Classification of Depression on social media using Distant Supervision

Kuldeep Vayadande[*], *Aditya Bodhankar, Ajinkya Mahajan, Diksha Prasad, Shivani Mahajan, Aishwarya Pujari, and Riya Dhakalkar*

Vishwakarma Institute of Technology, Pune, Maharashtra, India

Abstract. Amidst Covid-19, young adults have experienced major symptoms of anxiety and/or depression disorder (56%). Mental health issues have been spiking all over the world rapidly. People have taken up to social media as a platform to vent about their mental breakdowns. Twitter has seen enormous rise in depressive and anxious tweets in these times, but the downside being that majority of the population has neglected the importance of mental health issues and there are not enough resources to liberate people about it. Also, people hesitate to talk about their mental issues and seek help. So, a machine learning model using distant supervision to detect depression on Twitter is curated. Use of Sentiment140 dataset with 1.6 million records of different tweets. Our training data makes use of Twitter tweets included with emojis, which are classified as noisy labels on a dataset. Further, this paper mentions about how to use models like Support Vector Machine (SVM), Logistic Regression, Naive Bayes, Random Forest, XGBoost to distinguishing tweets between depressive or non- depressive. The purpose behind using multiple models is to achieve highest accuracy when trained with emoticon dataset. The paper's main contribution is the idea of using tweets with emoticons for distant supervised learning.

1 Introduction

Mental Health in medical terms is referred to as cognitive, behavioural and emotional well-being. In simpler terms, it is the way people think, feel, and behave. Mental health affects various aspects of life like personal relationships, working environment, etc. Surprising as it may sound, physical health can also affect mental state in several ways. Since 2020's lockdown period, mental health issues have been coming into picture more than ever. People have been isolating themselves and due to months and months of quarantining, deteriorating mental states have been observed in majority of world's population. The following mental health issues have risen:[*]

A. Clinical Depression: A mental health condition marked by a consistently sad mood or a loss of interest in activities, resulting in considerable impairment in everyday living.
B. Anxiety Disorder: Anxiety makes getting through the day tough. Nervousness, panic, and terror are common symptoms, as are sweating and a racing heart.

[*] Corresponding author: - kuldeep.vayadande@gmail.com

C. Bipolar Disorder: A mood swing illness characterized by bouts ranging from depressive lows to manic highs. Although the specific origin of bipolar disease is unknown, genetics environment, and changes in brain structure and chemistry may all have a role.
D. Dementia: A set of mental and social symptoms that makes daily life difficult. Dementia is a term used to describe a range of disorders in which at least two brain functions, such as memory and judgement, are impaired.
E. Schizophrenia: Schizophrenia is an excessive highbrow state of affairs in which patient's perceptions of reality are distorted. Schizophrenia can purpose hallucinations, delusions, and essentially aberrant wondering and behaviour, making daily existence challenging. Patients that suffer with this want a lifestyles lengthy treatment.

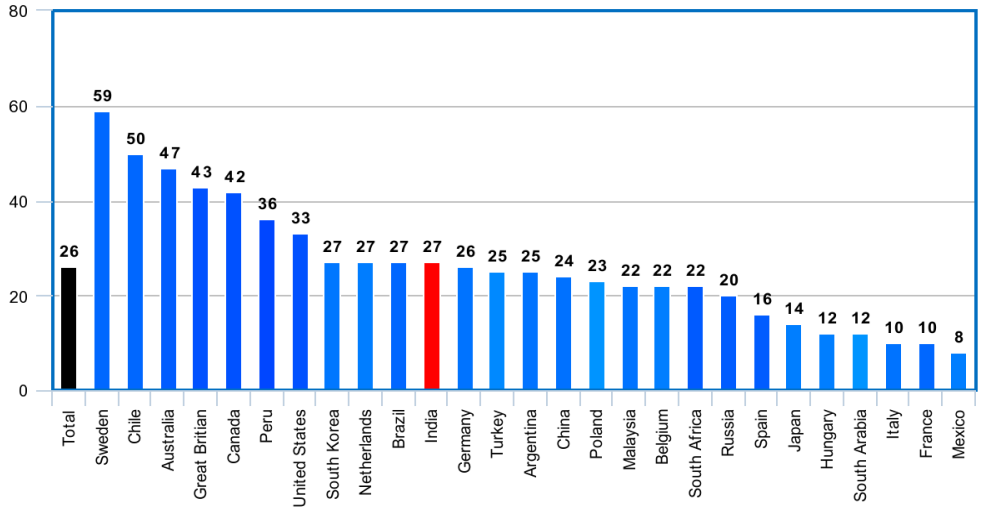

Fig. 1. Percentage of people who say mental health is top health problem (Source: Ipsos Global Health Monitor 2020)

Social media has become a crucial part in personal and professional lives of everyone. In lockdown it became extremely difficult for people to socialize outdoors which led to even more rise in online socializing through different platforms like Instagram, Facebook and Twitter.

Also, due to fear of judgement people lean more towards online interaction and expression rather than in-person. More and more people have been sharing their stories of dealing with mental issues on social media. In many places around the world society fails to realize the depth of mental health issues which makes it suffocating for the people dealing with those issues, therefore they choose to stay numb. But social media has come to their rescue in several cases and people have felt better talking about their problems. It has also helped organizations around the globe to track mental issues and thereafter provide aid to the people. Through someone's post, it is possible to predict how he/she is feeling. Machine Learning is a way of training models to predict outcomes. Through someone's post, it is possible to predict how he/she is feeling. So, with the help of it, the model can be trained in such a way that it can predict whether a given post depicts depression or not.

The emojis feature as loud labels. For example, the symbol ':' is included in a tweet (denotes accurate sentiment and:(denotes terrible sentiment. It is easy to extract vast quantities of tweets consisting of emoticons with the use of Twitter API. This is a vast development over manually labelling schooling facts that could take many hours. Educated

classifiers are placed on emoticon facts to the check on a hard and fast of tweets that can or might not include emoticons.

2 Problem Statement

As the rate of mental issues is spiking day by day, it becomes of utmost significance to identify the presence of depression as early as possible, in order to provide aid. The outcome of this model is to identify presence of depression early on social media that is done using Machine Learning approaches. In this research, tweets gathered from Twitter API (Sentiment140 dataset) have been classified as depressive or non-depressive. Five different ML classifiers have been applied to compare and achieve high accuracy.

3 Literature Review

Xujuan Zhou et al [1], these authors have proposed an analysis of tweets using (TSAM) Model that might spot the social interest. This paper includes the Australian federal election of year 2010 occasion turned into an instance for sentiment evaluation experiments. The used of framework of the TSAM includes of three modules:

- Module of Feature Extraction
- Sentiment identity module
- Sentiment aggregation and rating module.

Hatoon AlSagri et al [2], have presented research that indicated a correlation between excessive utilization with social media websites and depression. This examines objectives to make the most gadget gaining knowledge of strategies for identification a Twitter user that is likely to be depressed. The effects confirmed that the extra capabilities used result in higher accuracy and F- degree rating in detecting depressed users.

Guanghui Fu et al [3], worked on Suicide Risk Classification System with 2 million messages in a Chinese social platforms information supply that are generated every day. These messages were identified as crucial records that deliver for preventing suicide related to depression. The authors advise growing a device that could automatically discover textual remarks which can be indicative of excessive suicide risk.

Nirmal Varghese et al [4], proposed a system of the sentiment using Artificial Intelligence. Social media records might be applied for the complete procedure i.e., the evaluation and classification approach if it includes textual content records and emojis. It consists of Multiclass Classification with a Deep Learning algorithm that indicates better F-score price throughout the phrase evaluation.

Nafiz Al Asad et al [5], proposed a system to advise a facts-analytic, primarily depends upon the despair of any human being. In this proposed work, version facts that occur from the user's posts on famous social media websites: Twitter and Facebook are considered. The Depression degree of a consumer has been detected primarily based totally on his posts on social media.

J. Li et al [20], present a vector refinement model that is based on a genetic algorithm, which employs an improved genetic algorithm to optimize representations of word vector, so as to capture the word's sentiment. The suggested model can show experimental results and enhance existing binary classification and fine-grained classification SST.

H. Silva et al [21], analysed the opinions expressed on Twitter before and during the COVID-19 outbreak. Portuguese tweets about SUS that were posted. The tweets were processed under preliminary processing, classification, and analysis.

A. J. Nair et al [22], proposed a work that uses sentiment analysis using techniques like Logistic Regression, BERT sentiment analysis. The suggested analysis approach can be adapted based on the domain but are sensitive to sentiment expressions.

4 Methodology

The system will be using the Sentiment140 dataset for the training of our models. The models may vary with accuracy in terms of predicting the correct polarity of the given tweet because of the diverse algorithms. The system is designed to accept a tweet and predict if the tweet is positive or negative in nature. But, before the prediction of tweets, the training of models and dataset is required to go through various listed processes that are elaborated on in further sections of the paper:

1. Collection of Data
2. Analysis of Data
3. Pre-processing Data
4. Feeding processed data to different models for training
5. Feeding the testing data for testing
6. Acquiring results and predictions

In data analysis, the aim is to get a better understanding and hold on data available. The dataset is retrieved from Kaggle UCI [19]. The data set contains 1,600,000 rows with 6 different columns. The dataset used is Sentiment140. It consists of 1.6 million tweets collected from the existing API of Twitter. The tweets were marked as 0 indicates negative and 4 indicates positive and were used for sentiment detection. Here are the few steps that were followed accordingly throughout the process.

4.1 Dataset Collection

Table 1. The below table depicts the attributes in the dataset with their meaning.

Column Name	Description
Target	The nature of the tweet (0 = depressive, 4 = non- depressive)
Ids	Tweet ID
Date	Tweet Date
Flag	If there is empty query, then value will be NO QUERY
User	User that twitted
Text	Tweet text

4.2 Analysis attribute and correlation

```
Negative Sentiment Tweets Percentage is 49.89515000085244%
Positive Sentiment Tweets Percentage is 50.10484999914756%
```

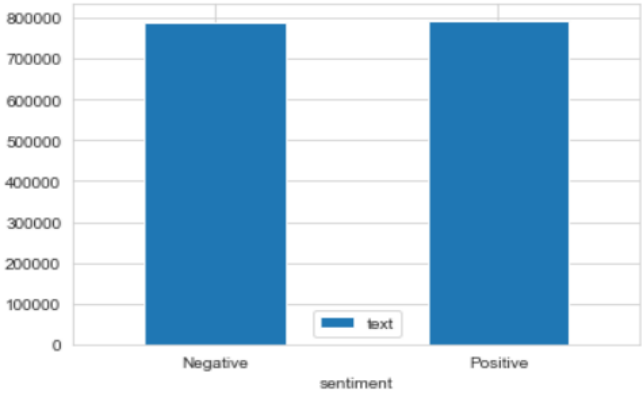

Fig. 2. Target count plot

4.3 Dataset Preprocessing

The Twitter language model contains a number of distinct characteristics. To reduce the feature space, use of the attributes listed below are cleared:

- Usernames: To direct their remarks, users frequently use Twitter usernames in their tweets. The @ symbol is commonly used before the username (for example, @alecmgo). All words that begin with the @ symbol are replaced with an equivalence class token (USERNAME).
- Link usage: Tweets with links are common. An equivalent class is used to match all URLs. A URL, such as "http://somewebsite.com," is turned to the token "URL."
- Repeating alphabets: Tweets use fairly informal language. On Twitter, for example, if you search "hey" with an arbitrary number of y's in the end (e.g., heyyy, heyyyy, heyyyyyy). We employ pre-processing to replace each letter that appears more than twice in a row with two occurrences. These words would be translated to the token yummy in the examples above.

The impact of these feature reductions is shown in Table 2. These three changes reduce the attribute set's actual size to 45.85%.

Table 2. Feature Reduction and Percentage.

Feature Reduction	No. of Features	Percent of Original
URLs	730152	91.86%
All	364464	45.85%
None	794876	100.00%
Repeated Letters	773691	97.33%
Username	449714	56.58%

5 Overview of Proposed Model

Machine Learning classifiers help predict possible outcomes for the required target variables when the data is properly processed, and models are trained for maximum accuracy.
Use of several different models like SVM (Support Vector Machine), XGBoost, Logistic Regression, Random Forest and Naïve Bayes to achieve high accuracy in predicting whether a given tweet is depressive or non-depressive. The models will be used independently of each other to predict the tweet and provide the accuracy each of them has achieved. The models and various factors like accuracy and F-score are provided in Section 8 of the paper.

6 Proposed Model and Architecture

Consideration of half of the dataset, i.e., the initial 8 lakh tweets with positive emoticons and the other initial 8 lakh tweets with negative emotions, to obtain a total of 16 lakh training tweets, after post-processing the data.
To collect test data, we employ the following procedure:

Certain queries are used to search the Twitter API. These searches were chosen at random from various domains. These queries, for example, include consumer products like kindle2 and businesses like at&t, and people like Elon Musk and Warren Buffet. The table below lists all the different categories of our searches.

Table 3. Twitter API categories of Searches

Category	Total	Percent
Misc	67	18.66%
Movie	19	5.29%
Product	63	17.55%
Person	65	18.11%
Location	18	5.01%
Company	119	33.15%
Event	8	2.23%
Grand Total	359	

Examining process of the query's result set. Classify of a result as good or negative if it has a sentiment. As a conclusion, the result set that is produced may or may not consist of emoticons and emojis in the required field. Application of various classifiers on this dataset and make a comparative analysis to find out which classifier provides best results on the given dataset keeping in mind other factors of dataset like weak annotations like emojis that is being covered in the distant supervision part.

7 Architecture

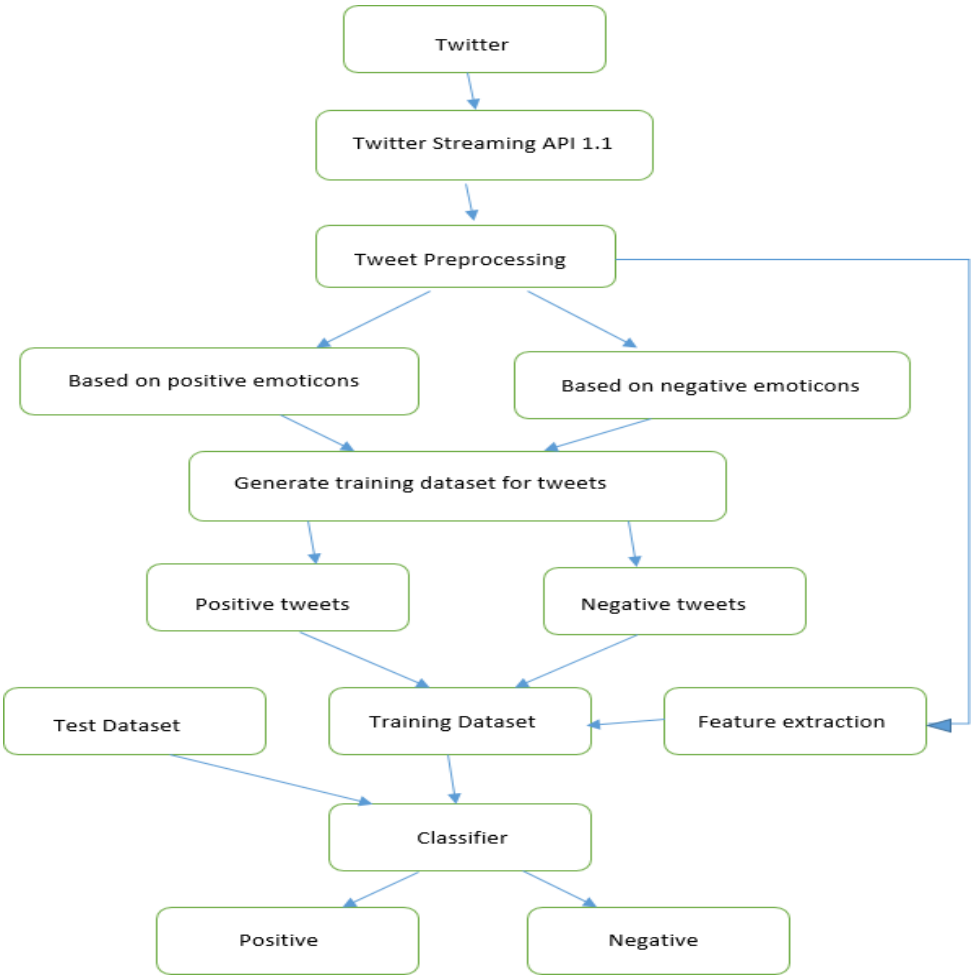

Fig. 3. Architecture of Algorithm

The Twitter Streaming API is used to extract the tweets from Twitter Dataset. The version of this API is 1.1. The collected data is raw and is pre-processed and removal of unwanted parameters that are not being used in the models via various steps mentioned above like removing usernames, hyperlinks, repeating alphabets and bringing alphabets to their present participle to ease the process of prediction. After the completion of the pre-processing task, a brief process of feature extraction is executed to help in the training data. The data that was divided into positive and negative emotions are combined to generate a mixed training dataset. The remaining data that was not part of training data is included in the mixed testing dataset. The training and testing dataset is now fed to all the models that are applied in the models and learn from the training data and predict positive and negative emotions on testing part.

8 Machine Learning Classifiers

There are 5 classifiers that are being used: Naive Bayes, XGBoost, Support Vector Machine, Logistic Regression and Random Forest classifier:

8.1 Support Vector Machines (SVM)

SVM stands for Support Vector Machines is a popular Supervised Machine Learning algorithms for problems like Classification and Regression. It is majorly applied in Machine Learning for classification issues. The SVM set of rules motive is to locate the most reliable selection boundaries for categorizing n-dimensional area into training in order that extra facts points may be comfortably positioned in an appropriate class in the future. The severe factors/vectors that help create the hyperplane are selected through SVM.

Accuracy achieved for this model is around 81.43%.

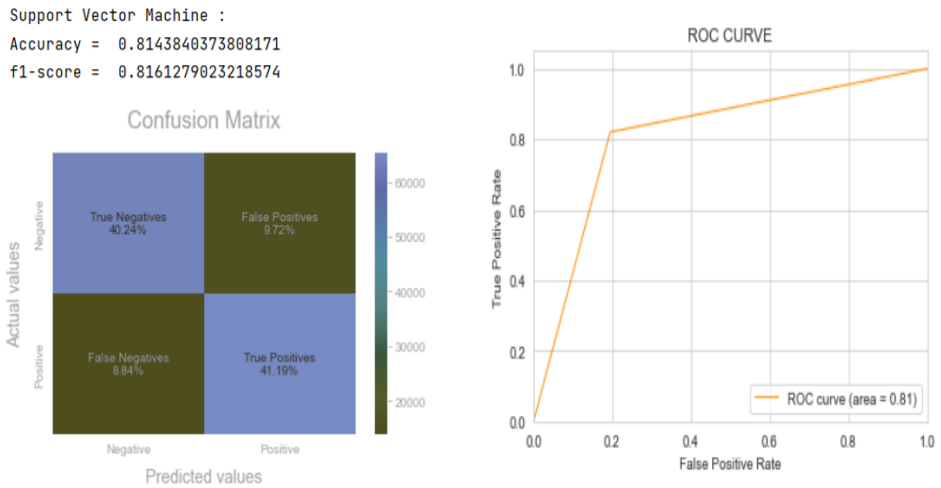

Fig. 4. Obtained Confusion Matrix and ROC Curve using SVM

8.2 Logistic Regression

Logistic regression is generally used to solve or predict binary classification problems. The target outcome is basically dichotomous in nature which means that there can be only 2 possible classes for the target variable is either 0 or 1.
Parameter(s):
1. fit_intercept: It has a Boolean value where its default value is 'true'. It basically specifies bias or intercepts constant must be combined with the decision function.
2. n_jobs: It has an int value where its default value is 'None'. If multi_class='ovr' (one vs rest), it generally represents the number of CPU cores used during parallelizing over classes. If solver='liblinear' then it's ignored.

The accuracy achieved for this model is 82.41%.

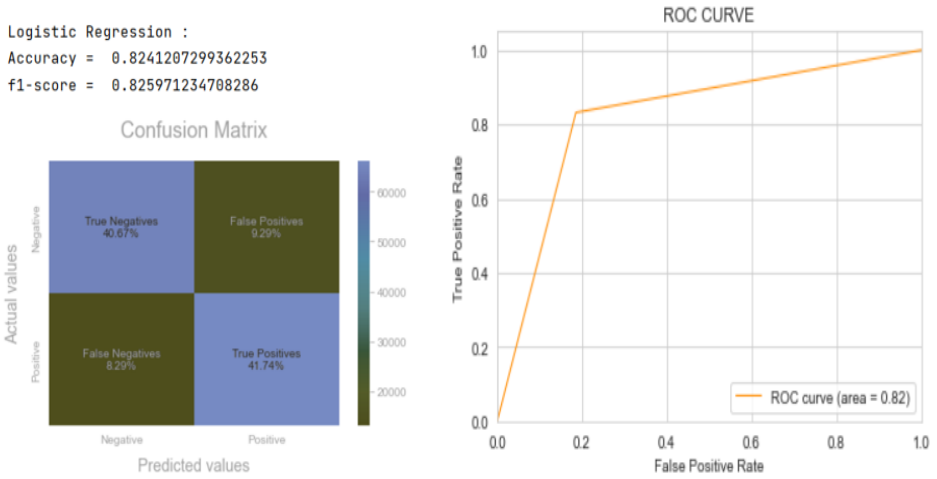

Fig. 5. Obtained Confusion Matrix and ROC Curve using Logistic Regression

8.3 Naïve Bayes

The Bayes Theorem is preferred when we wish to make a set of type algorithms referred to as Naive Bayes classifiers. It is one group of classifiers that can be compared in percentages, such that every couple of features being labeled is unbiased from the others. Models assign elegant labels to problem occurrences, expressed in form of vectors of characteristic values, the usage of the Naive Bayes technique. For education of such classifiers, there may be no one set of rules, however alternatively a ramification of algorithms that are based on the identical principle.

The accuracy achieved for this model is 80.53%.

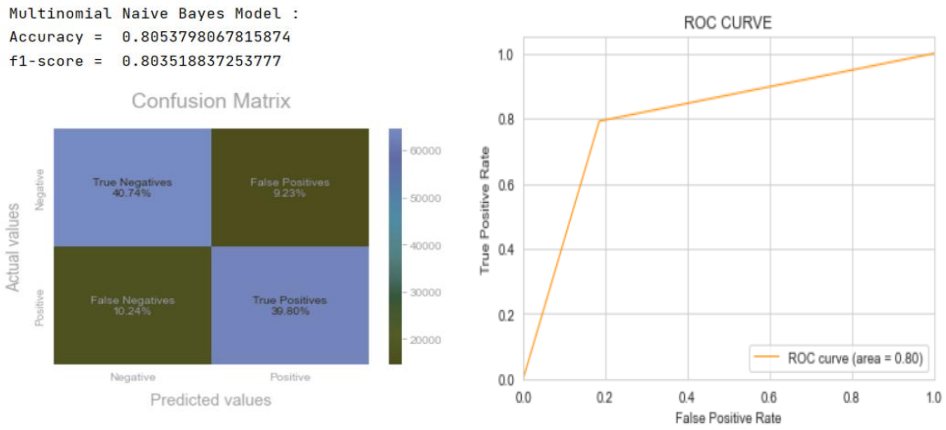

Fig. 6. Obtained Confusion Matrix and ROC Curve using Naïve Bayes

8.4 Random Forest

Random Forest is a popular device for knowing the set of rules that uses supervised gaining knowledge of techniques. In system gaining knowledge of, it could be used for both categorical and regressive issues. It is based on gaining knowledge, that's a method of integrating numerous classifiers to solve many relatively complex problems and therefore increase the model's result and accuracy. The greater the number of trees within the wooded area, the better the accuracy and the decrease the threat of overfitting.

The accuracy achieved for this model is 81.34%.

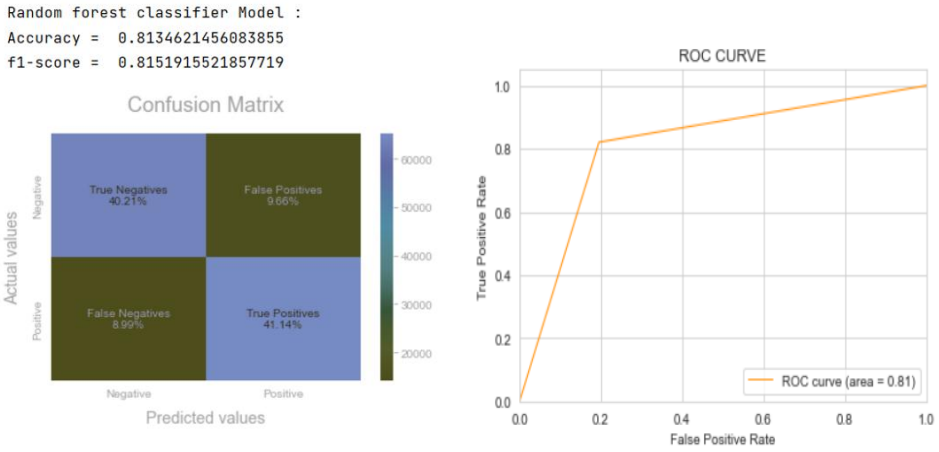

Fig. 7. Obtained Confusion Matrix and ROC Curve using Random Forest

8.5 XGBoost

XGBoost is an aggressive system mastering implementation of gradient boosted choice bushes optimized for speed and overall performance. XGBoost is a dispensed gradient boosting toolkit this is optimized for performance, flexibility, and portability. It makes use of the Gradient Boosting framework to implement the system gaining knowledge of algorithms. XGBoost uses parallel tree boosting (also known as GBDT or GBM) to tackle an expansion of data technology issues fast and as it should be. The same algorithm may also address troubles with billions of examples in disbursed surroundings (Hadoop, SGE, MPI).

The accuracy for this model is 80.56%.

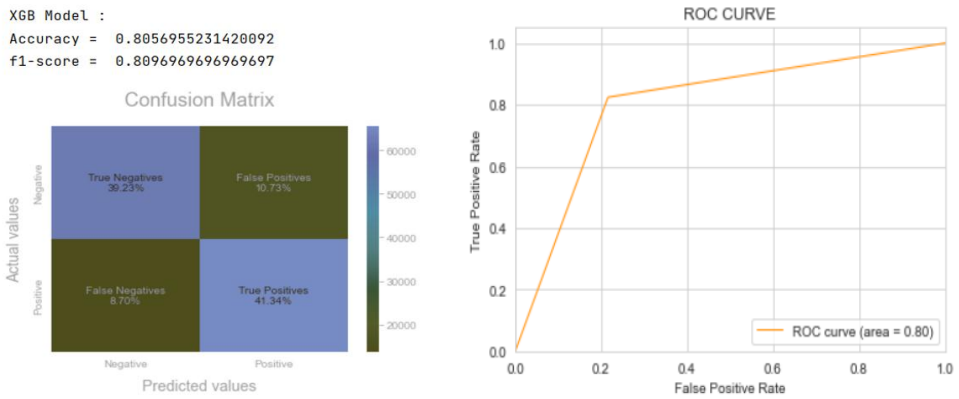

Fig. 8. Obtained Confusion Matrix and ROC Curve using XGBoost

9 Machine Learning

Machine gaining knowledge of ML is a subset of Artificial Intelligence (AI) that enables software to predict outcomes with high accuracies and that too without any manual programming to achieve the same. Machine getting to know algorithms use historic facts as input to expect new output values. Machine Learning is the field of examination that offers computer systems the capability to examine without explicit program. It is one of the most thrilling technologies that one might have ever encountered. As it's far glaring from the name, it offers a laptop that makes it more just like humans. Machine gaining knowledge is actively getting used today, perhaps in many extra places than one might count on. The following flow diagram explains all techniques thar are used in the project.

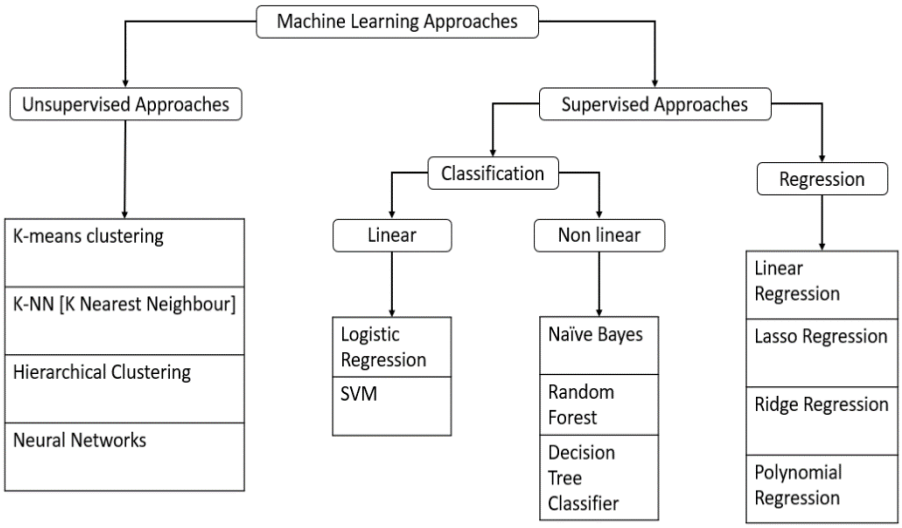

Fig. 9. Used Machine Learning Techniques Flowchart

9.1 Supervised Learning

Supervised learning technique is used for developing Artificial Intelligence (AI) that includes training a computer algorithm on given data which is labeled for some output. When given the never-before-seen data, the model trains until it can recognize the patterns and relations between the input data and the output values, permitting it to predict accurate results. However, to achieve this, the supervised learning models are needed to be properly trained using well labeled data so that the model can classify accurately and perform more precisely in the testing phase of the model.

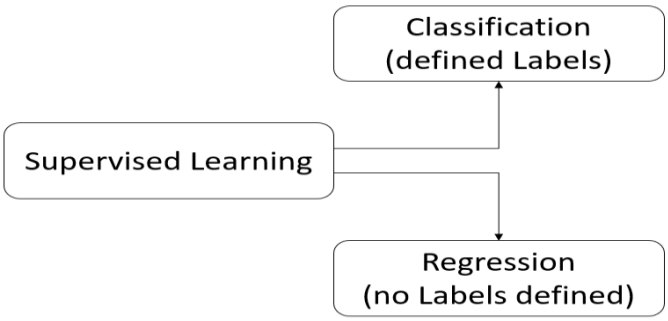

Fig. 10. Types of Supervised Learning based on Labels

Supervised learning excels in classification and regression issues such as determining media article's category or predicting the amount of sales in near future. The aim of supervised learning is to make the data insightful for a specific problem.

9.2 Unsupervised Learning

Most gadget mastering strategies require a fixed number of training statistics. A traditional approach for collecting education facts is to have human beings label a set of documents. An alternative technique to generating training information is distant supervision. In distant supervision, an already existing database is employed, along with Freebase or a site-particular database, to gather examples for the relation that is to be extracted further, and use of those examples to routinely generate our training data.

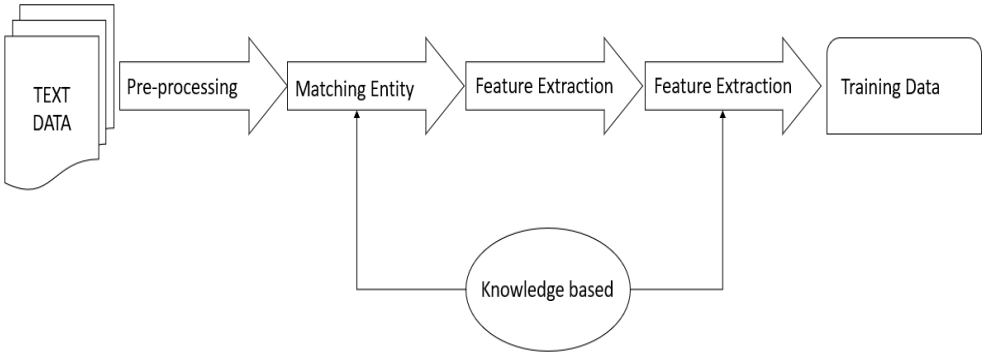

Fig. 11. Distant Supervision Data Flow techniques

The training data is processed data that is being fed to the algorithms. The training data itself needs to be processed because the raw text corpus may result in degradation of performance of the applied models. The most prominent process is pre-processing of data. The text here is cleaned and made simple for understanding for the models. Pre-processing includes removal of usernames, any hyperlinks and repeated letters that are present in text.

This data is matched with the corresponding entity represented as Entity Matching Process. The achieved data is used for feature extraction process to identify quantitative variables from the data that will be used for prediction. The data is labelled in the later stages because supervised learning algorithms are being applied over the data. The well-labelled data is now ready to be used as training data for XGBoost, Logistic Regression, Random Forest, Support Vector Machine (SVM) and Naïve Bayes algorithms for achieving high accuracy.

10 Results and Discussion

A publicly available Twitter dataset provided by Stanford University is used. Various feature extraction techniques were used to analyse the labeled datasets and compared the algorithms based on their accuracy. Employment of a framework in which a pre-processor is applied to raw sentences to make them more understandable.

Furthermore, several machine learning approaches train the models with attribute vectors and semantic analysis gives a large number of synonyms as well as similarities that help in determining the category of the given content.

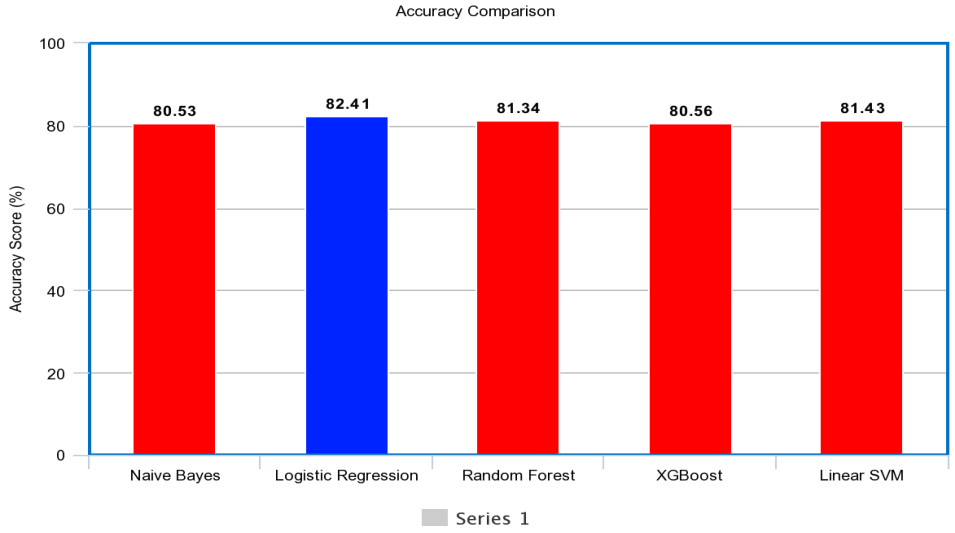

Fig. 12. Accuracy comparison of all used algorithms

11 Scope of Project

For classifying sentiment in tweets, machine learning algorithms work effectively. The precision could be enhanced further. A few points are listed down that can be useful in this case:

Semantics: The classifiers are used to categorize the sentiment behind a tweet. The feeling of the tweet may be determined from point of view of its interpretation. For instance, the

attitude is optimistic in the tweet "John wins against Bob". Here, for John, it's positive, whereas for Bob it's negative. Using a semantic role labeler can help you figure out which noun is mostly related to the verb; thus, classification is done accordingly.
Tweets from a given domain: For tweets across all domains, the best algorithm gives an accuracy of 83%. This is a substantial vocabulary. We believe our classifiers might perform better if they were limited to specific domains (such as movies).
Dealing with neutral tweets: Neutral tweets cannot be ignored in real-world applications. Neutral sentiment requires special consideration.
Internationalization: Only English sentences are examined yet, Twitter has a large worldwide following. It should be able to categorize the polarity of the tweet in other regional languages with the method that is being used.
Usage of emoticon data in testing set: Emoticons have been removed from the train data set before using in the test data set. This means that an emoticon feature in our test data has no effect on the classifier's classification. The emoticon features are quite valuable, so this should be handled.

12 Future Scope

The modelling techniques that are being used can further be improved by implementing few considerations:
1. Domain dependent: In numerous domains, the same sentence may have multiple meanings. For example, the word "unpredictable" is positive in the context of movies but it is detrimental in the context of car steering. The models should be well trained using huge data to understand domain context of any phrase.
2. Detecting Sarcasm: Sarcastic statements use positive words for expressing a negative view about a subject in a unique way. Natural Language Processing can be used to overcome these situations of sarcasm detection.
3. Internationalization: The models are limited to only predicting statements and phrases in English language only. The models can be trained in a way to support and predict accurately for regional languages also.

13 Conclusion

Using the dataset with a huge number of records has made our models more accurate in terms of predicting the correct polarity of the given tweet. Used several different classifiers in the model to achieve the highest accuracy possible. The Logistic Regression, however, has performed exceptionally well here. Twitter is one of the widely used platforms in today's world and with its growing usage of it, the model can help in prediction of early signs of depression, thus saving the youth for a better future. The models were successful in calculating and predicting accurate results from the testing data and understanding if the tweet was truly positive or negative. From the analysis and comparison, we conclude that the Logistic Regression model works best on the dataset with accuracy of 82.41% and an ROC count of 0.83. The model was best suitable for predicting correct outputs for its corresponding inputs. The Logistic regression model has achieved this due to the binary outcome of the predicting variable in the dataset. However, the Support Vector Machine model was also capable of achieving an accuracy of 81.43% and an ROC count of 0.82.

References

1. Yazdavar, A. H., Mahdavinejad, M. S., Bajaj, G., Romine, W., Sheth, A., Monadjemi, A. H., Thirunarayan, K., Meddar, J. M., Myers, A., Pathak, J., & Hitzler, P. (2020). Multimodal mental health analysis in social media. PloS one, 15(4), e0226248. https://doi.org/10.1371/journal.pone.0226248
2. Babu NV, Kanaga EGM. Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review. SN Comput Sci. 2022;3(1):74. doi: 10.1007/s42979-021-00958-1. Epub 2021 Nov 19. PMID: 34816124; PMCID: PMC8603338.
3. Uban, Ana Sabina et al. "An emotion and cognitive based analysis of mental health disorders from social media data." Future Gener. Comput. Syst. 124 (2021): 480-494.
4. AlSagri, Hatoon S. and Mourad Ykhlef. "Machine Learning-based Approach for Depression Detection in Twitter Using Content and Activity Features." ArXiv abs/2003.04763 (2020): n. pag.
5. Fu, Guanghui et al. "Distant Supervision for Mental Health Management in Social Media: Suicide Risk Classification System Development Study." Journal of medical Internet research vol. 23,8 e26119. 26 Aug. 2021, doi:10.2196/26119
6. Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. 2020. SenseMood: Depression Detection on Social Media. Proceedings of the 2020 International Conference on Multimedia Retrieval. Association for Computing Machinery, New York, NY, USA,407–411.DOI:https://doi.org/10.1145/3372278.3391932
7. Kawade, Dipak & Oza, Kavita. (2017). Sentiment Analysis: Machine Learning Approach. International Journal of Engineering and Technology. 9. 2183-2186. 10.21817/ijet/2017/v9i3/1709030151.
8. Raza, Hassan & Faizan, M. & Hamza, Ahsan & Mushtaq, Ahmed & Akhtar, Naeem. (2019). Scientific Text Sentiment Analysis using Machine Learning Techniques. International Journal of Advanced Computer Science and Applications. 10.10.14569/IJACSA.2019.0101222.
9. Nirag T. Bhatt, Asst. Prof. Saket J. Swarndeep (2020). Sentiment Analysis using Machine Learning Technique: A Literature Survey. IJERT. https://www.irjet.net/archives/V7/i12/IRJET-V7I12137.pdf
10. Mitra, Ayushi. (2020). Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset). Journal of Ubiquitous Computing and Communication Technologies. 2. 145-152.10.36548/jucct.2020.3.004.
11. Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A. Depression detection from social network data using machine learning techniques. Health Inf Sci Syst. 2018;6(1):8. Published 2018 Aug 27. doi:10.1007/s13755-018-0046-0
12. Kharde, Vishal & Sonawane, Sheetal. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications. 139. 5-15. 10.5120/ijca2016908625.
13. Asad, Nafiz & Pranto, Md. Appel Mahmud & Afreen, Sadia & Islam, Md. Maynul. (2019). Depression Detection by Analyzing Social Media Posts of User. 13-17-10.1109/SPICSCON48833.2019.9065101.
14. G. Geetha, G. Saranya, K. Chakrapani, J. G. Ponsam, M. Safa and S. Karpagaselvi, "Early Detection Of Depression from Social Media Data Using Machine Learning Algorithms," 2020 International Conference on Power, Energy, Control and

Transmission Systems (ICPECTS), 2020, pp. 1-6, doi: 10.1109/ICPECTS49113.2020.9336974.

15. K. A. Govindasamy and N. Palanichamy, "Depression Detection Using Machine Learning Techniques on Twitter Data," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 960- 966, doi: 10.1109/ICICCS51141.2021.9432203.
16. Chatterjee, Rinki & Gupta, Rajeev & Gupta, Bhavana. (2021). Depression Detection from Social Media Posts Using Multinomial Naive Theorem. IOP Conference Series: Materials Science and Engineering. 1022. 012095. 10.1088/1757-899X/1022/1/012095.
17. Cacheda, F., Fernandez, D., Novoa, F. J., & Carneiro, V. (2019). Early Detection of Depression: Social Network Analysis and Random Forest Techniques. Journal of medical Internet research, 21(6), e12554. https://doi.org/10.2196/12554
18. Chiong, Raymond & Budhi, Gregorius & Dhakal, Sandeep & Chiong, Fabian. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. Computers in Biology and Medicine. 135. 104499. 10.1016/j.compbiomed.2021.104499.
19. Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(2009), p.12.
20. J. Li and Y. Liang, "Refining Word Embeddings Based on Improved Genetic Algorithm for Sentiment Analysis," 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), 2020, pp. 213-216, doi: 10.1109/ITAIC49862.2020.9339058.
21. H. Silva, E. Andrade, D. Araújo and J. Dantas, "Sentiment Analysis of Tweets Related to SUS Before and During COVID-19 pandemic," in IEEE Latin America Transactions, vol. 20, no. 1, pp. 6-13, Jan. 2022, doi: 10.1109/TLA.2022.9662168.
22. A. J. Nair, V. G and A. Vinayak, "Comparative study of Twitter Sentiment On COVID - 19 Tweets," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1773-1778, doi: 10.1109/ICCMC51019.2021.9418320.
23. M. K. Patil, N. Chaudhari, B. V. Pawar and R. Bhavsar, "Exploring various emotion-shades for Marathi Sentiment Analysis," 2021 Asian Conference on Innovation in Technology (ASIANCON), 2021, pp. 1-5, doi: 10.1109/ASIANCON51346.2021.9544961.