IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

STRESS DETECTION USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING

[1] Jerripotula Priyanka, [2] Prof. K. Venkata Rao
[1]M. Tech Student, [2]Professor
[1,2]Department of CS & SE, AUCE, Andhra University, Visakhapatnam

ABSTRACT:

Nowadays many users posts tweets based on their mental condition about the things that happen in their day to day lives on the social media platforms. It is very important to detect and manage stress before it goes into a severe problem. A huge number of informal messages are posted every day in social networking sites, blogs and discussion forums. This paper describes an approach to detect the stress using the information from social media networking sites, like twitter. This project performs the operations involving data collection, data cleaning, training the machine and predicting the stressed and non-stressed users. This will be using the Natural Language Processing (NLP) and Machine Learning algorithms which include KNN ,Naïve bayes BernoulliNB, Random Forest, Decision tree and SVM. Psychological stress is threatening people's health. It is non-trivial to detect stress timely for proactive care. With the popularity of social media, people are used to sharing their daily activities and interacting with friends on social media platforms, making it feasible to leverage online social network data for stress detection. In this paper, we find that users stress state is closely related to that of his/her friends in social media, and we employ a large-scale dataset from real-world social platforms to systematically study the correlation of users stress states and social interactions. We first define a set of stress-related textual undergoes the training of the machine followed by the machine learning algorithms for better results. Thus the proposed system takes the tweets as input and decides whether it is stressed or non-stressed.

Keywords: NLP, KNN, SVM, Navie Bayes, Random Forest

1. INTRODUCTION:

The social media like Facebook, Twitter, and WhatsApp are the popular social networking website where the users create and use these social media application for conveying the message and expressing their individual feelings, thoughts to their friends and family on different subjects. The social media applications are highly influencing people of all age group peoples nowadays and changing the lifestyle of the people. As the continuous use of social media by the users, it is much more possible to identify the psychological state of the user by gathering their social media message and communication data timely and by analyzing the content.

Psychological stress which is a medical and physical illness that is a threat to the people health. Year by year the stress level was increasing to the people and sometimes the overwhelming stress leads to suicidal ideation. In the year 2018, about 15.7 per 100k people were suicide in India. Though the stress became more common in our day to day life with being harmful to the human life. So there is much important to predict the stress level of the people before it turns into a severe problem. Some of the traditional methods to detect the stress are actually reactive and has some limitation like hysteretic, time and labor consuming and computationally expensive. As the increased use of social media application by all age group it much more possible to find the emotional or stress state of the user in the earlier stage by using machine learning techniques which is much better than the traditional methods.

2. LITERATURE SURVEY:

A lot of astounding contributions have been made in the field of sentiment analysis in the past few years. Initially, sentiment analysis was proposed for a simple binary classification that allocates evaluations to bipolar classes. Alexander Pak and Patrick Paroubek [5] came up with a model that categorizes the tweets into three classes. The three classes were objective, positive and negative. In their research model, they started by generating a collection of data by accumulating tweets. They took advantage of the Twitter API and would routinely interpret the tweets based on emoticons used. Using that twitter corpus, they were able to construct a sentiment classifier. This classifier was built on the technique—Naive Bayes where they used N-gram and POS-tags. They did face a drawback where the training set turned out to be less proficient since it only contained tweets having emoticons. The papers [6–10] discuss effective data pre-processing techniques for social media content, specifically tweets. As the data contains the words which are most often used in a sentence but do not contribute to the analysis, such as stop words, symbols, punctuation marks. Removing these and converting different forms of the words to the base from is an essential step.

Sentiment analysis Apoorv Agarwal et al. [11] proposed a 3-way model for categorizing sentiments in three classes. The classes were positive, negative, and neutral. Models such as the unigram model, a feature constructed upon the model, and a tree kernel-based were used for testing. In the case of the tree kernel-centered model, tweets were chosen to be represented in the form of a tree. While implementing a feature-centered model over 100 features were taken into consideration. However, in the case of the unigram model, there were about 10,000 features. They concluded that features that end up combining previous polarization of words with their parts-of-speech (pos) tags are the most substantial. In terms of the result, the tree kernel-based model ended up performing better than the other two models. Certain challenges are made by a few researchers to classify public beliefs about movies, news, etc. from Twitter posts. V.M. Kiran Peddinti et al. [12] utilized the data from other widely accessible databases like IMDB and Blippr after appropriate alterations to benefit Twitter sentiment analysis in the movie domain. Davidov Dmitry et al. [13] projected a method to utilize Twitter user-defined hashtags in tweets as a classification of sentiment type using punctuation, single words, and patterns as disparate feature types. They are then combined into a single feature vector for the task of sentiment classification. They made use of the K-Nearest Neighbor approach to allocate sentiment labels by constructing a feature vector for each example in the training and test set. Tagging [14], in current times developed as a common way to sort out vast and vibrant web content. It usually refers to the act of correlating with or allocating some keyword or unit to a piece of data. Tagging aids to depict an article and lets it be located again by perusing. Scholars have established diverse methods and procedures for tagging corpus for numerous uses. Xiance et al. [15] offered a flexible and practical technique for the process of the recommendation of tags. They demonstrated documents and tags by implementing the tag-LDA model. Krestel et al. [16] recommended a method to customize the process of recommendation by tag. She proposed a method that amalgamates a probabilistic method of tags from the source. In this case, the tags were extracted from the user. She examined basic language models. Additionally, she performed LDA experimentations on a real-world dataset. The dataset was crawled from a vast tagging system which displayed that personalization progresses the process of tag recommendation. [17-27] These researchers have made significant contributions to stress detection and analysis through their innovations in natural language processing and sentiment analysis. Peters and Neumann introduced deep contextualized word representations in 2018, enhancing the understanding of stress-related language. Radford and Narasimhan's generative pre-training in the same year improved language comprehension, enabling more accurate stress detection. Devlin, Chang, Lee, and Toutanova's BERT, presented in 2019, has been instrumental in advancing stress analysis by pre-training deep bidirectional transformers for language understanding. Jin, Lai, and Cao's work in 2020 applied BERT and modified TF-IDF for multi-label sentiment analysis, aiding in nuanced stress assessment in text data

Stress/depression analysis

Arya and Mishra [28] present a review of the application of machine learning in the health sector, their limitation, predictive analysis, and challenges in the area and need advanced research and technologies. The authors reviewed papers on mental stress detection using ML that used social networking sites, blogs, discussion forums, Questioner technique, clinical dataset, real-time data, Bio-signal technology (ECG, EEG), a wireless device, and suicidal tendency. The study shows the high potential of ML algorithms in mental health [28]. Aldarwish et al used machine learning algorithms SVM and Naïve- Bayesian for Predicting stress from UGC- User Generated Content in Social media sites (Facebook, Twitter, Live Journal) they used social interaction stress datasets based on mood and negativism and BDI- questionnaire having 6773 posts, 2073 depressed, 4700 non-depressed posts (textual). They achieved an accuracy of 57% from SVM and 63% from Naïve- Bayesian. They also emphasized stress detection using big data techniques [29]. Cho G et al. presented the analysis of ML algorithms for diagnosing mental illness. They studied properties of mental health, techniques to identify, their limitations, and how ML algorithms are implemented. The authors considered SVM, GBM, KNN, Naïve Bayesian, KNN, Random Forest. The authors achieved 75% from the SVM classifier [30]. Reshma et.al proposed a Tensi Strength framework for detecting sentiment analysis on Twitter [31]. The authors considered SVM, NB, WSD, and n-gram techniques on large social media text for sentiment analysis and applied the Lexicon approach to detect stress and relaxation in large data set. The authors achieved 65% precision and 67% recall. Deshpande and Rao presented an emotion artificial intelligence technique to detect depression [32]. The authors collected 10,000 Tweet Using Twitter API.

3. METHODOLOGY:

3.1 Support Vector Machine (SVM):

An SVM methodology denotes its appearance between objects throughout region, which were displayed in such a way that either the groups were divided by either a great distance. The goal is to evaluate the maximum-margin hyperplane that provides the best class divide. The instances that are often nearest to a hyperplane with maximum margins are named vectors of assistance. The selected variables were focused onto the portion of a datasets that denotes a learning set. Dual-class support variables allow for all the formation with two simultaneous hyperplanes. Furthermore, the greater its boundary between other two hyperplanes, the classifier's failure of generalization would be stronger. Relative to several other machine learning methods SVMs are developed inside a special way.

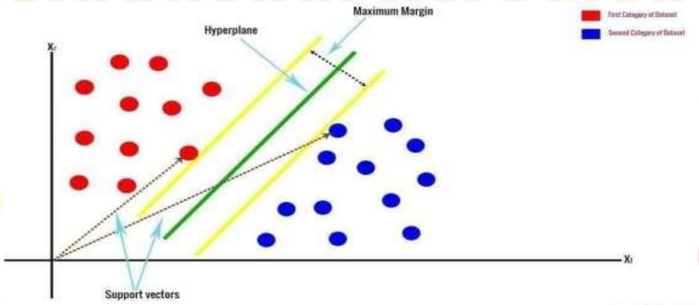

Figure 1: SVM

3.2 RANDOM FOREST

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. In our project for random forest algorithm, we used 21 estimators that is nothing but a decision tree to generate a model.

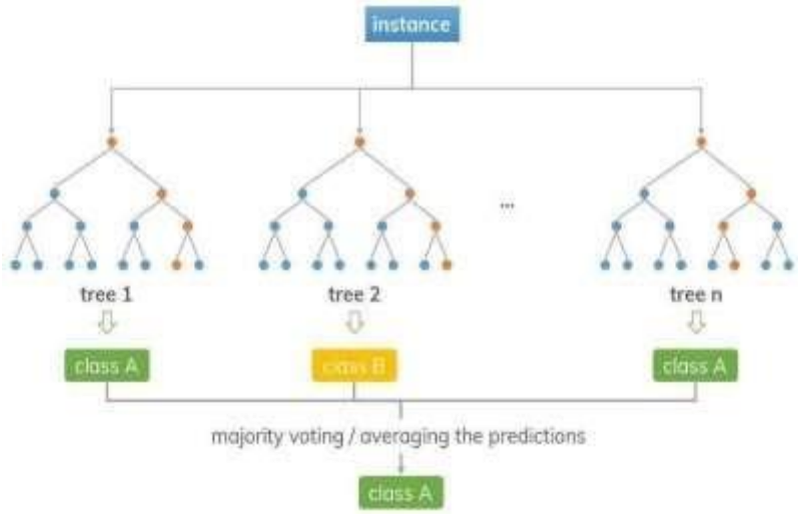

Figure 2: Random Forest

3.3 KNN

K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. We can understand its working with the help of following steps −

Step 1 − They want data frame for enforcing some method. And we must pack its instruction and perhaps even the relevant data within the first phase in KNN.

Step 2 − First, we will pick the K amount i.e., its closest information values. Some number could be K.

Step 3 − Does the preceding to every level well into the information −

− Using any of the methods notably: Manhattan, Euclidean or Hamming distance measure the distance among testing data so each line of sample data. Its most frequently utilized form for range calculation is Euclidean.

− Now, based on the distance value, sort them in ascending order.

− Next, list the top K lines from both the list you have ordered.

− Now, the category would be allocated to both the check points based on its most common classes of those lines.

Step 4 – End

In our project for KNN algorithm we used values of n as n=1,3,5,7,9 to generated a model.

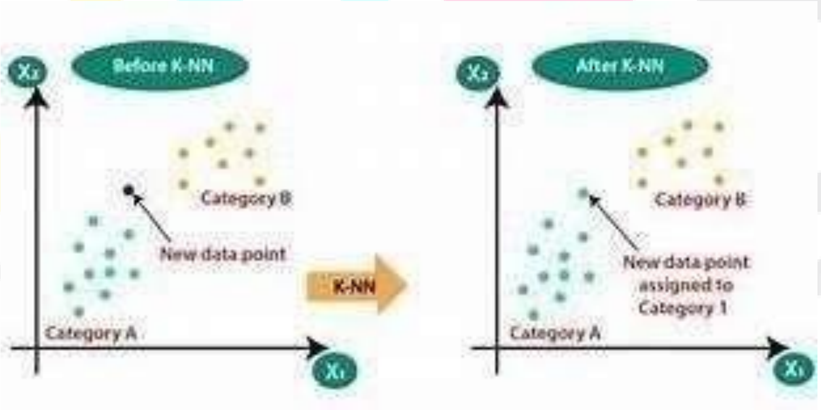

Figure 3 : KNN

3.4 DECISION TREES:

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.

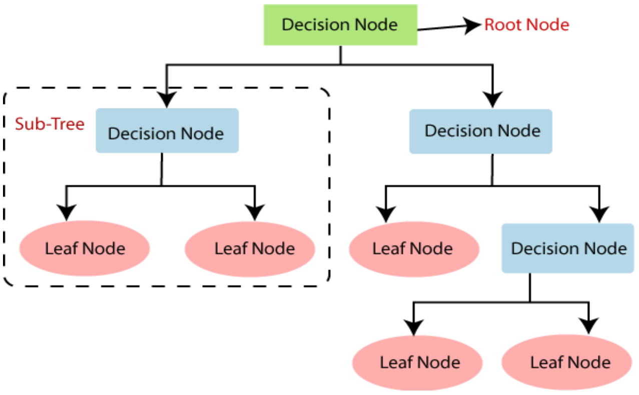

Figure 4 : Decision Tree

3.5 NAÏVE BAYES CLASSIFIER ALGORITHM:

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high- dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as Naïve. It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other. Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes' Theorem:

Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula for Bayes' theorem is given as:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A, B = events

$P(A|B)$ = probability of A given B is true

$P(B|A)$ = probability of B given A is true

$P(A), P(B)$ = the independent probabilities of A and B

Naïve Bayes Classifier Algorithm Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence P(B) is Marginal Probability: Probability of Evidence.

4. RESULTS AND DISCUSSION:

Figure 5: Dataset sample words

Stress Detection Model The label column in this dataset contains labels as 0 and 1. 0 means no stress, and 1 means stress. Stress and No stress labels were used instead of 1 and 0. So let’s prepare this column accordingly and select the text and label columns for the process of training a machine learning model:

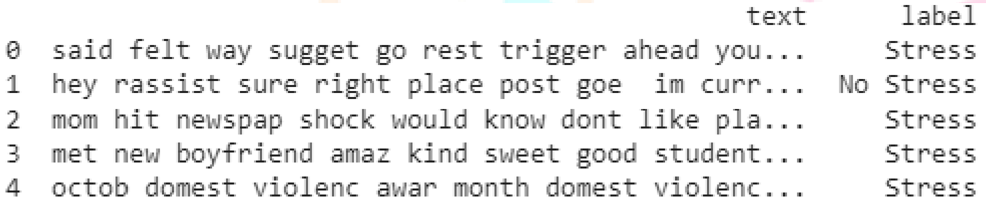
```
                                                text      label
0  said felt way sugget go rest trigger ahead you...     Stress
1  hey rassist sure right place post goe  im curr...  No Stress
2  mom hit newspap shock would know dont like pla...     Stress
3  met new boyfriend amaz kind sweet good student...     Stress
4  octob domest violenc awar month domest violenc...     Stress
```

Figure 6: input dataset with label

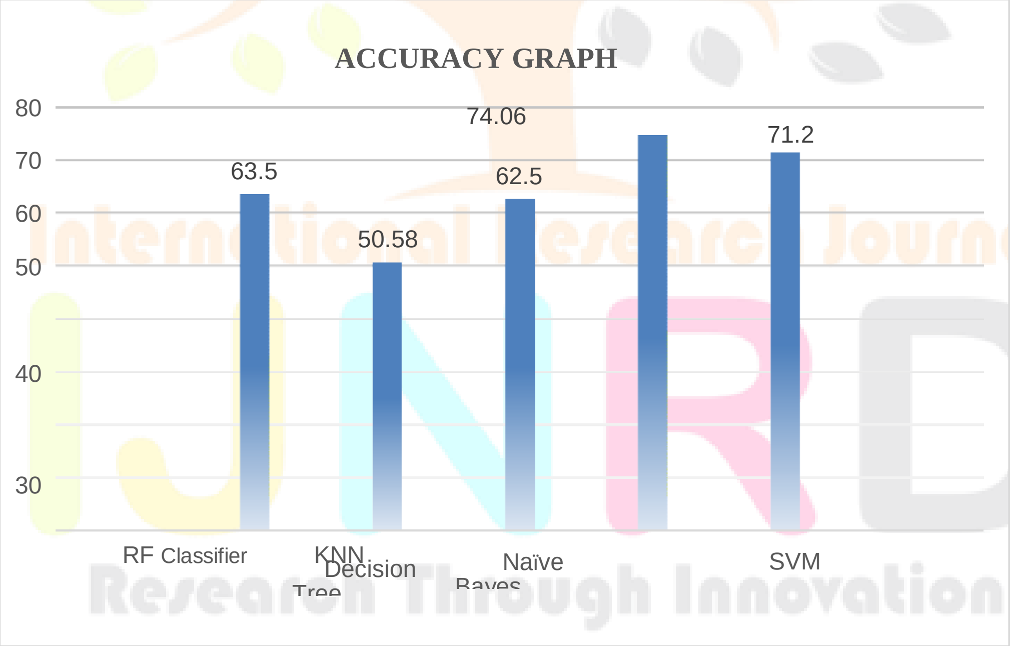

Figure 7: Accuracy Graph

5. CONCLUSION:

In today’s world, where mainly the youth and almost all of the population is suffering from surmounting stress, it is because of peer pressure, work load or other domestic tensions; it is very crucial to have a reality check about how stressed a person really is. Due to this reason that timely detection and prevention of stress is a dire need. We have come up with this project which assists people in scrutinizing the problem of stress. This project will be very beneficial for those who are not so comfortable in opening up about their problems to others. It will help these people get a reality check and may prompt them to reach out and get medical help, just based on their social interactions. We have utilized both human as well as machine learning and applied the concepts of Sentiment Analysis. The main characteristic of this system is its non-invasiveness and fast-oriented implementation in detecting stress when compared with the previous approaches.

References

1. Liang Y, Zheng X, Zeng DD. A survey on big data-driven digital phenotyping of mental health. Inform Fusion. 2019;52(1):290–307.
2. Liu B, Zhang L. A survey of opinion mining and sentiment analysis. Boston: Springer US. 2012; p. 415–463.
3. Munikar M, Shakya S, Shrestha A. Fine-grained sentiment classification using BERT. Artif Intell Transform Business Society. 2019;2019:1–5. https://doi.org/10.1109/AITB48515.2019.8947435.
4. Wang B, Liu Y, Liu Z, Li M, Qi M. Topic selection in latent Dirichlet allocation, 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). 2014. p. 756–760. https://doi.org/10.1109/FSKD.2014.6980931.
5. Alexander P, Patrick P. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC. 2010.
6. Jianqiang Z, Xiaolin G. Comparison research on text pre-processing methods on Twitter sentiment analysis. IEEE Access. 2017;5:2870–9. https://doi.org/10.1109/ACCESS.2017.2672677.
7. Pradha S, Halgamuge MN, Vinh NQT. Effective text data preprocessing technique for sentiment analysis in social media data, 2019 11th International Conference on Knowledge and Systems Engineering (KSE). 2019. p. 1–8.https://doi.org/10.1109/KSE.2019.8919368.
8. Deepa DR, Tamilarasi A. Sentiment analysis using feature extraction and dictionary-based approaches, 2019 Third International conference on I-SMAC (IoT in SociaMobile, Analytics, and Cloud) (I-SMAC). 2019. p. 786–790. https://doi.org/10.1109/I-SMAC47947.2019.9032456.
9. Chaturvedi S, Mishra V, Mishra N. Sentiment analysis using machine learning for business intelligence, 2017 IEEE International Conference on power, control, signals, and instrumentation engineering (ICPCSI). 2017. p. 2162–2166. https://doi.org/10.1109/ICPCSI.2017.8392100.
10. Ho J, Ondusko D, Roy B, Hsu DF. Sentiment analysis on tweets using machine learning and combinatorial fusion,2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). 2019. p. 1066–1071. https://doi.org/10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00191.
11. Apoorv A, Boyi X, Ilia V, Owen R, Rebecca P. Sentiment analysis of Twitter Data. Proceedings of the Workshop on Languages in Social Media. 2011.
12. Peddinti MK, Chintalapoodi P. Domain adaptation in sentiment analysis of Twitter, in Analyzing Microtext Workshop, AAAI, 2011.
13. Dmitry D, Oren T, Ari R. Enhanced sentiment learning using twitter hashtags and smileys. Coling 2010—23rd International Conference on Computational Linguistics, Proceedings of the Conference. 2. 2010; 241–249.
14. Anupriya P, Karpagavalli S. LDA based topic modeling of journal abstracts. Int Conf Adv Comput Commun Syst. 2015;2015:1–5. https://doi.org/10.1109/ICACCS.2015.7324058.
15. Xiance S, Maosong S. Tag-LDA for scalable real-time tag recommendation. J Comput Inform Syst. 2008;6:23.
16. Krestel R, Fankhauser P. Personalized topic-based tag recommendation. Neurocomputing. 2012;76:61–70. https://doi.org/10.1016/j.neucom.2011.04.034.
17. Peters ME, Neumann M. Deep contextualized word representations. 2018.
18. Radford A, Narasimhan K. Improving language understanding by generative pre-training. 2018.
19. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, vol 1.

Minneapolis; 2019. p. 4171–4186. https://doi.org/10.18653/v1/n19-1423.

20. Jin Z, Lai X, Cao J. Multi-label sentiment analysis base on BERT with modifed TF-IDF, 2020 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-CN), 2020. https://doi.org/10.1109/ISPCE-CN51288.2020.9321861.
21. Zubair M, Aurangzeb K, Shakeel A, Maria Q, Ali KI, Quan Z. Lexicon-enhanced sentiment analysis framework using rule-based classification scheme. Plos One. 2017;12(2):e0171649.
22. Zeng D, Dai Y, Li F, Wang J, Sangaiah AK. Aspect based sentiment analysis by a linguistically regularized CNN with gated mechanism. J Intell Fuzzy Syst. 2019;36(5):3971–80. https://doi.org/10.3233/JIFS-169958.
23. Alec G, Richa B, Lei H. Twitter sentiment classification using distant supervision. Processing. 2009; 150.
24. Alexandra B, Ralf S, Mijail K, Vanni Z, van der Erik G, Matina H, Bruno P, Jenya B. Sentiment analysis in the news. proceedings of LREC. (n-1). 2013
25. Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL Student Research Workshop. Association for Computational Linguistics, USA, 43–48. [N-2]. 2005.
26. Boiy E, Moens MF. A machine learning approach to sentiment analysis in multilingual web texts. Inf Retrieval. 2009;12:526–58. https://doi.org/10.1007/s10791-008-9070-z[N+1].
27. Li F, Huang M, Zhu X. Sentiment analysis with global topics and local dependency. In Proceedings of the TwentyFourth AAAI Conference on Artificial Intelligence, AAAI Press. 2010. 1371–1376.Nijhawan et al. Journal of Big Data (2022) 9:33 Page 24 of 24
28. Arya V, Mishra AK. Machine learning approaches to mental stress detection: a review. Ann Optimization Theory Pract. 2021;31(4):55–67.
29. Aldarwish MM, Ahmad HF. Predicting Depression Levels Using Social Media Posts, 2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS), Bangkok. 2017. pp. 277–280. https://doi.org/10.1109/ISADS.2017.41.
30. Cho G, Yim J, Choi Y, Ko J, Lee SH. Review of machine learning algorithms for diagnosing mental illness. Psychiatry Invest. 2019;16(4):262–9.
31. Baheti RR, Kinariwala SA. Survey: sentiment stress identification using tension/strength framework. Int J Sci Res Eng Dev. 2019;2(3):1–8.
32. Deshpande M, Rao V. Depression detection using emotion artificial intelligence, 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India. 2017. p. 858–862
33. Zucco C, Calabrese B, Cannataro M. Sentiment Analysis and Affective Computing for Depression Monitoring. In 2017 IEEE international conference on bioinformatics and biomedicine (BIBM). New York: IEEE. 2017. p. 1988–1995.