International Journal of Research in Engineering, Science and Management
Volume-2, Issue-5, May-2019
www.ijresm.com | ISSN (Online): 2581-5792

Predicting Depression Level using Social Media Sites

Shivani Kale[1], Pravin Borate[2], M. K. Nivangune[3]
[1,2]*Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India*
[3]*Professor, Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, India*

***Abstract*: Depression is the leading cause of many health issues and suicide attempts. There are plentiful reasons which cause depression, but it can be cured through early consultation of a psychologist. As it is observed that many of the patients do not consult the doctor at early stage they suffer from severe stage of depression. The other problem that can occur in such situations is the False Positive answers by the patients. These problems can be solved at a specific rate using the social media to get the depression rate. Social media is in trend for this generation. People usually tend to share their emotions and feelings using such mediums. These posts can be used to analyze the mental state of people. In this project we have used Natural Language Processing (NLP) and Machine Learning (ML) terminologies to solve such problems.**

***Keywords*: Depression, Prediction, Social Media Sites, Machine Learning, Python, Django 2**

1. Introduction

About 300 million people across the world are suffering from depression. According to analysis of Bradley Hospital, suicide attempts can be as high as 5 lakhs per year and World Health Organization has ranked suicide as the 2nd major cause for death, and it is observed that the most common reason behind this is depression. Depression at the early stage is more easily cured and would not cause any side-effect but as the severity increases the treatment becomes critical and this can deteriorate the patient mentally as well as physically. This makes depression the most devastating disease spreading around the globe.

Depression has the major contribution in the cause of diseases like cancer and heart attack. According to WHO 75% of individuals diagnosed from eating disorders are suffering from depression [2]. An individual suffering from any mental disease suffers from the snowball effect which in turn ends with Suicide attempts. 70% of the time patients do not consult a doctor at the early stage of depression which results in deterioration of their condition [1]. According to WHO two third of the depressed people do not get proper treatment as the condition is not encountered at early stage. However current methods to treat the depressed people are considered inadequate. Only 87% of the governments worldwide provide primary treatments for the illness [2].

In this modern world the tool which is used in a greater extent is the Social Networking Sites (SNS). People tend to open up and share their feelings with the help of such websites by posting blogs, photos and exchanging messages. According to the analysis carried out in the year 2018, Facebook has 2.19 billion active users and, Twitter counts up to 336 million users. Many researchers have been proving that social media can be effectively used to uncover the depression states of the people and maintain people’s mental health. Uncovering this needs high level of computation, data mining and analysis of the data which can be gained by scrapping the daily posts of the users. Classification of such posts depends on many factors like depressed people usually use 79% more negative use in their posts as compared to the normal posts and such posts mostly increase in the duration of 11pm to 5am [1].

Incredible research has already been done to solve this problem and some solutions have also been uncovered but these are mostly for the research purpose. We on the other side have built a website which is easily accessible to the end users and can help the doctors understand the mental state of the patients. The system takes data from Twitter bounded to an individual which can be gained simply by entering the twitted id of the user. The analysis does not consider all the parameters affecting depression due to which the final result cannot be considered 100% accurate. Depending on the different classification modules the rate of accuracy changes, to get the final result we consider the classification module which gives more accuracy.

2. Method

The predicting model is created using four different classifiers namely Support Vector Machine (SVM), Naive Bayes, KNN classifier and Logistic Regression. Natural Language Processing is used for pre-processing of data. It is carried out using Natural Language Toolkit (NLTK). Using NLTK the data is tokenized and is labelled either one of these: ‘negative’, ‘positive’ or ‘neutral’.

The front end is developed using Django 2. It is secure, fast, flexible and most popular but also the most expensive among all of the python frameworks [6]. Other than analyzing the depression using twitter the website provides add on features like registration of doctors, patients and self-analysis through questionnaire.

After getting the twitter id from front end, the data is processed at the backend using python. The posts are scrapped and saved in a CSV file. Latest 100 posts of the user are scrapped the number of posts can be increased according to the requirements and to increase the accuracy of the output. The data is then processed and classified using the four classifiers. The result of the classifier is then compared and the one having highest accuracy rate is displayed as the final result.

The data is divided in two parts while classification 80% of the data is given to the training phase and 20% is used for testing purpose. The accuracy of the classification is depending on the training data hence it is important to consider the edge cases of all the classes and not only the obvious ones [3].

3. Related Work

A. Pre- processing

Pre-processing of data is done transform the raw unstructured data into a clean data set and label the data. Natural Language Processing (NLP) terminologies are used for transforming the unstructured data in a meaningful format. NLP helps algorithmic system to integrate natural language understanding and language generation [7]. Tokenization is carried out to chop the scrapped posts into individual words called tokens and then these tokens are classified based on the pre-defined dataset and labelled accordingly.

B. Front-end

Django 2 framework of Python is used to develop the front end. Django has many merits over other web development frameworks. With the help of Django it is easy to develop every model independently with the help of Django apps. With transparent and clean code, the development of website becomes both efficient and effective [6].

4. Algorithms

A. Support Vector Machine (SVM)

Support Vector Machine constructs a hyperplane between the high-dimensional space to classify the emotions in two categories (positive or negative) [2]. We here use a simple SVM algorithm which utilizes a straight line between negative data points and positive data points.

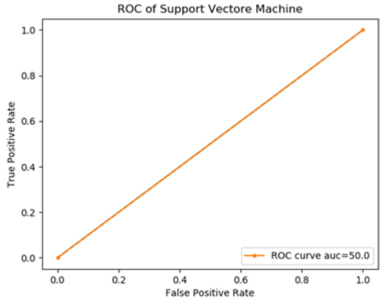

Fig. 1. ROC of support vector machine

B. Naive Bayes

In Naive Bayes every feature or class contributes independently which is hard to achieve in real life. To overcome this Multinomial Naive Bayes provided by Scikit-Learn is implemented [5].

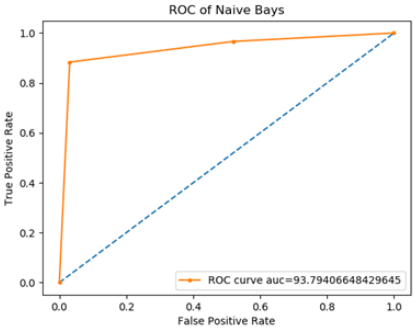

Fig. 2. ROC of Naïve Bays

C. Logistic Regression

Similar to the SVM algorithm the Logistic regression algorithm constructs a regression line between the two classes or dimensions of the data but here the line of regression is not linear and moves towards the extreme more rapidly. This makes Logistic Regression stronger and elaborated, but these predictions can turn out to be wrong.

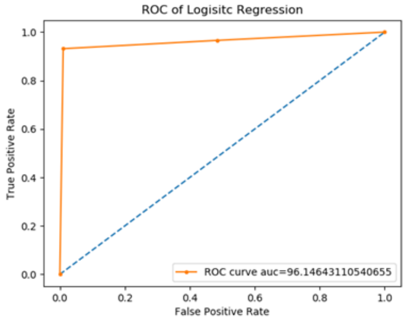

Fig. 3. ROC of logistic regression

D. KNN classifier

The classification of the data is based on the classes of its neighbours. The classification of data is directly based on the training examples which need to be executed at runtime due to which this is also called as example based classification. The value of "k" is the no of data elements considered as neighbours this can vary for model to model. The value of K is selected randomly and the one giving highest accuracy is selected. In our case the value 2 gives the highest accuracy rate.

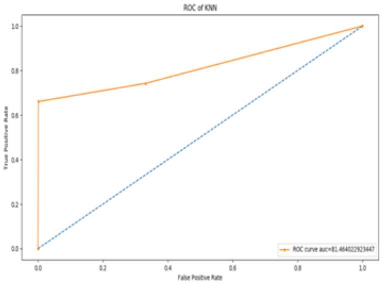

Fig. 4. ROC of KNN

E. Results

Table 1
Classification reports of every model;
bolded values demonstrate the highest values.

Classifiers	SVM	Naïve Bayes	Logistic Regression	KNN
Accuracy	50%	93.79%	**96.14%**	81.46 %
Precision	38.69%	90.35%	**95.29%**	79.12%
Completion Time	35.8s	**4.29s**	12.10s	5.85s

5. Conclusion

We have worked to utilize the ability of data from twitter to analyze the depression state of the users and make it easy to predict the condition of the patient. Other than social media analysis we have provided extra feature of depression questionnaire for self-testing.

Our aim was to establish a system that can help the doctors and family of the depressed people to predict the mental state of the patient easily and reduce the effect of false positive on the result of the analysis. We were also working to find the best fitted algorithm for the classification of emotion. According to our analysis Logistic Regression give better accuracy than any other model examined.

6. Future Scope

With the enhancement in the system, we look forward to understand how other models work and we hope to understand how deep learning will help enhance the system.

We hope to work with more amounts of data and understand its effect on the end model with the enhancement in the accuracy rate. We look forward to include many more feature of data which can help achieve more accurate results.

References

[1] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua and Wenwu Zhu, “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).

[2] Moin Nadeem, “Identifying Depression on Twitter,” in Social and Information Networks, 2016.

[3] M. M. Aldarwish and H. F. Ahmad, "Predicting Depression Levels Using Social Media Posts," *2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS)*, Bangkok, 2017, pp. 277-280.

[4] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, Johannes C Eichstaedt, “Detecting depression and mental illness on social media: an integrative review,” Current Opinion in Behavioral Sciences, Volume 18, 2017, Pages 43-49.

[5] O`scar Romero Llombart, “Using Machine Learning Techniques for Sentiment Analysis,” final project on computer engineering, school of engineering (EE), universitat auto` noma de barcelona (UAB), June 2017.

[6] B. Nithya Ramesh, Aashay R. Amballi, Vivekananda Mahanta, “Django the python framework,” in International Journal of Computer Science and Information Technology Research, Vol. 6, Issue 2, pp. 59-63, April - June 2018.

[7] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh, “Natural Language Processing: State of The Art, Current Trends and Challenges.”