ANALYZING SOCIAL BEHAVIOUR USING NLP

Shrey Garg
Computer Science & Engineering.
Amity University, Noida.

Poorva Sharma
Computer Science & Engineering,
Amity University, Noida.

One of the major reasons behind mental ill health in recent times has been depression. Depression also contributes towards the rising number of suicide rates all over the world and also leads to the impairment of daily life. Researchers and doctors have been trying to find out ways to detect depression in people in its early stages itself so that the person can be prevented from the more dangerous outcomes of depression in the later stages such as the danger of committing suicide and harming himself. In recent times social media has come to the rescue by emerging as a platform to annunciate information online. These online networks are used by legions of social media influencers to convey their personal experiences, thoughts and social ideals. Therefore, through this research paper we take a look at the future of social media to forecast, even prior to onset, depression in online personas. People suffering from depression have shown a tendency to communicate via online platforms, where they are not accountable to anyone for what they tweet or post, rather than discussing their problems with their families and friends. Also, previous studies have shown that depression also affects the language usage. In this paper, Naive-Bayes classifier has been used on twitter feeds for conducting emotion analysis focussing on depression. The results have been presented using the primary classification metrics including F1-score, accuracy and confusion matrix.

Keywords—Emotion Artificial Intelligence, Support Vector Machine, Naive Bayes, Depression Detection, Machine Learning, Natural Language Processing.

I. INTRODUCTION

For the developing world, mental health is and continues to be a prominent concern. The stresses of daily life events such as the stress of job, stress of competition in various fields, genes and family history further aggravate the problem. [1] Not only do nearly 300 million people worldwide suffer from clinical depression, but the probability for an individual to encounter a major depressive episode within a period of one year is 3 – 5% for males and 8 – 10% for females. Anyone suffering from depression is highly probable to get trapped in the vicious cycle of gloominess and ends up committing suicide.

Studies have shown that people suffering from depression are inclined towards expressing their experiences, opinions, emotions via a variety of social media platforms like LinkedIn, Facebook, Twitter, Instagram. They use various means to express such as through posting photos, videos, memes and mainly through text. This data which is being posted and texted daily by millions of users of various social media platforms can be put to use by undertaking explorative analysis. [1] Textual data is the best choice for doing analysis as it is the most widely used form of conveying things to others, it is easy to handle, quick to pre-process, can be quantized and has comparatively smaller memory storage size than videos and images. Further, Twitter proves to be the best among other social media platforms for applying emotion artificial intelligence as it has a limit on the amount of characters allowed in a single Tweet.

Machine learning algorithms and techniques can be used for the detection of emotion. Emotion artificial intelligence can further be enhanced to identify mental illness in the starting stages, market research, and in opinion mining. It can also be evolved further to include images and videos as well for the facial expression detection. In the textual emotion detection techniques, there are two levels for carrying out the process, one is through the analysis of the entire document that is called coarse level and the other is through the analysis of each and every sentence of the document, which is called attribute level analysis. [2] Since in this paper the source of data is twitter, therefore, we have done sentence level emotion analysis.

The matter which this paper provides is as follows:

1. We have merged 3 datasets from different sites such as “Kaggle, Socrata, Quandl”. All the users in the dataset are anonymized for privacy purposes.
2. We then screened the different models which can be used to quantify a person’s social media behaviour across a maximum of ten thousand tweets.
3. We have then compared the different models and shown how the model which we have used gives more accurate results than other models.
4. Our model showed an accuracy of 86%

The research paper has been organised as follows:

Part II contains the aim of writing this research paper. Part III has a brief account of the data and classifiers used in the execution, Part IV shows the result and Part V contains the conclusion and future works. Lastly, Part VI has references.

LITERATURE REVIEW

To recognise depression and its symptoms, a lot of research has been performed in psychology, psychiatry, sociolinguistic fields and medicine. Proposals of many quiz-based systems have been made to detect depression. Examples of some of the depression detecting quizzes which ask the user several questions and then based on the answers given by the user, estimate the probability of them suffering from depression are BDI[3], CES-D[3]

Biological attributes have been found out for the detection of depression in early stages by Redei and others. Blood samples collected for 26 candidates in the age group of fifteen to nineteen for transcriptomic indicators out-turned in the right detection of eleven out of fourteen people suffering from depression. From medical point of view, this was the first remarkable approach. [4]

Recent study has been done on log data about the activities of an individual to find out depression. A method has been prepared by Resnik and others to analyse textual data that has been written by individuals suspected of depression. They conducted an experiment on the college students by analysing essays written by college students and then applying “Latent Dirichlet Allocation” which is a known model of Machine Learning for topic extraction. [5]

A Shared Task was developed by Coppersmith and others for the Clinical Psychological Conference. Coppersmith et al., via this shared task, dispensed the dataset containing data from depressed users to all the participants to standardize fundamental computational technologies. The work concluded by the contenders can be summarized as:

1. To come across a set of data that gave maximal effect to distinguish between the classes provided for a user, “UMD” managed to use a managed topic model methodology. Moreover, they amalgamated all the tweets sensibly from a single week into a document and used it.
2. Simple regression models were used by “WWBP team”. This approach had a variety of features such as binary unigrams and deducing topics automatically. [6]
3. A “Character Language Model” was used by “Microsoft-IHMC-Qntfy joint team” to find out how a trail of characters that is already given, is to be created by every classification class and how to provide a score for every string. This approach was better than others since it scored even the shortest texts and abbreviations.

We have built our research paper on the above-mentioned work and our study contributes towards: (i) Investigating the potential of social media in identifying people suffering from depression (ii) comparing different models to predict depression in early stages.

II. AIM

The aim of this study is to make a model which can give more accurate results for detecting depression in people using their tweets. The accuracy of the model can be re-checked using standard “Precision”, “Recall”, and “F1 scores”. [7]

The above terms can be defined as:

A. Precision

“It is the fraction of retrieved documents that are relevant to the query. In our circumstances, it answers the question: *How many of the users we identified as depressed are actually depressed?”*

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

B. Recall

“Recall is the probability that a relevant document is retrieved by the query. Within our situation, it answers the question *Out of all of the depressed users, how many did we properly detect?”*

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

C. F1 Score (F-Measure)

“An F1 score is the harmonic mean of Precision and Recall; it therefore is commonly utilized as a classification evaluation metric due to weighing each metric evenly.”

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

III. DATA

A. Dataset

In this study we have made our own dataset by merging datasets of Twitter from 3 different websites. This dataset consists of tweets of around ten thousand users of twitter with public accounts. The users whose tweets have been included in this research work have been anonymized for the purpose of privacy and the users with whom they interacted have also been anonymized. Twitter terms of service and its policy has been taken care of while collecting data from there. From these tweets we have made the training and test dataset. The percentage of data used for training dataset is 80% and the percentage used for creating the test dataset is 20%. Similarly, different word lists have been generated for both types of datasets. The wordlist used for training dataset comprises of selected list words showing the signs of depression in them such as words like “depressed”, “drugs”, “suicide” [8] and the wordlist for test dataset has random tweets consisting of both neutral and negative elements.

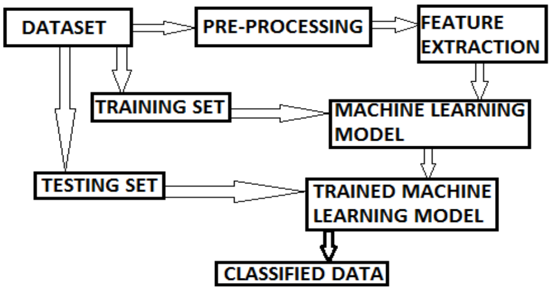

Fig.1

B. Features

The differences between conduct and language of the two classes-one of which has tweets indicating depression can be denoted using a set of attributes. All these measures have been taken on a document level which means that each user’s tweet has been analysed as an isolated document, thus making our work unique.

In this project a bag of words approach has been used to evaluate the text of a tweet. This approach uses the frequencies of the words occurring by depositing all the words within a bag and evaluating how frequently each word emerged.

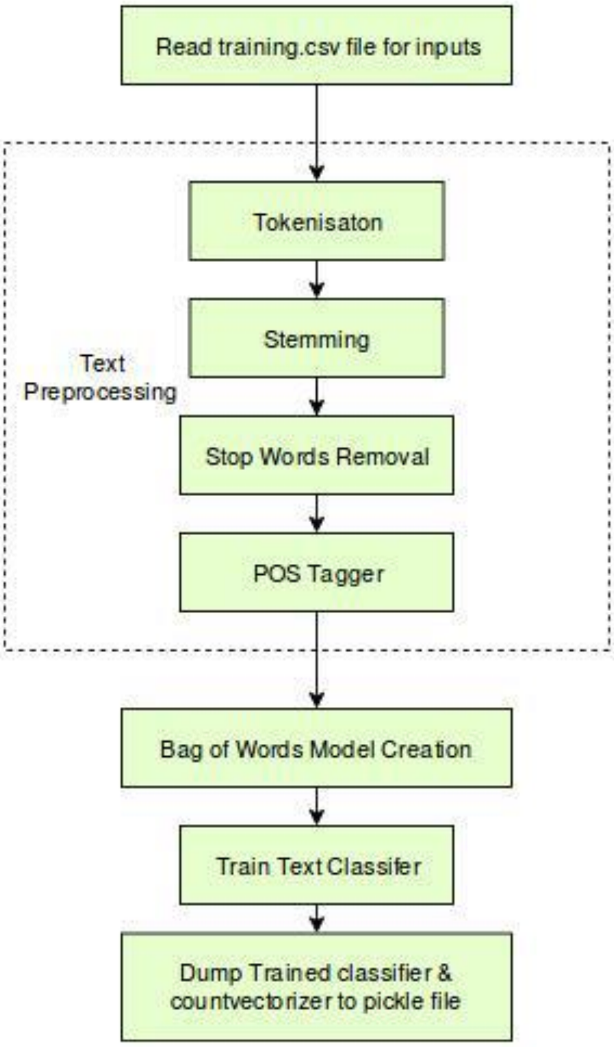
Read training.csv file for inputs
Text Preprocessing
Tokenisaton
Stemming
Stop Words Removal
POS Tagger
Bag of Words Model Creation
Train Text Classifer
Dump Trained classifier & countvectorizer to pickle file

Fig.2

C. Classifier

To assess the number of people affected by depression, we have used three types of classifiers. The three classifiers are support vector machine, tree-based approach and Naïve Bayes theorem. These models have been used against our dataset to get the results. In the coming section, a brief about all the models and how we have used them in our dataset has been given.

1. NAÏVE BAYES THEOREM

It is a classifier which puts into practice Bayes Theorem. It does not has a single algorithm but is an agglomeration of a number of algorithms sharing a common trait, that is, in all the algorithms, all the features which are classified are independent of one another. [9]

Bayes' Theorem basically "finds the probability of an event occurring given the probability of another event that has already occurred", i.e., it works on conditional probability. Naïve Bayes classifier is a common classifier as it gives high reliability, calls for a paltry training data to gauge the variables requisite for classification and calculates the result very quickly.

"The equation is as follows:

$$P(H|M) = \frac{P(E1|H) * P(E2|H) * P(En|H) * P(H)}{P(M)}$$

Here,

H is the probability of a classification

E1 to En are the Evidence variables

M is the Set of all evidences"

"Naive Bayes Classifier has different types and they are:

1. Gaussian Naive Bayes
2. Multinomial Naive Bayes
3. Bernoulli Naive Bayes"

Multinomial Naive Bayes has been used in this paper as the classifier because it performs better than the others on text data and gives better result for polynomial distributed data.

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})} \qquad (2)$$

$$p(C_k|x_1,\dots,x_n) = \frac{1}{Z}p(C_k)\prod_{i=1}^{n} p(x_i|C_k) \qquad (3)$$

$$p(\mathbf{x}|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!}\prod_i {p_{ki}}^{x_i} \qquad (4)$$

2. SUPPORT VECTOR MACHINE

It is an algorithm that inspects the recognizing patterns and data that is to be used for classification. When a set of input is given, SVM categorizes them in different categories. It can do both non-linear and linear classification. In this paper we have used Linear SVM. Algorithmically:

"Provided a training dataset of *n* points of form (*X1*, *Y1*), ….., (X*n*, Y*n*), where Y*i* is either 1 or -1, indicating each possible class of which the point X*i* may belong. Each Xi is a *p*-dimensional real vector, where we desire to determine the "maximum-margin hyperplane" which divides the group of points X*i* for which Y*i* = 1 from the points for which Y*i* = -1, such that the distance between the hyperplane and the nearest point Xi from either group is maximized. We define our hyperplane as a set of points *X* satisfying *W*•*X*–*B*=0, where *W* is the normal vector to the hyperplane. The parameter *w* determines the offset of the hyperplane from the origin along normal vector *W*." [9]

3. LOGISTIC REGRESSION

To approximately find the binary response probability grounded on one or more features we can use logistic regression. It is not a classification model technically but still we have used it because it stands for a distinct choice model. The connection between the features and a binary dependent variable are represented through the following equation:

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \qquad (1)$$

In the above equation, the success case boundaries of best fit, hence depression are shown by β0 + β1*x*. Therefore, the probability of t which is a dependent variable shows the depressed case and takes over a bias which is non-depressed. [9]

IV. RESULT

This section has been dedicated to inquire into and analyse the degree of accuracy established from the attributes drawn out from a user's linguistic history and exhibited by the existence of active depression in text data set.

The tokens used to find out depression in a set of data are F-measure, recall and accuracy.

Classification Algorithm	Precision	Recall	F-1 Score	Accuracy	Samples
Multinomial Naïve Bayes	0.836	0.83	0.8329	86%	10000
Support Vector Machine	0.804	0.79	0.7973	79%	10000
Linear Regression	0.85	0.82	0.81	82%	10000

TABLE 1

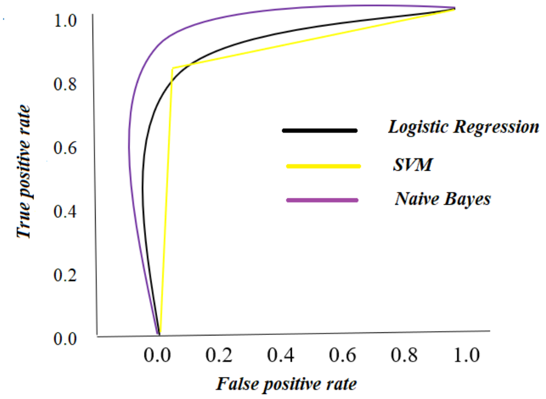

Fig.3

The classifiers' exactness by which they detected depression from a set of text is shown in the above graph. Count Vectorizer has been used in implementation.
Moreover, table 1 shows that Multinomial Naïve Bayes Theorem was able to achieve the maximum accuracy of 86% for the detection of depression in a set of data. It was Multinomial Naïve Bayes theorem that brought a basic Naïve Bayes Theorem's capability to the likes of Support Vector Machine. Even the precision score of Naïve Bayes Theorem was more than that of the other two classifiers (scoring 0.836).
Logistic regression attained a value very close to Naïve Bayes theorem in accuracy scoring 0.86. Recall value for Logistic regression was better than Naïve Bayes theorem but since the F-measure and precision score of Naïve Bayes was better than other two models therefore it has outdone the other two models.
Further, Area under Curve criterion was used to measure quantitatively the achievement of the classifiers' Receiver Operating Characteristic Curve. The below equation is used to find out the area under the curve for each classifier. [10]

$$A = \int_{\infty}^{-\infty} \mathrm{TPR}(T)\mathrm{FPR}'(T)\,dT = \qquad (5)$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(T' > T) f_1(T') f_0(T)\,dT'\,dT = P(X_1 > X_0)$$

Naïve Bayes theorem again achieved the highest score among the other two classifiers with a score of 0.94, second came Logistic Regression with a score of 0.91 and the last was SVM with 0.80 score. A score of 1 is considered the best for ROC AUC and Naïve Bayes was the closest to this score.
To detect depression, we have determined some scale for an absolute classifier: [11]
1. Accuracy over F1 score has been given more importance because identifying depression is more important than turning undependable because of a multitude of false positives.
2.Since our research deals with recognition of depression we target for a model which is control biased.
3. Time and computational resources are to be taken care of.
Linear Support Vector Machine would be best if (2) keeps more importance than (1) and Naïve Bayes is the best if (1) is given more importance since it gives high accuracy and high precision.

V. CONCLUSION AND FUTURE WORK

We have illustrated the prospective of using twitter as a tool for calculating and predicting major depressive disorder in individuals. At first, we compiled 3 datasets together to form a dataset which can be used in the model. Next, we made a bag of words approach towards quantifying this dataset. Finally, we grasp these different attributes to construct, compare and contrast several statistical classifiers which may foretell the likelihood of depression within an individual.

Our aim was to establish a method by which recognition of depression through analysis of large-scale records of user's linguistic history in social media may be possible. Future work may include detecting depression through images and through emoji's people use on their social media platforms to express themselves more precisely. [10] Future research can be done with possible improvements such as more refined data and more accurate algorithm.

VII. ACKNOWLEDGEMENT

We the students of Amity University Noida would like to acknowledge the efforts of all the individuals whose kind support and guidance made this Research Paper possible. We would like to extend our sincere thanks to all of them. We are highly indebted to **Prof. Mr. Nirbhay Kashyap** for taking us under his guidance and supervision. We are grateful that we shared his vast ocean of knowledge and experience of many years with us.

VI. REFERENCES

[1] M.S.Neethu, Rajsree, *"Sentiment analysis in twitter using machine learning techniques"*,Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013

[2] Mitali Desai, Mayuri A. Mehta,*"Techniques for sentiment analysis of Twitter data: A comprehensive survey"*, International Conference on Computing, Communication and Automation (ICCCA), 2016

[3] Krystian Horecki and Jacek Mazurkiewicz, *"Natural Language Processing Methods Used for Automatic Prediction Mechanism of Related Phenomenon"*, Springer International Publishing Switzerland ICAISC, 2015

[4] T. Al-Moslmi, N. Omar, S. Abdullah and M. Albared, "Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review," in *IEEE Access*, vol. 5, pp. 16173-16192, 2017.

[5] C. Clavel and Z. Callejas, "Sentiment Analysis: From Opinion Mining to Human-Agent Interaction," in *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 74-93, 1 Jan.-March 2016.

[6] K. Lee, A. Agrawal and A. Choudhary, "Mining social media streams to improve public health allergy surveillance" In *Proc. of the 2015 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining*, Paris, France, August 2015, pp. 815-822.

[7] V. Carchiolo, A. Longheu and M. Malgeri, "Using twitter data and sentiment analysis to study diseases dynamics" In

Proc. of the Int. Conf. on Information Technology in Bio-and Medical Informatics, Springer, Cham, Valencia, Spain, September, 2015, pp. 16-24.

[8] S. Liu, X. Cheng, F. Li and F. Li, "TASC:Topic-Adaptive Sentiment Classification on Dynamic Tweets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 6, pp. 1696-1709, 1 June 2015.

[9] M. Chen, W. Chen and L. Ku, "Application of Sentiment Analysis to Language Learning," in *IEEE Access*, vol. 6, pp. 24433-24442, 2018. doi: 10.1109/ACCESS.2018.2832137.

[10] R. Mehra, M. K. Bedi, G. Singh, R. Arora, T. Bala and S. Saxena, "Sentimental analysis using fuzzy and naive bayes," In *Proc. of the 2017 Int. Conf. on Computing Methodologies and Communication (ICCMC)*, Erode, 2017, pp. 945-950.