Multi-class Classification gives more precise and precise Classification. In Multi-class Classification, the data would be divided into multiple sub-classes predicated on the polarities. Machine Learning and Deep Learning Techniques would be utilized for the classification process. Utilizing Social media, sentiment levels can be monitored or analysed. This paper shows a review of the sentiment analysis on Social media data for apprehensiveness or dejection detection utilizing various artificial intelligence techniques. In the survey, it was optically canvassed that social media data which consists of texts,emoticons and emojis were utilized for the sentiment identification utilizing various artificial intelligence techniques. Multi Class Classification with Deep Learning Algorithm shows higher precision value during the sentiment analysis.

Kamil Zeberga, Muhammad Attique, Babar Shah, Farman Ali, Yalew Zelalem Jembre, and Tae-Sun Chung et al[5]With the current advancement in the Internet, there has been a growing demand for building intelligent and smart systems that can efficiently address the detection of health-related problems on social media, such as the detection of depression and anxiety. [12] discussed that Live wire with Active Appearance model (AAM) strategy is called Oriented Active Appearance Model (OAAM). The Geodesic Graph-cut calculation creates much better division results than some other completely programmed strategies distinguished in writing in the expressions of exactness and period preparing.. In addition, we propose a knowledge distillation technique, which is a recent technique for transferring knowledge from a large pretrained model (BERT) to a smaller model to boost performance and accuracy. We also devised our own data collection framework from Reddit and Twitter, which are the most common social media sites. Finally, we employed word2vec and BERT with Bi-LSTM to effectively analyze and

detect depression and anxiety signs from social media posts. Our system surpasses other state-of-the-art methods and achieves an accuracy of 98% using the knowledge distillation technique.

III. PROPOSED SYSTEM

This study presents the framework intended to retrieve, process, evaluate, and classify social networking data about mental health such as depression and anxiety. Figure 3 illustrates a general architecture that is proposed in this research work. Monitoring comments and posts on social media platforms can provide insights into how individuals self-reveal and talk about mental health issues. This type of information disseminated and shared across such platforms can be utilized as a data source for mental health problem identification. However, these types of data collected from social media are usually unstructured and include informal expressions, vague, and contentiously changing topics. It is extremely difficult to generate meaningful information from social media platforms that could be consumed in the domain of mental health evaluation and depression detection. Therefore, the proposed practical framework contains various modules such as data collection, data preprocessing, labeling techniques, word embedding, and classification. The primary objective of this research is to develop an automated system capable of discovering and evaluating mental health conditions using deep learning techniques such as XLNet and RNN

A. Dataset Preparation

The data collection procedure from one sources is discussed.

Here we collected all the latest twitter data from Kaggle .

B. WordVec

The word2vec algorithm accepts a text corpus and generates an equivalent vector associated with the given input. The generated vector will have high dimensions, and each word included in the text will be given an associated vector in vector space. The vectors are placed in space such that words with similar contextual meaning will stay in closer proximity to one another. However, the word2vec algorithm is unable to represent a new word with a vector if it is not included in the training data set.To find word representations that can infer adjacent words within a context *c* with high accuracy, the skip-gram model for a given series of words \{w1, w2,w3 , …wn \} increases the objective of the average log probability over all *N* target words and their respective contexts

$$\frac{1}{N} = \sum_{n=1}^{N} \sum_{-c \leqslant j \leqslant c_j \neq 0} \log\ P\left(w_{n+j} | w_n\right).$$

(1)

The probability predicted for a given center word is highly dependent on the inner product of the vectors used to represent the input and output candidates \{wi and wo, respectively, normalized to the requirements of a probability distribution over all words in the vocabulary of size *N* using the softmax function as follows:

$$P\left(w_o | w_i\right) = \frac{\exp\left(v_{w_o}^T v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left(v_w^T v_{w_I}\right)}. \quad (2)$$

For the entire data set collected from Twitter, we passed each pair into the neural network and trained our model to represent the collected texts

C. XLNet

XLNet, a generalized autoregressive pretraining method that enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order and overcomes the limitations of BERT thanks to its autoregressive formulation.Furthermore, XLNet integrates ideas from Transformer-XL, the state-of-the-art autoregressive model, into pretraining. Empirically, under comparable experiment settings, XLNet outperforms BERT on 20 tasks, often by a large margin, including question answering, natural language inference, sentiment analysis, and document ranking. The performance boost obtained by XLNet is mainly because of the novel language modelling objective used. This shows that there is scope for improvement of the models by improving the language models used as the pre-training objectives for the models.The following equation formally describes the language modeling objective using XLNet, for a given text sequence x, and set of all permutations of the sequence Zr and z € Zr

$$\max_{\theta} E_{z \sim Z_T}\left[\sum_{t=1}^{I} \log p_\theta\left(x_{z_t} | x_{z<t}\right)\right]$$

(3)

When these models are trained on large datasets, it achieves state-of-art performances on downstream LP tasks. This uses permutation language modelling, which trains an autoregressive model on all possible permutation of words in a sentence (3). During prediction of a word in a sequence, it takes into account bidirectional

context and predicts the masked tokens on the basis of the words/tokens to the right as we as the left of the masked token in the sequence. Another important aspect of XLnet is two-stream attention; this refers to attention streams working in parallel, one which encodes the content of the tokens and the other which incorporates the positional information. This property would be useful and so is exploited to perform the sentiment analysis on code-mixed data. [2] discussed that Biomedical and anatomical data are made simple to acquire because of progress accomplished in computerizing picture division. More research and work on it has improved more viability to the extent the subject is concerned. [7] discussed about the combination of Graph cut liver segmentation and Fuzzy with MPSO tumor segmentation algorithms. The system determines the elapsed time for the segmentation process.

V. RESULT ANALSIS AND DISCUSSION

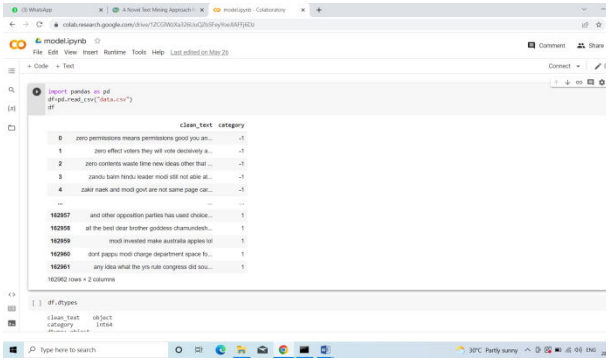

Fig.2 Loading and Exploration of input

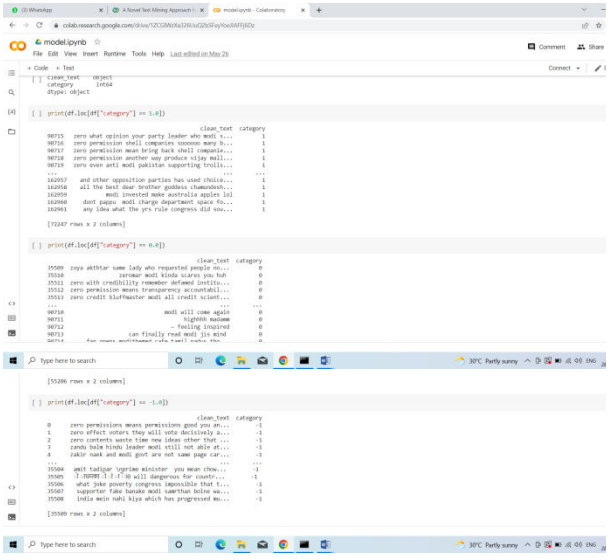

Fig.3 The collected data is separated based on their polarity.

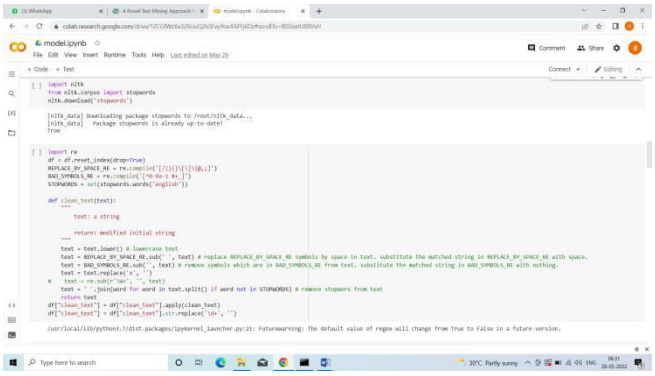

Fig. 4 Preprocessing

Most tweets consist of messages along with usernames, empty spaces, special characters, stop words, emoticons, abbreviations, hash tags, time stamps, URL's ,etc. Thus to make this data fit for mining we pre- process this data by using various functions of NLTK. [8] discussed that Automatic liver tumor segmentation would bigly influence liver treatment organizing strategy and follow-up assessment, as a result of organization and joining of full picture information. Right now, develop a totally programmed technique for liver tumor division in CT picture.

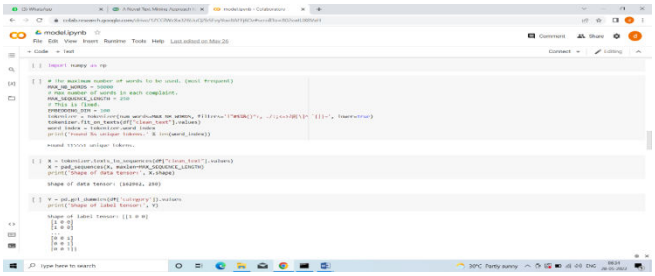

Fig.5 Data Labeling

Using a Tokenizer to assign indices and filtering out unfrequent words. Tokenizer creates a map of every unique word and an assigned index to it.

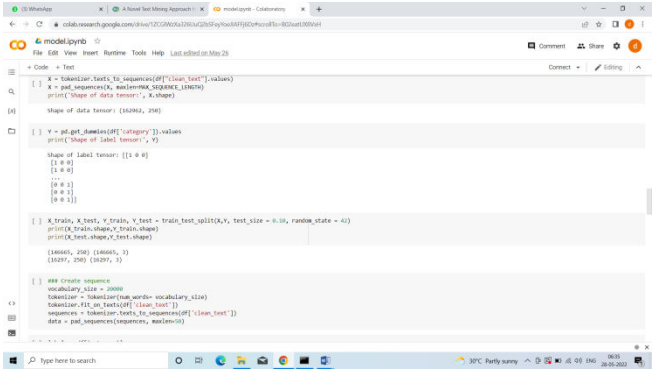

Fig. 6 Training and Testing

The training and Testing of dataset is done in the ratio 80:20

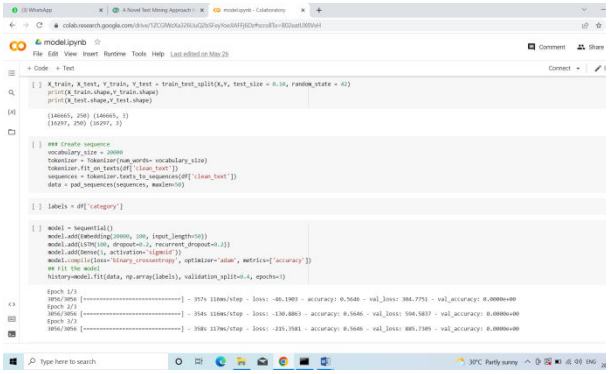

Fig. 7 Text Embedding

In the **Text embedding** block, the tokens are transformed to vectors, called word embeddings, with numerical representation of the word´s semantic meaning. That is, the word embedding inherits the meaning of the word. [13] discussed that Tumor segmentation required also the identical automatic initialization as regarding the liver. This phase was applied only in order to liver volume, obtained following automatic delineation of lean meats surface: this latter, used to original dataset quantity, was used as a new mask in order to be able to prevent processing overloads and even avoid errors related to be able to arsenic intoxication surrounding tissues delivering similar gray scale droit.

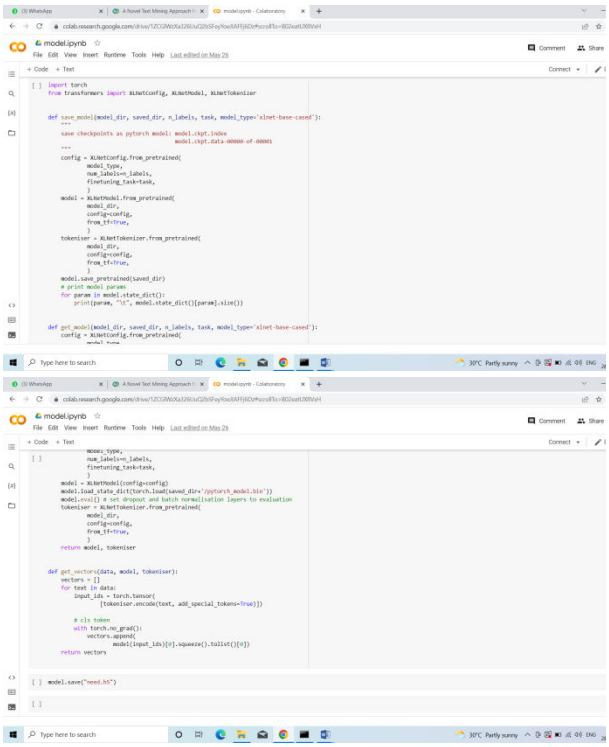

Fig. 8 Classification

Pretraining a model using XL Net

VI. CONCLUSION

The results indicates that our presented system precisely handles the mixed data and enhances the performance of mental health classification.In this study, we developed a strongly constructed framework for the detection of mental health problems using deep learning techniques such as XLNet , RNN which enhances the accuracy of smart healthcare systems to detect mental-health-related problems mainly depression and anxiety.Finally , we compared the results that is obtained from the proposed RNN with other classification models which were designed to analyze and predict mental-health-related problems from social media data.we will compare Our Proposed Approach versus State-of-the-Art Algorithms. In that experiment, the proposed RNN will be compared with ,RF, LG, SVM, AdaBoost, MLP, and LSTM algorithms for identifying depression-/anxiety-related sentiments using the data.

ACKNOWLEDGMENT

We are deeply indebted to Dr. P. Maragathavalli, Assistant Professor, Department of Information Technology, Puducherry Technological University, Puducherry, for her valuable guidance throughout the project work.

REFERENCES

[1] Koyel Chakraborty, Siddhartha Bhattacharyya, Rajib Bag “A Survey of Sentiment Analysis from Social Media Data” IEEE Transactions on Computational Social Systems https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=6570650

[2] Christo Ananth, S.Aaron James, Anand Nayyar, S.Benjamin Arul, M.Jenish Dev, “Enhancing Segmentation Approaches from GC-OAAM and MTANN to FUZZY K-C-MEANS”, Investigacion Clinica, Volume 59, No. 1, 2018,(129-138)

[3] Ananna Saha, Ahmed Al Marouf, Rafayet Hossain “Sentiment Analysis from Depression-Related User-Generated Contents from Social Media” Proceedings of the International,ICCCE https://ieeexplore.ieee.org/document/9467214

[4] Nirmal Varghese Babu, E. Grace Mary Kanaga “Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review” SN Computer Science 3:74 https://link.springer.com/article/10.1007/s42979-021-00958-1

[5] Kamil Zeberga, Muhammad Attique, Babar Shah, Farman Ali, Yalew Zelalem Jembre, and Tae-Sun Chung “A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT

Model”
https://www.hindawi.com/journals/cin/2022/7893775/
[6] P. V. Rajaraman, A. Nath, P. R. Akshaya, and G. Chatur Bhuja, “Depression detection of tweets and A comparative test,” International Journal of Engineering Research, vol. V9, no. 03, pp. 422–425, 2020.
[7] Christo Ananth, D.R.Denslin Brabin, “ENHANCING SEGMENTATION APPROACHES FROM FUZZY K-C-MEANS TO FUZZY-MPSO BASED LIVER TUMOR SEGMENTATION”, Agrociencia, Volume 54, No. 2, 2020,(72-84).
[8] Christo Ananth, D.R.Denslin Brabin, “ENHANCING SEGMENTATION APPROACHES FROM GAUSSIAN MIXTURE MODEL AND EXPECTED MAXIMIZATION TO SUPER PIXEL DIVISION ALGORITHM”, Sylwan, Volume 164, No. 4, 2020,(15-32).
[9] H. S. Alsagri and M. Ykhlef, “Machine learning-based approach for depression detection in twitter using content and activity features,” IEICE - Transactions on Info and Systems, vol. E103.D, no. 8, pp. 1825–1832, 2020.
[10] A. Yadav and D. K. Vishwakarma, “Sentiment analysis using deep learning architectures: a review,” Artificial Intelligence Review, vol. 53, no. 6, pp. 4335–4385, 2020.
[11] F. Ali, S. El-sappagh, S. M. R. Islam et al., “An intelligent healthcare monitoring framework using wearable sensors and social networking data,” Future Generation Computer Systems, vol. 114, pp. 23–43, 2021.
[12] Christo Ananth, Ruchin Jain, “Enhancing Segmentation Approaches from Oaam to Fuzzy K-C-Means”, The Journal of Research on the Lepidoptera, Volume 51, No. 2, 2020,(1086-1108).
[13] Christo Ananth, Bindhya Thomas, Priyanka Surendran, Dr.A.Anitha, “Enhancing Segmentation Approaches from GC-GGC to Fuzzy K-C-Means”, Annals of the Romanian Society for Cell Biology, Volume 25,Issue 4, April 2021,pp. 2829 – 2834

Dataset :
https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset

Dr. P. Maragathavalli
She received her B.E. degree in CSE from Bharathidasan University, M. Tech and Ph.D. degree in CSE from Pondicherry University. She joined Pondicherry Engineering College in 2006 and currently working as Assistant Professor in the Department of Information Technology. She has published several research papers in various refereed journals and international conferences. Her area of interest includes Security Testing, Optimization Techniques, Genetic Algorithm, Machine Learning and Information Security. She is a Life member of ISTE.

Divyabharathy .S
She is pursuing her B. Tech degree in the Department of Information Technology, Puducherry Technological University.

Ann Sandra.J
She is pursuing her B. Tech degree in the Department of Information Technology, Puducherry Technological University.

Monica Emmanuel
She is pursuing her B. Tech degree in the Department of Information Technology, Puducherry Technological University