ISSN No:-2456-2165

Detecting Mental Distress through User's Social Media Activity

Isha Raina, B. Indra Thannaya
IGDTUW, Delhi

Abstract:- Users of social networking sites can approach their friends who are interested and expressing their thoughts, feelings, and sentiments through ideas, photographs, and videos. This opens the door to studying online information for user emotions and feelings in order to gain a better understanding of their emotions and attitudes when utilizing these online platforms.

Depression may be dangerous to one's health, particularly if it is recurring and of moderate or severe degree. It can make the individual suffer a lot and make them perform poorly at job, school, and at home. Suicide is a possibility when depression is severe. It is one of the leading causes of death among those between 15 to 29 of age.

Machine learning algorithms and Natural Language As the facts state that around 700,000 people in one year kill themselves.

Processing will be employed in the proposed problem statement to detect if a person is going through mental distress. The main aim is to discover that commonality within the tweets that can help in identifying whether the individual is on the edge of mental distress so that there's no delay is reaching out and helping the person who is suffering.

I. INTRODUCTION

Depression, being one of the most common mental illness, impacts about 300 million individuals throughout this world. Early identification is crucial for prompt action, which can help to prevent the illness from worsening. As per WHO stats, around 280 million people in the whole word are suffering from mental distress. In many nations, depression is still underdiagnosed and untreated, resulting in negative self-perception and, in the worst-case scenario, suicide [1]. The need to detect mental distress in individuals is alarming. Hence, this project is aimed at detecting if a person is depressed by analyzing their social network posts and tweets.

People have begun to express their experiences and struggles with mental health illnesses via online forums, microblogs, and tweets as the Internet has grown in popularity. Many researchers were influenced by their online activities to develop new types of prospective health-care solutions and approaches for early depression detection systems. They attempted to get a greater performance increase by employing several Natural Language Processing (NLP) methodologies and text categorization methods. [2]

In present era, data has turned into unstructured data from numerous businesses, social media, organizations, banking, and so on includes hidden patterns that, when studiedthoroughly, can expose new extents of research and development. However, reading all this vital material and coming to a decision is not an easy task. Text mining, opinion mining, text recognition, and so on all come into play here.

Natural Language Processing (NLP)

It is the practice of using software to recognize and deploy natural language such as speech and text automatically.

The following are the two primary components of NLP that are defined and described: -

- Natural Language Generation (NLG): The use of artificial intelligence (AI) programming to generate written or spoken narratives from a data collection is known as natural language generation. NLG incorporates computational linguistics, NLP, and NLU, as well as human-to-machine and machine-to-human interaction.
- Natural Language Understanding (NLU): It primarily entails the following two major tasks:
  - The provided natural language input is mapped to eloquent representations.
  - Recognizing different patterns in Natural Language.

NLP Pipeline

The aim is to break the problem down into little chunks and then use machine learning to address each one individually. Then intricate things can be performed by chaining together numerous machine learning models that feed into each other.

The steps to build a NLP Pipeline are:

- Sentence Segmentation: The first and the initial most stage in creating a natural language processing pipeline is this one. Breaking up the content into discrete phrases is the very initial step in the pipeline.
- Word Tokenization: After splitting out sentence, breaking the material into discrete phrases is the first stage in the pipeline. The sentences are broken down into words in order to define and understand the semantic meaning of each word independently in this step.
- Parts of Speech Predictions for Each Token: This step is to find out the part of speech for each work as they get converted into tokens now.
- Text Lemmatization: The major goal of this stage is to figure out what each word's basic form is so it can be acknowledged if different sentences are talking about the same entity or not. This technique is known as lemmatization or determining the most fundamental form or lemma of each word in the phrase.
- Identifying Stop Words: Before undertaking any statistical analysis, filtering out terms called stop words is important. In

ISSN No:-2456-2165

this step identification and elimination is done in this step. We mainly remove all the stop words present inn our data corpus.

- Dependency Parsing: This stage governs how the words in a sentence are associated to one another. Dependency parsing is the term for this. The aim is to create a tree that gives each word in the text a single parent word. The key verb in the expression will be the tree's root.
- Named Entity Recognition (NER): The goal of Named Entity Recognition is to identify these nouns and assign them towards the real-world concepts they reflect.

Machine learning (ML) is the ponder of computer calculations that can learn and create on their claim with involvement and information. It could be a component of manufactured insights. Machine learning algorithms create a model based on trained information to create forecasts or judgments without having to be unequivocally modified to do so. Machine learning calculations are utilized in a wide run of applications, such as medication, mail filtering, discourse acknowledgment, and computer vision, when existing algorithms are difficult or incomprehensible to construct.

The algorithms used in this project are: Logistic Regression & Support Vector Machine (SVM) and Random Forest.

Logistic Regression: This is a kind of linear classification technique which is used to estimate the possibility of a binary answer based on one or more predictors. [3][4]. The method of modelling the likelihood of a discrete result given an input variable is known as logistic regression. The most frequent logistic regression models have a binary result, which might be true or false, yes or no, and so forth. Multinomial logistic regression can be used to model circumstances with more than two distinctconclusions. When it comes to classification jobs, logistic regression is a useful analytical method for assessing if a new instance fits best into a category. Because components of cyber security, such as threat detection, are classification issues, logistic regression is a valuable analytic tool.

Support Vector Machine (SVM): Support-vector machines are a part of supervised learning models that analyse data for classification and regression analysis using learning techniques in machine learning. The Support Vector Machine (SVM) model portrays examples as points in a high-dimensional space used for classification, with the points of the various categories separated by a large distance.Unused occurrences are at that point mapped into the same space and classified concurring to which side of the crevice they arrive on[5]. By implicitly converting their inputs into high-dimensional vector space, the kernel technique allows SVMs to execute non-linear classification successfully. The purpose of the support vector technique is used to find a hyperplane that differentiates amongst data points in an N-dimensional space (N = the number of characteristics).

Random Forest: Random Forest (RF) is a set of decision tree classifiers trained using the bagging approach, in which a number of different learning models are combined to improve the overall output. [6]

Implementation Steps are given below:

- Data Pre-processing step
- Fitting into the training set our algorithm which isRandom Forest
- Predicting the test result
- Testing the accuracy of the result obtained
- Visualizing the test set result.

II. BACKGROUND LITERATURE

Improved feature selection and combination helps in improving classifier performance and accuracy.Raza Ul Mustafaa, Noman Ashrafb, Fahad Shabbir Ahmedd, Javed Ferzunda, Basit Shahzadc, Alexander Gelbukh, 2020 concluded in their paper A Multiclass Depression Detection in social media based on Sentiment Analysis [7] Using Neural Network, SVM,RFand 1D Convolutional Neural Networks they achieved 91% accuracy. In the similar fashion Michael M. Tadesse, Hongfei Lin, Bo Xu ,Liang Yang in his paper “Detection of Depression - Related Posts in Reddit Social Media Forum”, 2019 [2] obtained an accuracy of 80% using Random Forest, SVM, Decision tree, Logistic Regression, Adaptive Boosting, Multilayer Perceptron and stated that the model's accuracy is excellent. Machine learning and deep learning algorithms can be used to improve the model and further study.Akshi Kumar, Aditi Sharma, Anshika Arora found a way to improvise the models to get a better result using Boosting, Random forest, Multinomial, Naive Bayes. [9] Their research was implemented on Twitter data Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution [10] algorithm provided Depression behavior discovery. Multimodal Depressive Dictionary Learning (MDL) method achieved the best performance with 85% in F1-Measure with NB and MSNL algorithms (2017).

K-Nearest Neighbours , Decision tree, SVM, Ensemble, the algorithms when put together prove to be sufficient and efficient in the detection of the depression with accuracy between 60 and 80% [11] as seen in the paper “Depression detection from social network data using machine learning techniques” where Facebook data was used.

On the other hand Munmun De Choudhury, Michael Gamon, Scott Counts, Eric Horvitz in 2018 used PCA, Support Vector Machine classifier(SVM) and concluded that Findings and methods used in the research are useful in developing toolsfor identifying the onset of major depression, for use byhealthcare agencies. [12]

Using Convolutional Neural Network the accuracy was found out to be 82% on Twitter data in Multimodal mental health analysis in social media [13]. Deep learning and neural networking were the major components of this method. David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt, Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt[14] compared the results with algorithms like SVM

ISSN No:-2456-2165

and Linear Regression and obtained an accuracy of ~70% in the paper Detecting depression and mental illness on social media: an integrative review.

Hoyun Song∗Jinseon You∗Jin-Woo Chung Jong C. Park [15] proved that FAN considers only four featureswhich are not sufficient on themselves entirely to detect depression as they used Feature Attention Network, Multilayer Perceptron(MLP), GloVe, GRU, one of theRecurrent Neural Network(RNN) variants, L2 Regularization, Adam Optimizer, ConvolutionalNeural Network(CNN-E,CNN-R) it was observed that it outperforms all themodels except the CNN-R model, FAN shows a similar F1- score to the baseline methodologies.In 2020 Zhenpeng Chen, Yanbin Cao, and HuihanYaodeve developed a model DeepMojiModel,SEntiMojiModel [16]SEntiMoji was beneficial for tasks that mainly depend on emotion identification. The method of construction of variousdatasets can be different. so, the performance should be analysedrationally was the main objective and aspect behind their research.

MyStem, Udpipe, Linis-Crowd SentimentDictionary, Random Forest, Support Vector Machine used SVM+PM-r model with 75.1% ROC-AUC score using Vkontakte dataset[17]. They discovered that the n-gram and tf-idf based features did not perform as expected over the dataset.

With a F1 score of 0.51 Random Forest, Logistic Regression, Naive bayes, CNN, GloVe W+N, etc algorithms were used but did not come out to be a suitable metric for this task they used Reddit data [18]. Kali Cornn in her research paper "Identifying Depression on social media" [19] used Logistic Regression, Support Vector Machine, BERT- based model, Character based CNN model without embeddings, Character-based CNN with pre trained word embeddings and got BERT accuracy 85.7% & CNN accuracy 92.5%. Use of word embedding proved to be a disadvantage. The major issue with CNN model is the high amount of increase in the training time. Random forest with two threshold functions, two independent RF classifiers was used in "Early detection of depression: social network analysis and random forest techniques" [20] They concluded that Time-based approach is effective and that different model combinations can be compared and studied in future for better results and performance. Again in "Study of depression analysis using machine learning techniques" by Devakunchari Ramalingam, Vaibhav Sharma, Priyanka Zar [21] SVM and RF was used applied on Weibo and Twitter dataset which gave an accuracy of 82% but the lack of a perfectly accurate model was a big disadvantage.

Hatoon S. AlSagri, Mourad Ykhlef [22] used SVM, Naive Bayes, Decision Tree and Machine learning based approach for depression detection in twitter using content and activity features where they concluded that the Decision tree is more comprehensive and evaluative but failed to detect depression with a good accuracy The best results are shown by SVM with a precision of 73.6% and accuracy of 77.5%.

Some of the challenges with current models:

- There is a need to find out more features that can relate to human behavior and help in the detection of depression.
- The n-gram and tf-idf based features did not perform as expected over the dataset.
- There are several improvements to be made for better optimizations.
- Use of word embedding proved to be a disadvantage and the main issue with CNN model is the high amount of increase in the training time.
- Fine grain emotion analysis can be done for the purpose of anxiety detection.
- There is a requirement to work on the ethical aspects and terms to extend this form of study (i.e. depression detection).
- There is a need to build a smart AI system that can analyze the symptoms from tweets accurately. The lack of a perfectly accurate model is a big disadvantage.

III. DATASET DESCRIPTION

The dataset used is Sentiment140 [23] dataset with 1.6 million tweets.

The data consists of 4 columns namely Target, User_Name, ID&Tweet_Text. We have combined some part of the Sentiment140 and the scraped depressive tweets to form a new dataset.

Column	Description	Data Type
Target	Polarity of tweet (4 - depressed, 0 - positive)	Int
User_Name	The user that tweeted (armotley)	Int
ID	The id of the tweet (2087)	Object
Tweet_Text	The text of the tweet (about to file taxes)	Object

Table 1: Shows the dataset description with column definition

According to the creators of dataset "Our method was unusual in that our training data was generated automatically rather than by people manually annotating tweets. We took the view that any tweet containing positive emoticons, such as:), was positive, and any tweet with negative emoticons, such as:(, was negative. We gathered these tweets using the Twitter Search API and a keyword search"

IV. METHODOLOGY

The proposed method is based on machine learning processing algorithms like Support Vector Machine, Random Forest and Logistic Regression and for text preprocessing we will be using NLP.

The process of the methodology that will be followed is:

- Data Collection: In this step the data is collected based on the problem statement. The result of this phase is often a data representation that will be used for training.

ISSN No:-2456-2165

- Data Preparation: This is a crucial part in the process, and people generally spend up to 80% of their time here. Having a clean data collection improves the accuracy of model in the long run. The obtained data is subsequently cleaned up by deleting any redundant, undesired, or null variables that might impair the model's or algorithm's accuracy.
- Model Selection: In ML, there are a variety of algorithms tochoose from. The need is to figure out which algorithm is the best out of all the options. In this project SVM, Logistic Regression, and Random Forest will beused.
- Model Training: In this step the data set is linked to an algorithm, which learns and develops predictions using advanced mathematical modelling. These algorithms are usually classified into one of three groups:
  - Binary – Divide into two groups.
  - Classification - Sort into a variety of categories.
  - Predict a numeric value using regression.
- Model Evaluation: Model evaluation means finding out the given factors namely: accuracy, precision and recall. These are the three basic measures used to evaluate a classification model.
- Parameter Tuning: Tuning is the process of enhancing a model's performance while avoiding overfitting or excessive variance. The model is fine-tuned by altering the learning rates or the values of the test and train datasets. This is performed in machine learning by picking appropriate "hyperparameters."
- Predictions: Data is used in machine learning to answer problems. So, inference, or prediction, is the point when we get to answer certain questions. This is the culmination of all of the efforts, and it is here that the benefit of machine learning is apparent.

V. IMPLEMENTATION STEPS

- Deleting columns that are not important
- Data Cleaning: Removing any links, twitter handles (username), punctuations & special characters present in the data
- Text Processing: Removing stopwords by importing nltk library
- Text Tokenization

VI. RESULT

The accuracy of the model with different algorithms is given in table-

Metrics	Logistic Regression	SVM	Random Forest
Accuracy	0.72373279	0.72630636	0.70706053
Precision	0.72401699	0.72946045	0.70738364
Recall	0.72358226	0.72586216	0.70718665
F1 Score	0.72354649	0.72509602	0.70701695

VII. CONCLUSION AND FUTURE WORK

If we totally depend on the findings of our model and algorithm, we may come to the conclusion that there is a clear need to realize that individuals are going through a variety of different situations, which can be observed in their statements.

The future potential in this subject might include developing a model for sending frightening information to individuals, as well as the government and medical authorities, in order to minimize depression instances in the country and worldwide

REFERENCES

[1.] EEEE - M. J. Friedrich, ‘‘Depression is the leading cause of disability around the world,’’ JAMA, vol. 317, no. 15, p. 1517, Apr. 2017
[2.] Tadesse, Michael M., et al. "Detection of depression-related posts in reddit social media forum." IEEE Access 7 (2019): 44883-44893.
[3.] Gortmaker, Steven L. "Theory and methods--Applied Logistic Regression by David W. Hosmer Jr and Stanley Lemeshow." Contemporary sociology 23.1 (1994): 159.
[4.] Agresti, Alan. An introduction to categorical data analysis. John Wiley & Sons, 2018.
[5.] Noble, William S. "What is a support vector machine?." Nature biotechnology 24.12 (2006): 1565-1567.
[6.] Xu, Baoxun, Yunming Ye, and Lei Nie. "An improved random forest classifier for image classification." 2012 IEEE International Conference on Information and Automation. IEEE, 2012.
[7.] Mustafa, Raza Ul, et al. "A multiclass depression detection in social media based on sentiment analysis." Proceedings of the 17th IEEE International Conference on Information Technology—New Generations. Springer, 2020.
[8.] Stephen, Jini Jojo, and P. Prabu. "Detecting the magnitude of depression in Twitter users using sentiment analysis." International Journal of Electrical and Computer Engineering 9.4 (2019): 3247.
[9.] Kumar, Akshi, Aditi Sharma, and Anshika Arora. "Anxious depression prediction in real-time social data." International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India. 2019.
[10.] Shen, Guangyao, et al. "Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution." IJCAI. 2017.
[11.] Islam, Md Rafiqul, et al. "Depression detection from social network data using machine learning techniques." Health information science and systems 6.1 (2018): 1-12.
[12.] De Choudhury, Munmun, et al. "Predicting depression via social media." Seventh international AAAI conference on weblogs and social media. 2013.
[13.] Yazdavar, Amir Hossein, et al. "Multimodal mental health analysis in social media." Plos one 15.4 (2020): e0226248.
[14.] Guntuku, Sharath Chandra, et al. "Detecting depression and mental illness on social media: an integrative review." Current Opinion in Behavioral Sciences 18 (2017): 43-49.
[15.] Song, Hoyun, et al. "Feature attention network: interpretable depression detection from social media."

Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation. 2018.
[16.] Chen, Zhenpeng, et al. "Emoji-powered sentiment and emotion detection from software developers' communication data." ACM Transactions on Software Engineering and Methodology (TOSEM) 30.2 (2021): 1-48.
[17.] Stankevich, Maxim, et al. "Depression detection from social media texts." Elizarov, A., Novikov, B., Stupnikov., S (eds.) Data Analytics and Management in Data Intensive Domains: XXI International Conference DAMDID/RCDL. 2019.
[18.] Trotzek, Marcel, Sven Koitka, and Christoph M. Friedrich. "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences." IEEE Transactions on Knowledge and Data Engineering 32.3 (2018): 588-601.
[19.] Cornn, Kali. "Identifying depression on social media." Department of Statistics Stanford University Stanford, CA 94305 (2020).
[20.] Cacheda, Fidel, et al. "Early detection of depression: social network analysis and random forest techniques." Journal of medical Internet research 21.6 (2019): e12554.
[21.] Ramalingam, Devakunchari, Vaibhav Sharma, and Priyanka Zar. "Study of depression analysis using machine learning techniques." Int. J. Innov. Technol. Explor. Eng 8.7C2 (2019): 187-191.
[22.] AlSagri, Hatoon S., and Mourad Ykhlef. "Machine learning-based approach for depression detection in twitter using content and activity features." IEICE Transactions on Information and Systems 103.8 (2020): 1825-1832.
[23.] https://www.kaggle.com/kazanova/sentiment140 - Dataset