Anxiety and Depression Posts in Reddit Social Media

[1]Anita Harsoor, [2]Kavita Siddgonda

[1]Associate Professor, [2]PG student
Department of Computer Science & Engineering
Poojya Doddappa Appa College of Engineering
Kalburagi, India.

***Abstract*: Depression is known to be the biggest contributor to global illness, and a significant cause of suicide. In this paper, our study's main purpose is to review posts from Reddit users to identify any factors that may expose the depression attitudes of local online users. NLP is a field in machine learning with the ability of a computer to understand, analyze, manipulate, and potentially generate human language. Here we used sentiment analysis. Instead of constructing all resources from scratch, NLTK does have all that NLP tasks. The proposed framework is developed by using Logistic Regression, Support Vector Machine, Random Forest, Adaptive Boosting and Multilayer Perceptron classifier. Logistic Regression (LR) is a linear classification approach used to estimate the probability occurrence of binary response based on one or more predictors and features. We used the clustering techniques k-mean to Cluster user data, and provide the user with accuracy stress levels. Using clustering of k-means, we developed four levels of depressive symptoms to match the classifications based on the norms. By using k-means algorithm, we achieve 99 percent accuracy and 0.99 F1 scores of detecting the depression. The results showed that our automatic classification system worked well in comparison with the approach based on the standard.**

***Index Terms*: Natural language processing, Reddit social media, Stress level prediction, Artificial intelligence, Anxiety.**

I. INTRODUCTION

Knowledge about the disclosure on social media of mental health issues is minimal. This project focuses on the Reddit website, which aims to fill a void between other social media sites such as Twitter or Facebook-frequently synonymous with anonymous permanent identities-and health fora.Reddit is a unique forum through which users can opt to build "throwaway" accounts that are not connected to their main account to make posts or comments that reveal confidential details.Younger generation often turn to mental health networking fora on social media. Looking at comments and articles on these sites will provide insight into how people reveal themselves, and address topics of mental health such as depression. Using a scraped Reddit comment dataset, this project aims at classifying depression into commentary.Depression has long been described as a common mental health condition as a single illness, with a collection of diagnostic criteria.It also co-occurs with anxiety or other psychological and physical disorders; and affects the emotions and behaviors of the people affected. There are estimated 322 million people suffering from depression, equal to 4.4 per cent of the global population, according to the WHO report.Nearly half of the people at risk live in the area of South-East Asia (27%) and the West Pacific (27%), like China and India. Depression remains undiagnosed in many countries and is left without adequate treatment that can lead to serious self-perception and, at its worst, suicide. The social stigma surrounding depression often prevents many affected individuals from finding sufficient medical support. As a result they are moving to less structured outlets like social media.Via online forums, micro-blogs or tweets, people have started to share their experiences and struggles with mental health disorders.Their online activities inspired many researchers to introduce new forms of potential solutions to health care and methods for early detection of depression. Folowing Tasks we are using here: Machine Learning:Machine learning is an artificial intelligence (AI) application which provides systems ssswith the ability to learn and improve automatically from experience without being explicitly programmed. Machine learning focuses on computer programs being created that can access data and use it to learn for themselves.

a) Dataset :A data set is a data collection. We need a collection of data for training in Machine Learning projects. This is the real collection of data used to train the model for the specific actions.

b) Classification : In machine learning, classification is a supervised principle of learning that basically categorizes a collection of data into groups.

c) K-Means Algorithm:K-means clustering is one of the simplest and most common machine-learning algorithms.Clustering is the method of splitting data space or data points into a number of groups, making data points in the same groups more similar to other data points in the same group, and distinct from data points in other groups.Pycharm and Scikit-learn implementationUsing sklearn in Python needs only four lines to apply the algorithm: import the classifier, construct an instance, fit the training set data, and predict test set results:

II. RELATED WORK

According to the new estimates released by the World Health Organization WHO [1], the number of people living with depression increased by more than 18% between 2005 and 2015. More than 80% of this disease burden is among people living in low- and middle-income countries. This booklet provides latest available estimates of the prevalence of depression and other common mental disorders at the global and regional level, together with data concerning the consequences of these disorders in terms of lost health. For instance**,** According to, M. J. Friedrich[2], ‘‘Depression is the leading cause of disability around the world,’’ JAMA, vol. 317, no. 15, p. 1517, Apr. 2017. The proportion of the global population living with depression is estimated to be 322 million people—

4.4% of the world's population—according to a new report, "Depression and Other Common Mental Disorders: Global Health Estimates," released by the World Health Organization. The report also includes data on anxiety disorders, which affect more than 260 million people—3.6% of the global population. The prevalence of these common mental disorders is increasing, particularly in low- and middle-income countries, with many people experiencing both depression and anxiety disorders simultaneously
For instance, According to M. Nadeem [3], Social media has recently emerged as a premier method to disseminate information online. Through these online networks, tens of millions of individuals communicate their thoughts, personal experiences, and social ideals. For instance, According to S. Paul, S. K. Jandhyala, and T. Basu, [4] ''Early detection of signs of anorexia and depression over social media using effective machine learning frameworks,'' in Proc. CLEF, Aug. 2018, pp. 1–9. The CLEF eRisk 2018 challenge focuses on early detection of signs of depression or anorexia using posts or comments over social media. The eRisk lab has organized two tasks this year and released two different corpora for the individual tasks. The corpora are developed using the posts and comments over Reddit, a popular social media. The machine learning group at Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI), India has participated in this challenge and individually submitted five results to accomplish the objectives of these two tasks. The paper presents different machine learning techniques and analyze their performance for early risk prediction of anorexia or depression. The techniques involve various classifiers and feature engineering schemes. The simple bag of words model has been used to perform Ada boost, random forest, and logistic regression and support vector machine classifiers to identify documents related to anorexia or depression in the individual corpora. We have also extracted the terms related to anorexia or depression using met map, a tool to extract biomedical concepts. The experimental analysis on the training set shows that the Ada boost classifier using bag of words model outperforms the other methods for task1 and it achieves best score on the test set in terms of precision over all the runs in the challenge. Support vector machine classifier using bag of words model outperforms the other methods in terms of f-measure for task2. The results on the test set submitted to the challenge suggest that these framework achieve reasonably good performance. For instance, according to A. Benton, M. Mitchell, and D. Hovy. [5] (2017). ''Multi-task learning for mental health using social media text.'' We introduce initial groundwork for estimating suicide risk and mental health in a deep learning framework. By modeling multiple conditions, the system learns to make predictions about suicide risk and mental health at a low false positive rate. Conditions are modeled as tasks in a multi-task learning (MTL) framework, with gender prediction as an additional auxiliary task. We demonstrate the effectiveness of multi-task learning by comparison to a well-tuned single-task baseline with the same number of parameters. Our best MTL model predicts potential suicide attempt, as well as the presence of atypical mental health, with AUC > 0.8. We also find additional large improvements using multi-task learning on mental health tasks with limited training data.

III. SYSTEM ARCHITECTURE

A) EXISTING SYSTEM

In the existing system, the natural language processing techniques (NLP) and machine learning approaches are used to train the data and evaluate the efficiency of the system. The single bigram feature with support vector machine (SVM) is used to detect depression. The strength and effectiveness of the combined features (LIWC+LDA+bigram) are demonstrated with multilayer perceptron (MLP) classifier resulting in the top performance for depression detection reaching 91% accuracy and 0.93F1 scores.

B) PROPOSED SYSTEM

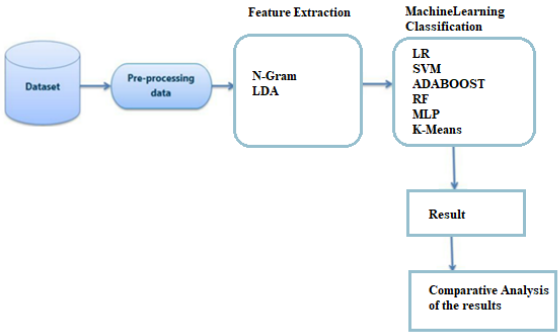

Fig. 1 Proposed System

The proposed framework as shown in fig.1 is developed by using Logistic Regression, Support Vector Machine, Random Fo rest, Adaptive Boosting and Multilayer Perceptron, K-Means classifier. Logistic Regression (LR) is a linear classification approach used to estimate the probability occurrence of binary response based on one or more predictors and features. Support Vector Machine (SVM) model is a representation of the examples as points in a highly dimensional space utilized for classification, where the points of the separate categories are widely divided. New examples are then mapped into the same space and predicted to belong to a category based on which side of the gap they fall [54]. Random Forest (RF) is an ensemble of decision tree classifiers trained with the bagging method where a combination of learning models increases the overall result. Adaptive Boosting (AdaBoost) is an ensemble technique that can combine many weak classifiers into one strong classifier [56]. It is widely used for binary class classification problems. Multilayer Perceptron (MLP) is a special case of the artificial neural network often used for modelling complex relationships between the input and output layer. Due to its multiple layers and non-linear activation it can distinguish the data that is not only non-linearly separable.

The present techniques rely on clinician's review of the patient in person. Those methods are subjective, done on interview and depend on reports by the patient. With rise in the depression, some automatic and reliable means of depression recognition is required. Efforts are being made in this direction to assess and detect depression through computer vision and machine learning. Hence in this paper we present our methods of text feature extraction, followed by decision level fusion which help identify depressed subjects with appreciable accuracy as evident in results obtained.

IV. METHODOLOGY

Information Retrieval, Information Extraction, Machine Translation, Text Simplification, Sentiment Analysis, Text Summarization, Auto-Predict, Auto-Correct, Speech Recognition, Question Answering Natural Language Generation. Here we used sentiment analysis. (Natural Language Toolkit)NLTK: NLTK is a common open source Python package. Instead of constructing all resources from scratch, NLTK does have all that NLP tasks. The best single feature is bigram with the Support Vector Machine (SVM) classifier for 80 percent accuracy and 0.80 F1 scores to detect depression. The proposed framework is developed by using Logistic Regression, Support Vector Machine, Random Forest, Adaptive Boosting and Multilayer Perceptron classifier. Logistic Regression (LR) is a linear classification approach used to estimate the probability occurrence of binary response based on one or more predictors and features. We used the clustering techniques k-mean to Cluster user data, and provide the user with accuracy stress levels. By using k-means algorithm, we achieve 99 percent accuracy and 0.99 F1 scores of detecting the depression. Using clustering of k-means, we developed four levels of depressive symptoms to match the classifications based on the norms. We then evaluated the validity of the classifications by contrasting them with the standard-based approach (and its variations) in terms of their robustness, model efficiency (precision, AUC, and sensitivity), and convergent construct validity (i.e. correlations with established correlates). The results showed that our automatic classification system worked well in comparison with the approach based on the standard.

1. Registration: The user have to register in to the social media system. In the registration phase the user will have to fill the details consisting in the registration phase. After registration the user can create his own tweet.
2. Data collection: Collection of user data from the Social media. It is not directly access the user posts on their social media page. In order to obtain the user data from social media page, we creates create tweet page through which the user can create their own tweets. All the information posted are stored in the analysis database.
3. Clustering: The posts from different user's arevr2collected together and separated by clustering techniques. The cluster comprises of sentiment based separation and classification k-mean algorithm, SVM, LR, Adaboost, RF have to use in this module.
4. Stress level prediction: Finding stress level of the user indifferent states. Find depression posts, Positive and Negative, like and dis-like posts.

V. EXPERIMENTAL RESULTS:

In the results, first we create a login for users to access the permission from admin. Once user is registered, they can tweet or re-tweet. The figures from fig 2 to fig 6 shows the execution of the proposed system, by classifying the tweet posts. Here we used the k-means algorithm for classifying the depression posts into positive posts, negative posts and depressed posts and also to achieve the 99 percent accuracy.

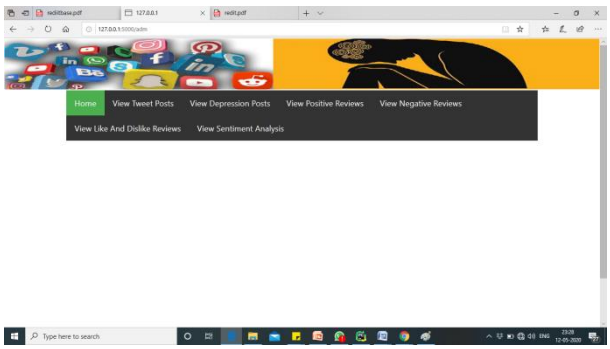

Fig 2: Main menu

Fig 3: Depression post using SVM classification

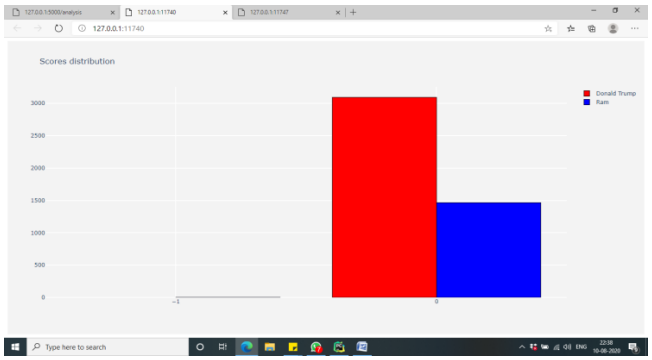

Fig4: scores distribution using k-means classification

Fig 5: sentiment distribution

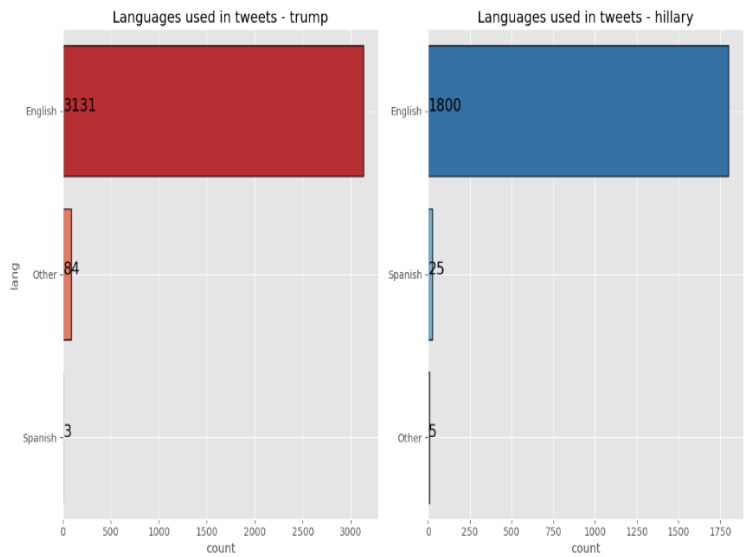

Fig 6: languages used in tweets

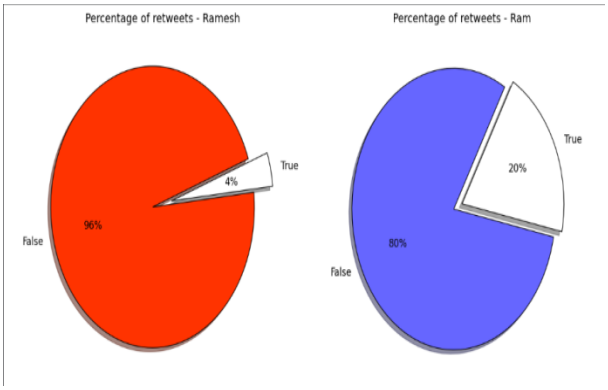

Fig 7: tweet percentage

In the table fig 8, we have achieved accuracy approximately for each algorithms applied. The maximum accuracy is achieved by applying the K-Means algorithm.

Data source	Methods	Accuracy (Approx.)
Reddit	SVM & AdaBoost	98%
Reddit	Random Forest	98 %
Reddit	Linear Regression& MLP	96 %
Reddit	K-Means	99 %

Fig 8: Table showing accuracy

CONCLUSION

In this paper, we have tried to identify the presence of depression in Reddit social media; and have been searching for affective performance enhancing depression detection solutions. The proposed framework is developed by using Logistic Regression, Support Vector Machine, Random Forest, Adaptive Boosting and Multilayer Perceptron, K-Means classifier. Logistic Regression (LR) is a linear classification approach used to estimate the probability occurrence of binary response based on one or more predictors and features. Support Vector Machine (SVM) model is a representation of the examples as points in a highly dimensional space utilized for classification, where the points of the separate categories are widely divided. New examples are then mapped into the same space and predicted to belong to a category based on which side of the gap they fall [54]. Random Forest (RF) is an ensemble of decision tree classifiers trained with the bagging method where a combination of learning models increases the overall result. Adaptive Boosting (AdaBoost) is an ensemble technique that can combine many weak classifiers into one strong classifier [56]. It is widely used for binary class classification problems. Multilayer Perceptron (MLP) is a special case of the artificial neural network often used for modelling complex relationships between the input and output layers. Due to its multiple layers and non-linear activation it can distinguish the data that is not only non-linearly separable.

FUTURE ENHANCEMENTS

In future work, fine grain emotion analysis can be done to detect anxiety indicators instead of using SentiWordNet which categorizes the words into three polarities. Further, the model can be tested on different user base: geographic, age, profession etc. Neuro-fuzzy and deep learning models can be explored for superlative prediction performance.

REFERENCES

[1] W. H. Organization. (2017). Depression and Other Common Mental Disorders: Global Health Estimates. Geneva: World Health Organization. [Online].Available: http://www.who.int/en/newsroom/fact-sheets/detail/depression
[2] M. J. Friedrich, ‘‘Depression is the leading cause of disability around the world,’’ JAMA, vol. 317, no. 15, p. 1517, Apr. 2017.
[3] M. Nadeem. (2016). ‘‘Identifying depression on twitter.’’ [Online]. Available: https://arxiv.org/abs/1607.07384
[4] S. Paul, S. K. Jandhyala, and T. Basu, ‘‘Early detection of signs of anorexia and depression over social media using effective machine learning frameworks,’’ in Proc. CLEF, Aug. 2018, pp. 1–9.
[5] A. Benton, M. Mitchell, and D. Hovy. (2017). ‘‘Multi-task learning for mental health using social media text.’’ [Online]. Available: https://arxiv.org/abs/1712.03538
[6] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, ‘‘From ADHD to SAD: Analyzing the language of mental health on twitter through self-reported diagnoses,’’ in Proc. 2nd Workshop Computer. Linguistics Clin. Psychol. Linguistic Signal Clin. Reality, 2015, pp. 1–10.
[7] D. Maupomés and M. Meurs, ‘‘Using topic extraction on social media content for the early detection of depression,’’ in Proc. CLEF (Working Notes), vol. 2125, Sep. 2018. [Online]. Available: https://CEUR-WS.org
[8] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, and J. Boyd-Graber, ‘‘Beyond LDA: Exploring supervised topic modeling for depression-related language in twitter,’’ in Proc. 2nd Workshop Computer. Linguistics Clin. Psychol. Linguistic Signal Clin. Reality, 2015, pp. 99–107

[9] D. Preotiuc-Pietro et al., ‘‘the role of personality, age, and gender in tweetingaboutmentalillness,’’inProc.2ndWorkshopComput.Linguistics Clin. Psychol. Linguistic Signal Clin. Reality, 2015, pp. 21–30.

[10] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, ‘‘Affective and content analysis of online depression communities,’’ IEEE Trans. Affect. Comput. vol. 5, no. 3, pp. 217–226, Jul. 2014.

[11]H.A.Schwartzetal. ‘‘towardsassessingchangesindegreeofdepression through Facebook,’’ in Proc. Workshop Comput. Linguistics Clin. Psychol. Linguistic Signal Clin. Reality, 2014, pp. 118–125. [12] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, ‘‘Recognizing depression from twitter activity,’’ in Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst., Apr. 2015, pp. 3187–3196.

[13] J. Wolohan, M. Hiraga, A. Mukherjee, Z. A. Sayyed, and M. Millard, ‘‘Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with nlp,’’ in Proc. 1st Int. Workshop Lang. Cognition Comput. Models, 2018, pp. 11–21.

[14] Y. Tyshchenko, ‘‘Depression and anxiety detection from blog posts data,’’ Nature Precis. Sci., Inst. Comput. Sci., Univ. Tartu, Tartu, Estonia, 2018.

[15]S.C.Guntuku, D.B.Yaden, M.L.Kern, L.H.Ungar, andJ.C.Eichstaedt, ‘‘Detecting depression and mental illness on social media: An integrative review,’’ Current Opinion Behav. Sci., vol. 18, pp. 43–49, Dec. 2017.

[16] R.A.Calvo, D.N.Milne, M.S.Hussain, and H.Christensen, ‘‘Natural language processing in mental health applications using non-clinical texts,’’ Natural Language Eng., vol. 23, no. 5, pp. 649–685, 2017.

[17] Á. Hernández-Castañeda and H. Calvo, ‘‘Deceptive text detection using continuous semantic space models,’’ Intel. Data Anal., vol. 21, no. 3, pp. 679–695, Jan. 2017.

[18] S. Freud, the Psychopathology of Everyday Life. London, U.K.: Hogarth, 1901.

[19] A. T. Beck, Depression: Clinical, Experim. Theoretical Aspects. Philadelphia, PA, USA: Univ. Pennsylvania Press, 1967.

[20] T. Pyszczynski and J. Greenberg, ‘‘Self-regulatory perseveration and the depressive self-focusing style: A self-awareness theory of reactive depression,’’ Psychol. Bull., vol. 102, no. 1, p. 122, Jul. 1987.

[21] E. Durkheim and A. Suicide, A Study in Sociology. Abingdon, U.K.: Routledge, 1952. [22] A. G. Reece, A. J. Reagan, K. L. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer, ‘‘Forecasting the onset and course of mental illness with twitter data,’’ Sci. Rep., vol. 7, no. 1, p. 13006, Oct. 2017. [23] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, ‘‘recognizing depression”

[22] W. S. Noble, ‘‘What is a support vector machine?,’’ Nature Biotechnol., vol. 24, no. 12, p. 1565, May 2006