ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

Performance Evaluation Of Random Forest Classifier For Depression Detection From Tweets With Different Word Embedding Models

RESEENA MOL N.A[1]
Research Scholar,Dept.Of Computer Science
Karpagam Academy of Higher Education
Coimbatore,India
reseena.n.a@gmail.com

Dr.S.VENI[2]
Professor,Dept.Of Computer Science
Karpagam Academy of Higher Education
Coimbatore,India
venikarthik04@gmail.com

Abstract: **Due to spontaneous growth in internet usage , immense amount of data is generated in social media. And these online data can used for various purposes, in this paper data generated as tweets are used for detecting depression of twitter account holders at an early stage. Early detection of depression is very crucial while considering mental health issues. Machine learning and deep learning classifiers are more commonly used in the area of sentiment analysis. As sentiment analysis is the base for detecting depression from tweets, here we consider Random Forest classifier for the objective and evaluated performance with different word embedding models like CBOW ,Skip Gram and FastText .**

***Keywords** : Social media,Depression detection,Machine learning,Sentiment analysis,Random Forest.*

I. INTRODUCTION

As per WHO more than 264 millions of people were affected by Depression ,which is a common disorder that will hurt a person critically and can cause to perform poorly at their profession and even in their family. And to some extent it can become a reason for suicide [1]. In most of the times, Depression is not detected at an early stage, but by analyzing tweets it become now possible to recognize the depressive nature of the text like sentiment analysis extract the sentiment polarity of text. By identifying depressive tweets, treatment can be started early and which may be significant for the patient.

Sentiment analysis combines text mining, NLP and computational intelligence in order to classify texts as per their emotion polarity. Different Machine learning techniques and deep learning models were used in this area and ensemble techniques also provides better results. In this paper an Ensemble classifier Random Forest resolves the classification problem, combines base classifiers and decision is based on majority voting or on an average depending on the combination, which in turn results in better accuracy than base classification methods. In addition to the classifier ,preprocessing and word embedding models also have vital role in the classification process and they all are also explored in this paper.

Rest of the paper is arranged as section 2 covers Related works, Proposed Methodology in Section 3, Experimental Results in Section 4 and paper is concluded in Section 5.

II. RELATED WORKS

[12]In the proposed work, detection of depressive and non depressive tweets from Arabic tweet set is performed. Different machine learning models Random forest,Naive bayes,Adaboost and Liblinear were used for implementing the classification process and Liblinear provides higher accuracy.

[13]Authors focused on analyzing the sentiments on tweets and thus by determining the depressive or non depressive tweets. Different machine learning

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

algorithms were used here for classification process- Tf- IDF,Bag of words and multinomial naive bayes and they reached at the conclusion that multinomial naive bayes algorithm shows better accuracy when compared with the rest.

[14]As per authors, machine learning classifiers can be used for classifying tweets as depressed or not depressed. Processes that done on the dataset were preprocessing, feature extraction, taking account measures based on user activity in their twitter accounts and then applied classifiers for the final result. SVM and Naive bayes are used in this work and found that SVM gave better result.

[15]In this paper authors proved through their experiment that traditional vectorisers affect the accuracy of classifiers in sentiment analysis, so they used word embedding model ,word2vec for generating high dimensional vectors. From the results it is found that accuracy of classification process improves when vectorizers are replaced with word embeddings.

[16]In the proposed work, sentiment polarity classification on four data set is performed. Different vector representations, lexicon based,word embedding based and hybrid approaches were used. Based on experiments, it is concluded that hybrid approach results better accuracy.

[17]Through this work, authors stated that word embeddings helps to classify different types of words in an efficient way than the old bag of words approach while implementing with supervised machine learning approaches. Analysed patterns of negativity in Australian parliamentary speeches. In the work accuracy of machine learning classifier employing word embedding is compared with that of bag of words vectorisation. It is stated that word embedding helps to identify semantics of the words even if they are not included in the training data set.

[19]In this work sentiment analysis on amazon data set is performed to recognize the polarity of sentiment involved in the text. Here a hybrid approach for classification is employed by combining SVM with Random Forest and through experiment , it is found that the combined classifier provides better results than pure SVM or RF.

[20] Authors worked on tweets for sentiment analysis by employing different machine learning techniques using various embedding models like Word2Vec,Fasttext and Glove.GaussianNB,Linear SVC,NuSVC,Logistic Regression,SGD and Random forest were explored and reached at the conclusion that Fasttext embeddings performance out weigh other embedding models.

III. PROPOSED METHODOLGY

This section presents the methodology used in the proposed work. Process is completed as a sequence of steps. Fig. 1 depicts the system architecture.

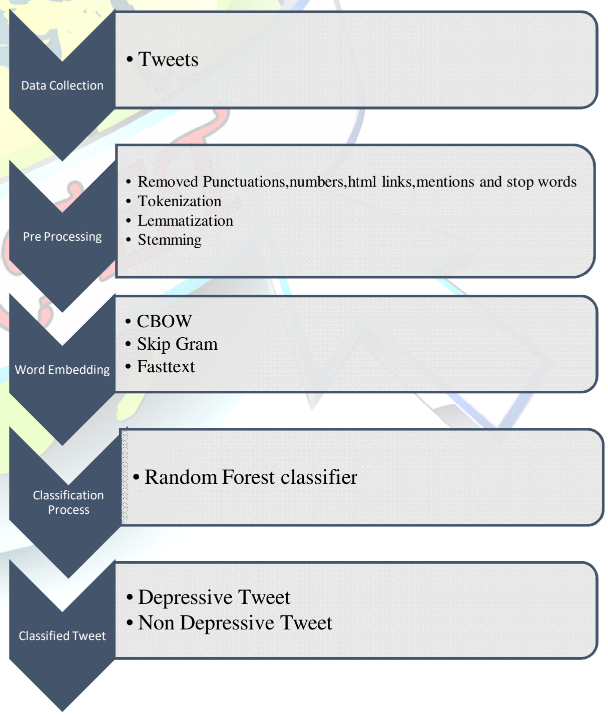

FIG. 1. SYSTEM ARCHITECTURE

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

A. Data Collection

Our work is on Tweets and here we collected tweets from an open Kaggle data set ,containing 3657 depressed tweets and 3356 non depressed tweets its count is represented in Fig. 2.

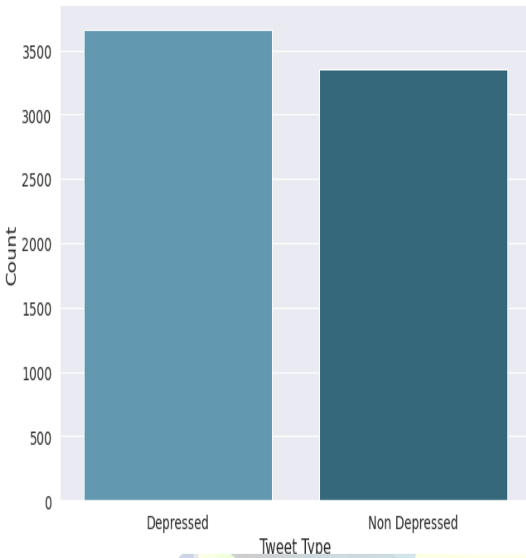

FIG. 2. NUMBER OF TWEETS UNDER EACH CLASS

B. Data Preprocessing

Quality of classification process depends on the quality of data [2]. Usually online data contains lots of noises and hence preprocessing is required to clean the data for better results [2]. Tasks that are done during preprocessing are Removed numbers, punctuations,stop words,html links, and mentions like #,@,https:// etc. ,done tokenization and performed lemmatization and stemming.

C. Word Embedding

After the preprocessing task, data is to be converted into numerical values as the classifiers could not able to process texts in its raw form. In earlier classification methods vectorisers like one hot encoding,count vectorizers etc. were used but due to its large dimensional size and similarity issues word embeddings were brought into usage[3]. They are dense, distributed, fixed-length word vectors, built using word co-occurrence statistics as per the distributional hypothesis[4]. In this work three embedding methods were used for evaluating performance of Random Forest classifier with different embedding models.

1. WORD2VEC
   Syntactic meaning of words are maintained in Word2Vec model and words are organized by their syntactic similarity[5]. For prediction of supplementary words in a sentence, Word2Vec have two algorithms[6]:
   - CBOW(Continuous Bag of Words) – In this model context is given by multiple surrounding words and the center ,target word is predicted with the help of these surrounding words[7].
   - SG(Skip Gram) – In this model, trying to predict the surrounding context words with the help of center word[18]

2. FASTTEXT
   This model , for each n-grams generates vectors ie. for sub parts of words[8]. Its base is Skip Gram model, but provides more accuracy for classification process.

D. Classification

For classification process data set available will be split up into training and testing set[9]. A classifier model is developed with the help of training dataset and accuracy of the model is checked with testing dataset.

Different classifiers were popularly used in the area of sentiment analysis including Machine learning models, ensemble models and Deep learning models. In this proposed work an ensemble model is used. Ensemble model helps to improve the performance of the classifier by merging up of different base classifiers. Random Forest is used in this work as classifier.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

- Random Forest

It is one of the most popular ensemble algorithm. Random forest algorithm trains on multiple decision trees driven on slightly different subsets of data. In Random forest classification method, many classifiers are generated from smaller subsets of the input data and after that their individual results are aggregated based on a voting mechanism to generate the desired output of the input data set[10].

E. *Performance Evaluation*

After classification model is developed, their accuracy is to be measured. Metrics considered for performance evaluation are Accuracy, Precision, Recall and F1 score and these measures are based on true positivity and true negativity on result of classification process [11].

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (3)$$

$$\text{F1Score}=2\text{x}\frac{Recall x Precision}{Recall+Precision} \quad (4)$$

IV. EXPERIMENTAL RESULTS

In this work, Random Forest classification is employed for implementation. Before training the model ,pre processed data set need to be vectorised and for that three word embedding models are applied here-CBOW,Skip Gram and FastText. Vectors obtained from these three embedding models were passed to the training phase and then the result accuracy of these three were compared and shown in Table 1.

TABLE 1. PERFORMANCE EVALUATION METRICS FOR RANDOM FOREST CLASSIFIER

Classifier	Word Embedding Models	Accuracy	Precision	Recall	F1-Score
Random Forest	CBOW	87.17	84.21	92.56	88.18
	Skip Gram	93.65	93.21	94.62	93.91
	FastText	94.15	93.75	95.04	93.36

From the results obtained, it is observed that both Skip gram and Fasttext shows better accuracy, but Fasttext exhibit a slight greater accuracy . Fig. 3 shows chart representation of accuracy.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

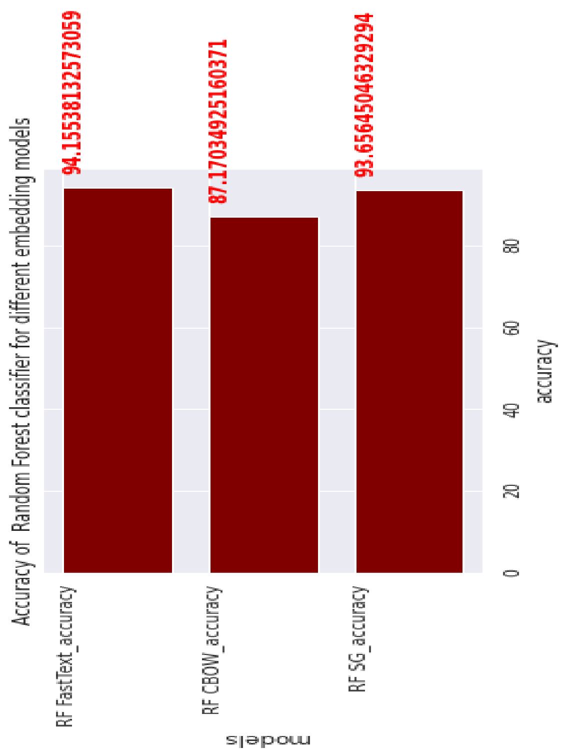

FIGURE. 3. ACCURACY MEASURES OF RANDOM FOREST CLASSIFIER FOR DIFFERENT EMBEDDING MODELS

V. CONCLUSIONS

In this work , researchers tried to identify depressive tweets from set of tweets using an ensemble classification model Random Forest with three different word embedding models-CBOW,Skip gram and FastText. As word embedding models also have a significant role in the performance of the classifier we tried with different embedding models. From the result obtained, we arrived at the conclusion that optimal accuracy performance was with FastText word embedding model with Random Forest classifier. With this work we got the assurance that selection of word embedding model for a classifier is also vital in concern with its accuracy. In our future work we will try to find out an optimized machine learning model with suitable word embedding model by exploring various classifier models with different embedding models.

REFERENCES

[1]S. Bennett, C. Coggan, and P. Adams, "Problematising depression: Young people, mental health and suicidal behaviours," Soc. Sci. Med., vol. 57, no. 2, pp. 289–299, 2003, doi: 10.1016/S0277-9536(02)00347-7.

[2]E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," Procedia Comput. Sci., vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.

[3]L. C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings for sentiment analysis," EMNLP 2017 - Conf. Empir. Methods Nat. Lang. Process. Proc., pp. 534–539, 2017, doi: 10.18653/v1/d17-1056.

[4]F. Almeida and G. Xexéo, "Word embeddings: A survey," arXiv, no. 1991, 2019.

[5]Md Al-Amin, M. S. Islam, and S. Das Uzzal, "Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words," ECCE 2017 - Int. Conf. Electr. Comput. Commun. Eng., pp. 186–190, 2017, doi: 10.1109/ECACE.2017.7912903.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 8, Special Issue 1, August 2021

[6]B. Xue, C. Fu, and Z. Shaobin, "A study on sentiment computing and classification of sina weibo with Word2vec," Proc. - 2014 IEEE Int. Congr. Big Data, BigData Congr. 2014, no. 2013, pp. 358–363, 2014, doi: 10.1109/BigData.Congress.2014.59.

[7]S. Al-Saqqa and A. Awajan, "The Use of Word2vec Model in Sentiment Analysis: A Survey," ACM Int. Conf. Proceeding Ser., pp. 39–43, 2019, doi: 10.1145/3388218.3388229.

[8]N. Nedjah, I. Santos, and L. de Macedo Mourelle, "Sentiment analysis using convolutional neural network via word embeddings," Evol. Intell., pp. 2–6, 2019, doi: 10.1007/s12065-019-00227-4.

[9]M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," 2013 4th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2013, 2013, doi: 10.1109 /ICCCNT .2013. 6726818.

[10]Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, "Random forest and support vector machine based hybrid approach to sentiment analysis," Procedia Comput. Sci., vol. 127, pp. 511–520, 2018, doi: 10.1016/j.procs.2018.01.150.

[11]Ankit and N. Saleena, "An Ensemble Classification System for Twitter Sentiment Analysis," Procedia Comput. Sci., vol. 132, no. Iccids, pp. 937–946, 2018, doi: 10.1016/j.procs.2018.05.109.

[12]S. Almouzini, M. Khemakhem, and A. Alageel, "Detecting Arabic Depressed Users from Twitter Data," Procedia Comput. Sci., vol. 163, pp. 257–265, 2019, doi: 10.1016/j.procs.2019.12.107.

[13]M. Ambika, K. V Devakrishnan, A. Divya, R. G. Raj, and K. Kaviyaa, "Detection of Depression and Mental illness of Twitter users using Machine Learning," Int. J. Eng. Adv. Technol., vol. 9, no. 4, pp. 1331–1335, 2020, doi: 10.35940/ijeat.d8314.049420.

[14]P. S. J. Pachouly, G. Raut, K. Bute, R. Tambe, and S. Bhavsar, "Depression Detection on Social Media Network (Twitter) using Sentiment Analysis," pp. 1834–1839, 2021.

[15]O. B. Deho, W. A. Agangiba, F. L. Aryeh, and J. A. Ansah, "Sentiment analysis with word embedding," IEEE Int. Conf. Adapt. Sci. Technol. ICAST, vol. 2018-Augus, no. March, pp. 1–4, 2018, doi: 10.1109/ICASTECH.2018.8506717.

[16]M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings," Expert Syst. Appl., vol. 69, pp. 214–224, 2017, doi: 10.1016/j.eswa.2016.10.043.

[17]E. Rudkowsky, M. Haselmayer, M. Wastian, M. Jenny, Š. Emrich, and M. Sedlmair, "More than Bags of Words: Sentiment Analysis with Word Embeddings," Commun. Methods Meas., vol. 12, no. 2–3, pp. 140–157, 2018, doi: 10.1080/19312458.2018.1455817.

[18]M. Tolba, S. Ouadfel, and S. Meshoul, "Hybrid ensemble approaches to online harassment detection in highly imbalanced data," Expert Syst. Appl., vol. 175, no. January, p. 114751, 2021, doi: 10.1016/j.eswa.2021.114751.

[19]Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, "Random forest and support vector machine based hybrid approach to sentiment analysis," Procedia Comput. Sci., vol. 127, pp. 511–520, 2018, doi: 10.1016/j.procs.2018.01.150.

[20]I. Kaibi, E. H. Nfaoui, and H. Satori, "A comparative evaluation of word embeddings techniques for twitter sentiment analysis," 2019 Int. Conf. Wirel. Technol. Embed. Intell. Syst. WITS 2019, pp. 1–4, 2019, doi: 10.1109/WITS.2019.8723864.