Depression Detection on Social Media using Machine Learning Techniques: A Survey

Suyash Dabhane[1], Prof. Pramila M. Chawan[2]

[1]M. Tech Student, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India
[2]Associate Professor, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

---***---

Abstract - *Depression is a common but serious mental health disorder. Still, most people dealing with depression do not approach doctors for this problem. On the other hand, the use of Social Media Sites like Twitter is expanding extremely fast. Nowadays, people tend to rely on these social media applications to share their emotions and feelings. Thus, this readily available content has become helpful for us to analyze the mental health of such users. We can apply various machine learning techniques on social media data to extract the mental health status of a user focusing on Depression. Detecting texts that express negativity in the data is one of the best ways to detect depression. In this paper, this problem of depression detection on social media and various machine learning algorithms that can be used to detect depression have been discussed. The Ensemble Learning approach for solving this problem has been enlightened. We aim to find and implement the most appropriate approach and algorithm to solve this problem.*

***Key Words*: Depression Detection, Natural Language Processing, Machine Learning, Ensemble Learning, Twitter, Social Media**

1. INTRODUCTION

Depression is a dysfunctional behavior that can influence anybody regardless of old enough, gender, status, and so forth. It extremely brunt a person's life affecting what they think about themself, their sleeping cycle, eating cycle, etc. It is the worst state of a person's mind when they feel sad and loses interest in nearly doing every productive thing and they can't simply move from that state. Factors like Social, Biological, and psychological factors are responsible for causing depression. Depression also causes other physical illnesses. Self-destruction is the subsequent reason for death in 15-29-year-olds due to depression.

Using a machine learning approach to detect depression will surely help social media users for detecting and predicting depression risk. It additionally helps social media users to seek early help to overcome depression. A machine learning approach like supervised learning can analyze and build a model on social media posts like Twitter or Reddit posts. There are many factors like users' posts, tweets, replies, post time, emotions, etc. which can contribute to detecting depression. To classify the data or tweets are depressive or not we will use machine learning algorithms and natural language processing. Before making a model, we will do exploratory data analysis on our dataset to thoroughly understand it.

In this paper, I have discussed different techniques that can be used to detect depression from social media data for e.g. Twitter data and how these techniques differ from each other and so on.

1.1 Supervised Machine Learning

Machine learning algorithms especially are of two types supervised and unsupervised machine learning algorithms. The supervised machine learning model gets trained on the labeled dataset. The dataset which has input as well as a label is called a labeled dataset. Classification and regression are two types of supervised machine learning algorithms. As the dataset for this project is labeled dataset, we will review supervised classification algorithms.

1.2 Naïve Bayes Classifier

Naive Bayes assumes that the presence of one feature is independent of other features. The model is easy to build. A large dataset can be trained using naive Bayes. For making real-time predictions we can use the Naive Bayes algorithm as it is fast.

1.3 Support Vector Machine

SVM algorithm plots a dataset on 2-dimensional space. Depending on data there are linear and non-linear separators. For non-linear separators, we need to use kernel. Proper kernel value should be chosen. The hyperplane is drawn which separates different classes of datasets. The data points which are present nearer to the hyperplane are called a super vector. With the help of a maximum marginal classifier, we can easily explain the support vector machine. Linear kernel, polynomial kernel, and radial kernel are 3 types of kernel used in SVM.

2. LITERATURE REVIEW

In machine learning, there are supervised machine learning classification algorithms for example Support Vector Machine, K Nearest Neighbor, Naive Bayes, Decision Tree, etc. Based on the different tasks and available data we can use those algorithms. The accuracy of the model can be improved using ensemble learning methods.

There are different methods in machine learning which can be used to detect depression from posts of users. But each of them has its pros and cons which may affect a system in terms of accuracy and efficiency.

There are simple ensemble methods like max voting, averaging, weighted averaging. But here we will use advanced ensemble techniques. We will first experiment with different ensemble learning models like bagging, boosting and tagging.

2.1 Classification using Naïve Bayes and SVM

Each classification algorithm with its pros and cons. For depression classification or detection, no method can be considered perfect. Depending on data in some cases Support Vector Machine is giving better accuracy and in other cases Multinomial Naive Bayes giving better accuracy.

The supervised classification algorithms don't perform at the human level hence give a limited performance. A voting model with several classifiers can be used. With the help of this voting model, we can select features that are giving the maximum number of votes. But this method is not effective. For the model to be perfect we should have bias and variance in balance. One way to increase the performance of machine learning models is to ensemble them.

Consequently, different further research is expected to concentrate on issues of improving the accuracy of the model to make accurate depression detection.

3. PROPOSED SYSTEM

3.1 Problem Statement

"To analyze and detect depression of social media data of users like Twitter feed by using machine learning techniques."

3.2 Problem Elaboration

Depression is a leading cause of mental ill-health, which has been found to increase the risk of early death. However, 70% of the patients would not consult doctors at a stage of depression. Meanwhile, people increasingly rely on social media for sharing emotions, and daily life activities thus helpful for detecting their mental health. We aim to apply Machine Learning Techniques on Social Data of a user like Twitter feeds for performing analysis focusing on depression detection.

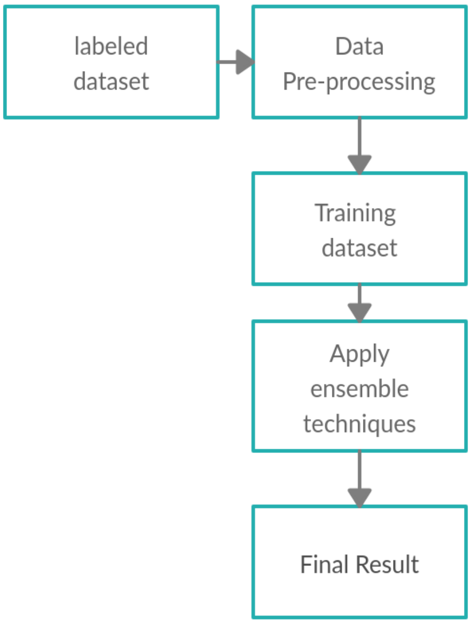
3.3 Proposed Methodology

Machine learning supervised algorithms like naive Bayes, support vector machine, K nearest neighbor, decision tree and logistic regression, etc. As studied, we observed that naive Bayes and support vector machine algorithms are giving better performance than other models. But the accuracy which we are getting is not good. The accuracy which was obtained is stuck between 80 to 85 %. So, to improve the performance of depression detection systems we need to use other powerful algorithms. The traditional machine learning algorithm follows steps like data collection, preprocessing, model selection, model training, evaluation, parameter tuning, and prediction. With the help of ensemble learning, we combined different individual models to get the improved and powerful final model.

In an existing system given in this paper [3], only an individual classifier is used as given in the training diagram. A multinomial naive Bayes classifier is used and gives 83 % accuracy. We will use different individual classifiers and then we will apply ensemble methods with some modification to get accurate results.

1. **Data Collection: -** We will collect the data of Twitter posts using Twitter API. The dataset is in the JSON format. We want the dataset in a CSV file for that we will convert the JSON file into a CSV file. AS our dataset needs a label, we will add them manually.

2. **Data Preprocessing: -** Data preprocessing involves steps like tokenization, stemming, lemmatization, stop word removal, PoS tagging, and finally, we convert the text data in vector format using TF-IDF or Bag of Words.



3. **Training: -** Before training a model, we will split our dataset into two sets, training set, and testing set. As we are going to use ensemble learning techniques, we will use different classification algorithms as base learners

with three types of ensemble methods. Ensemble learning techniques include stacking, bagging, boosting.

In stacking, we train each base learner (different machine learning algorithms) with a complete training dataset. We use meta learners to combine the results of all base learners. The output of base learner models is given to the training meta learner. In bagging, we divide a dataset into several times base learners. We give this sub-dataset to each base learner and then we take the mean of all predictions obtained from base learners. With the help of bagging, we can reduce a variance. In boosting, there are strong learners and weak learners. With the help of boosting we can decrease bias error. Boosting converts weak learners to strong learners. It is a continuous process of building models. In boosting, base learners are trained using an iterative process. The disadvantage of boosting includes it sometimes overfits the data.

In the end, we will select the best method which will give better accuracy.

4. **Testing: -** We will test our model on testing data and we will evaluate the performance of our model using evaluation metrics like accuracy, precision, recall, F1-score, etc.

4. CONCLUSION

In this paper, we have introduced an ensemble learning method that we are going to experiment on existing methods of depression detection. After that, we will get the model with more improved accuracy which can detect depression more accurately. Existing methods were using classification algorithms like multinomial naive Bayes and support vector machines which are giving less accurate prediction of a system. With the help of this proposed method, we will try to build a more efficient and accurate model to detect depression on social media data.

REFERENCES

[1] A. Noureen, U. Qamar and M. Ali, "Semantic analysis of social media and associated psychotic behavior," 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Guilin, 2017, pp. 1621-1630, doi: 10.1109/FSKD.2017.8393009.

[2] P. Arora and P. Arora, "Mining Twitter Data for Depression Detection," 2019 International Conference on Signal Processing and Communication (ICSC), NOIDA, India, 2019, pp. 186-189, doi: 10.1109/ICSC45622.2019.8938353.

[3] M. Deshpande and V. Rao, "Depression detection using emotion artificial intelligence," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, 2017, pp. 858-862, doi: 10.1109/ISS1.2017.8389299.

[4] P. Gupta and B. Kaushik, "Suicidal Tendency on Social Media: A Case Study," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 273-276, doi: 10.1109/COMITCon.2019.8862236.

[5] Dfgdf Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, & Wenwu Zhu (2017). Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17 (pp. 3838–3844).

[6] N. A. Asad, M. A. Mahmud Pranto, S. Afreen, and M. M. Islam, "Depression Detection by Analyzing Social Media Posts of User," 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), Dhaka, Bangladesh, 2019, pp. 13-17, doi: 10.1109/SPICSCON48833.2019.9065101.

[7] S. Jain, S. P. Narayan, R. K. Dewang, U. Bhartiya, N. Meena, and V. Kumar, "A Machine Learning based Depression Analysis and Suicidal Ideation Detection System using Questionnaires and Twitter," 2019 IEEE Students Conference on Engineering and Systems (SCES), Allahabad, India, 2019, pp. 1-6, doi: 10.1109/SCES46477.2019.8977211.

[8] O. Obulesu, M. Mahendra and M. ThrilokReddy, "Machine Learning Techniques and Tools: A Survey," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, 2018, pp. 605-611, doi: 10.1109/ICIRCA.2018.8597302.

[9] T. N. Rincy and R. Gupta, "Ensemble Learning Techniques and its Efficiency in Machine Learning: A Survey," 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, 2020, pp. 1-6, doi: 10.1109/IDEA49133.2020.9170675.

[10] Dong, X., Yu, Z., Cao, W. et al. A survey on ensemble learning. Front. Comput. Sci. 14, 241–258 (2020). https://doi.org/10.1007/s11704-019-8208-z

[11] https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models

[12] https://www.who.int/news-room/fact-sheets/detail/depression

BIOGRAPHIES

Suyash Dabhane,
MTech Computer Engineering
VJTI, Mumbai.

Prof. Pramila M. Chawan,
Associate Professor,
Department of CE and IT,
VJTI, Mumbai.