

Location-Based Depression Analysis among the Various Cities of India

Kashaf Khan, Shailja Agarwal, Vijay Singh and Bhasker Pant
Department of Computer Science and Engineering
Graphic Era University, Dehradun, Uttarakhand, India

Abstract: Depression is one of the most common mental illness and it can affect the people of all ages. According to the WHO depression is the second leading cause of death among the 15-29 age group. More than 36% Indian are depressed due to various reasons and the current life style plays a vital role in it. In this study, we analyzed depression among the various cities of India and finds the people who lived in a metro cities are more depressed than the small cities using the data mining techniques. Compare the results with Naive Bayes and random forest finds that the users of the metros are more depressed than the naturally rich cities like Shimla and Thiruvananthapuram.

Key words: Data mining, depression, Naive Bayes, random forest, illness, WHO

INTRODUCTION

Depression is serious mood disorder. It affects the person's daily life what they feel, how they react on daily activity, sleeping, eating or working. Depression makes everything tougher and less meaningful. The theme of the world health day April 7, 2017 is "depression" and this mental disorder affects one in every 20 people in India and one of the major cause of suicide. Globally around 322 million people are under the condition of with this mental disorder depression. According to WHO depression is ranked as the single largest contributor to global disability that is around 7.5%. The major problem associated with depression is social reasons and lack of awareness, generally people do not want to share their condition with the other person that is also the major symptoms of depression as well. In the initial level, general counselor can rectify the problem but in the severe level specialized psychiatrists are required. In this study we find the cities where users are more depressed.

Naive Bayes and random forest: In this study introduction of Naive Bayes and random forest are given for better understanding of the study.

Naive Bayes: Naive Bayes is a Bayesian classifier, it uses supervised learning approach for predicting class membership probabilities (Swetapadma and Yadav, 2016). The assumption in the Naive Bayes is

that the features are the independent given in the class and only have a parent class (Taheri *et al.*, 2011) shown in Eq. 1:

$$p(C_j/d) = \frac{p(d/C_j)p(C_j)}{p(d)} \quad (1)$$

Where:

$p(C_j/d)$ = Probability of instance d being in class C_j

$p(d/C_j)$ = Probability of generating instance d given class C_j

$p(C_j)$ = Probability of occurrence of class C_j

$p(d)$ = Probability of instance d occurring

Random forest: The main purpose of the random forest is combining of trees for classification and regression (Guenther and Fritsch, 2010; Patel and Giri, 2016). Random forest is one of the most effectively and heavily used supervised learning approach and can perform classification and regression. Random forest is considering as the combination of tree predictors in such a way that each tree depends on the value of random vector sampled independently and with the distribution for all the tree in the forest (Breiman, 2001), basically random forest ensembles learning methods where a combination of weak models become to a powerful model. The main demerit of the random forest is that, we have very little control on what the model does generalized random forest is shown in Fig. 1.

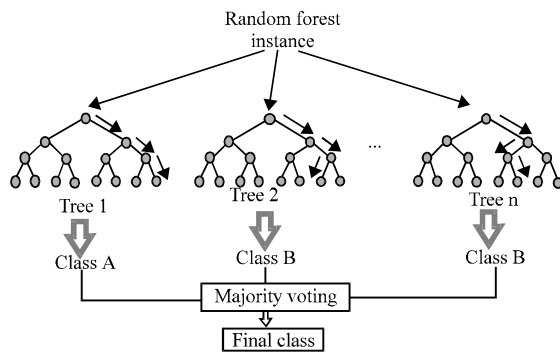


Fig. 1: Sample random forest

Literature review: Twitter is becoming interesting investigating medium of mass number of people and helps in analysing various kind of studies. Mental health is one of them. O’Dea *et al.* (2015) proposed a method of detecting suicidality on Twitter. They use human coders and machine classifier to differentiate the concern level among the suicide related Tweets. Epps (2014) proposed an automatic assessment of depression from speech and behavioral signals, he used a method for mitigating unwanted variability in speech. De Choudhury (2014) focused on the tools for identifying the depression disorder for the use of healthcare agencies. His investigation reveals that social media can be used for effectively identifying the mental status or condition of the user. Suhara and Pentland (2017) used the recurrent neural networks on the self-reported histories of the user and they find that the accuracy of the self-reported histories is higher than the support vector machine. They envisaged that if the long-term history is available of the user, then it is comparatively easy to predict the depressed condition of the user. Wahed and Hassan (2017) performed a study based on questionnaire among the medical students of the university. Their finding shows that higher stress and anxiety which is the early stage of the depression are strongly associated with female, older age or obese people. Orsal *et al.* (2013) did a study on internet addiction and depression among the university students and they find a significant positive correlation between the level of internet addiction and depression. Wolters *et al.* (2014) developed a system “Help4Mood” to support the diagnosis and treatment of Major depressive disorder. The system was developed with the help of specialized people in this domain and the person affected with depression. Wilson *et al.* (2014) used the microblogging services to detect the depression among the users. They used the linguistic approach

using the various word classes and find the mean and standard deviation against depression Tweets and non depression Tweets, their major contribution is linguistic based comparison between depression related Tweets and other Tweets. Gulhane *et al.* (2011) used the fuzzy logic for the stress analysis. They used the various speech features like loudness, short time energy, fundamental frequency, zero crossing rate etc. to detect the stress. Joshi (2012) proposed a multimodal approach for depression analysis for this she considered various features like the body gestures, ECG measures and Speech characteristics to detect the depressive disorder. The main finding of her work is multimodal approach improves performance over the single modality approaches. Dibeklioglu *et al.* (2015) used the multimodal approach to detect the depression in the clinical interviews or self-report rating. They used logistic regression on various movement dynamics and detect that facial movement dynamics was higher than the head movement dynamics during depression and they, envisaged that multimodal measures afford most effective detection. Scherer *et al.* (2014) worked on acoustic features and the vocal timings to detect the depression level among the interviewee and the finding suggests that the interviewers modify their acoustic feature in response to the depression severity.

MATERIALS AND METHODS

Proposed algorithm: To understand the depression among the users of various cities of India, we decided to collect location-based Tweets from the four different cities named Bangalore, New Delhi, Shimla and Thiruvananthapuram. Twitter is one of the effective medium to share the feelings and further these emotions can be used to various dimensional analysis (Fig. 2).

In the first step, download the location specific Tweets using the python library shown in. The downloaded Tweets contain lots of unnecessary or unwanted information or simple words downloaded tweets contains noisy data.

Then apply noise removal techniques, resulting clear and reduced data. Then remove the stop words which doesn’t contain itself any meaning using the “tm” text mining package of R. This will result the more informative dataset of Tweets. Thereafter, we divide the dataset in ratio of 80% for training and 20% for testing and applied Naive Bayes and random forest algorithms.

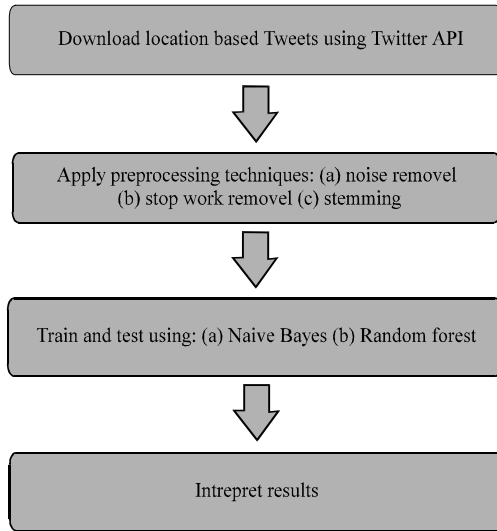


Fig. 2: Main steps

Sample Tweets:

- I feel angry, lonely, worthless and hopeless. I hate my job and everything
- I feel so pressurized, hopeless and sad
- I feel so depressed, don't want to wake up next morning
- Life is good when you have a wonderful company
- Walk in my shoes. Feel what I feel. Surely u will be suicidal useless and hopeless
- Feels good always
- First time I feel, so hopeless, so lonely, so uncared, so unloved, so angry so sad, so suicidal
- Feeling positive, energetic non-depressed
- I feel so dump, lonely and worthless. I don't know what to do anymore
- Distress-extreme anxiety, sorrow or pain. That's what I feel right now
- Fun night out don't want to wake up early next morning

RESULTS AND DISCUSSION

We consider for different random sets of each city and applied the Naive Bayes and random forest, results of both the algorithms are almost similar and envisages that the users of New Delhi and Bangalore are more depressed than the cities Shimla and Thiruvananthapuram. Figure 3 shows the results using the Naive Bayes and Fig. 4 shows the results using the random forest. In the Fig. 5 combined result is shown.

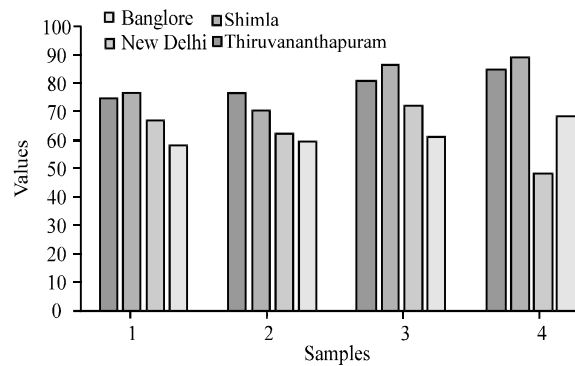


Fig. 3: Using Naive Bayes

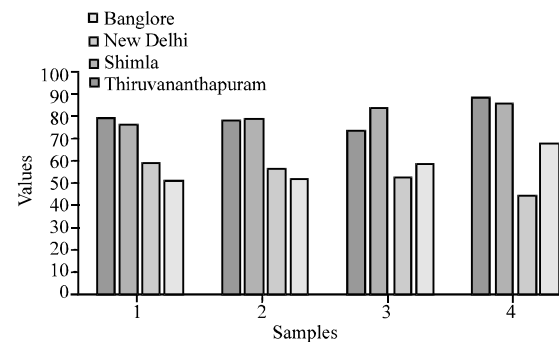


Fig. 4: Using random forest

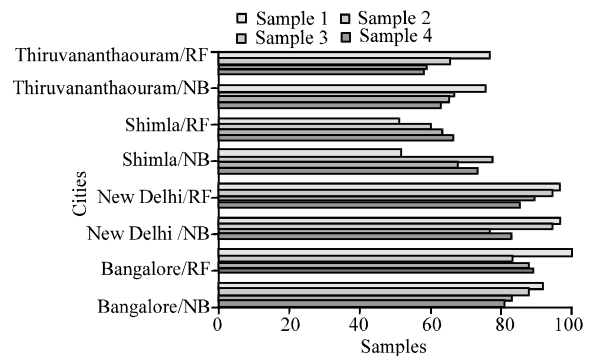


Fig. 5: Naive Bayes vs. random forest

CONCLUSION

The study confirms that the Twitter can be used to express the depression and Naive Bayes and random Forest both are effective technique to classify the textual contents and provide the promising results. The extracted results indicated that the metros cities are more prone to depression rather than the naturally rich cities. In the future work, we would implement more classification algorithms on the more diverse set of cities.

REFERENCES

- Breiman, L., 2001. Random forests. *Mach. Learn.*, 45: 5-32.
- De Choudhury, M., 2014. Can social media help us reason about mental health?. *Proceedings of the 23rd International Conference on World Wide Web*, April 07-11, 2014, ACM, Seoul, Korea, ISBN:978-1-4503-2745-9, pp: 1243-1244.
- Dibeklioglu, H., Z. Hammal, Y. Yang and J.F. Cohn, 2015. Multimodal detection of depression in clinical interviews. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, November 09-13, 2015, ACM, Seattle, Washington, USA., ISBN:978-1-4503-3912-4, pp: 307-310.
- Epps, J., 2014. Automatic assessment of depression from speech and behavioural signals. *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, November 7, 2014, ACM, Orlando, Florida, USA., ISBN:978-1-4503-3119-7, pp: 1-1.
- Gulhane, Y.H., S.V. Rode and S.A. Ladhake, 2011. Application of fuzzy logic in stress analysis. *Proceedings of the International Conference and Workshop on Emerging Trends in Technology*, February 25-26, 2011, ACM, Mumbai, India, ISBN:978-1-4503-0449-8, pp: 679-685.
- Gunther, F. and S. Fritsch, 2010. Neuralnet: Training of neural networks. *R. J.*, 2: 30-38.
- Joshi, J., 2012. Depression analysis: A multimodal approach. *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, October 22-26, 2012, ACM, Santa Monica, California, USA., ISBN:978-1-4503-1467-1, pp: 321-324.
- O'Dea, B., S. Wan, P.J. Batterham, A.L. Caley and C. Paris et al., 2015. Detecting suicidality on Twitter. *Internet Interventions*, 2: 183-188.
- Orsal, O., O. Orsal, A. Unsal and S.S. Ozalp, 2013. Evaluation of internet addiction and depression among university students. *Proc. Soc. Behav. Sci.*, 82: 445-454.
- Patel, R.K. and V.K. Giri, 2016. Feature selection and classification of mechanical fault of an induction motor using random forest classifier. *Perspect. Sci.*, 8: 334-337.
- Scherer, S., Z. Hammal, Y. Yang, L.P. Morency and J.F. Cohn, 2014. Dyadic behavior analysis in depression severity assessment interviews. *Proceedings of the 16th International Conference on Multimodal Interaction*, November 12-16, 2014, ACM, Istanbul, Turkey, ISBN:978-1-4503-2885-2, pp: 112-119.
- Suhara, Y., Y. Xu and A.S. Pentland, 2017. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. *Proceedings of the 26th International Conference on World Wide Web*, April 03-07, 2017, ACM, Perth, Australia, ISBN:978-1-4503-4913-0, pp: 715-724.
- Swetapadma, A. and A. Yadav, 2016. Protection of parallel transmission lines including inter-circuit faults using Naive Bayes classifier. *Alexandria Eng. J.*, 55: 1411-1419.
- Taheri, S., M. Mammadov and A.M. Bagirov, 2011. Improving naive Bayes classifier using conditional probabilities. *Proceedings of the 9th Conference on Australasian Data Mining Vol. 121*, December 01-02, 2011, ACM, Ballarat, Australia, ISBN:978-1-921770-02-9, pp: 63-68.
- Wahed, W.Y.A. and S.K. Hassan, 2017. Prevalence and associated factors of stress, anxiety and depression among medical Fayoum University students. *Alexandria J. Med.*, 53: 77-84.
- Wilson, M.L., S. Ali and M.F. Valstar, 2014. Finding information about mental health in microblogging platforms: A case study of depression. *Proceedings of the 5th International Symposium on Information Interaction in Context*, August 26-30, 2014, ACM, Regensburg, Germany, ISBN:978-1-4503-2976-7, pp: 8-17.
- Wolters, M.K., A.S. Tatar, S. Matu, R. Moldovan and D. David et al., 2014. eHealth support for people with depression in the community: A case study series. *Proceedings of the Conference on HCI Korea*, December 10-12, 2014, Hanbit Publishing Network, Seoul, Korea, ISBN:978-89-6848-752-1, pp: 138-144.