A

Training Project Report

On

DEPRESSION DETECTION USING TEXT

### BACHELOR OF TECHNOLOGY

Degree

In

**Information Technology**

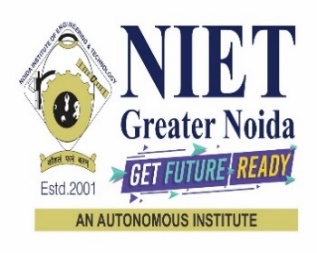
Under the Supervision of:

Dr. Vivek Kumar, Dr. Hitesh Singh

By:

Raqim Fraz, Prince Kumar Singh, Swati Mishra

Pankaj Kumar, Pushp Kunwar



NOIDA INSTITUTE OF ENGINEERING & TECHNOLOGY, GREATER

NOIDA

**Depression Detection Using Text**

Dr. Vivek Kumar, Dr. Hitesh Singh, Raqim Fraz, Prince Kumar Singh, Swati Mishra, Pankaj Kumar, Pushp Kunwar

**Abstracts-**

In recent years, there has been a lot of interest in a relatively new field of research called "depression detection using tweets from Twitter." Researchers have discovered that it is possible to detect and categories people who may be at risk for depression or who are currently exhibiting symptoms of depression using Twitter data.

As reported by the World Health Organization, depression is the most prevalent mental disorder, affecting more than 300 million individuals around the world. Depression is a major contributor to more from over three quarters of lives lost each year. However, because of self-deception upon that side of some patients and a total lack of knowledge of the issue, depression can go undiagnosed or untreated for a very long time. Failure to diagnose and treat the condition could worsen it, lowering quality of life and, in extreme cases, making it hard to maintain employment.

There is a consensus among many studies in the literature that social media sites, where users freely express their ideas and feelings, could be a valuable resource for tracking trends and health-related problems. Thanks to articles on social networking sites such as Twitter, researchers can examine a wide range of psychological problems and aspects of human behavior and attitude; however, on some other platforms, it is accurate that individuals on social media who seem to be depressed are highly improbable to utilize these terms directly; consequently, associated research in this area frequently relies on targeted keywords like "loneliness" and "troubleshoot" when using the data.

**Problem statement -**

The psychological health of the populace is under threat from stress, worry, and despair. There is still no uniform screening procedure that can be used consistently, and many caretakers, including parents and teachers, are ill-equipped to identify depression. Technology called stress detection systems is used to gauge and monitor an individual's level of stress.

**Keywords**:- Depression Detection, Machine Learning, SVM, MultinomialNaive Bias, Logistic Regression, ECG, WHO, Accuracy, Precision.

**Introduction**-

Millions of individuals all around the world are impacted by the common problem of depression. It's critical to develop efficient techniques for identifying and managing stress because it can significantly affect someone's health and wellbeing. A promising tool for enhancing stress detection and management is machine learning.

Variations in pulse rate, hypertension, and skin response ( gsr are just a few of the many ways that the human body reacts to depression. Many physiological sensors, including electrocardiogram (ECG) and electrodermal activity sensors, can be used to assess these changes (EDA).

**Literature Review**

One of the major global variables influencing suicide is depression. [1] However, because they are not correctly identified, a sizable percentage of depression cases get it untreated. Previous studies have shown that major depressive disorder patients' social media posts can be examined to see if they are presently depressed or at risk of becoming depressed. This study investigates whether social media users' posts, primarily those that lack obvious terms like "depression" or "diagnosis," may be utilized to accurately determine depressive symptoms using machine learning. We investigate various text processing and document featuring techniques along with classification algorithms based on machine learning, which would include single and ensemble models, in order to put forward a generalized strategy for depression.

Social networking sites have developed as a fantastic medium for consumers to communicate with pals who discuss their preferences and trade viewpoints, expressive images, and moving videos. [2] In order to better understand user sentiments and dispositions when utilizing these online communication platforms, it is now possible to analyze user sentiments and emotions in social network data. In this study, we use Twitter information collected from a shared digital source to analyze depression. We recommend using algorithms to investigate the effects of depression detection because it is efficient and scalable.

Fb, Twitter, and Instagram are just a few of the social media sites that have profoundly altered our world. [3] Individuals are more connected nowadays and have started to take on an online image of sorts. Unquestionably, social media has some incredible benefits, yet there are additionally unavoidable negatives. Several studies have revealed a connection between frequent usage of social media and increased melancholy. The objective of the current study is to detect Twitter users who might be sad based on the both their network behavior and tweets. We developed and assessed classifiers to identify whether a user is sad or not using features obtained from that user's network activity and tweets. The results showed that as more features are used, the precision increases. detection of mental diseases like depression. This study's main contribution is the exploration part of the features and its impact on detecting the depression level.

By 2030, depression illness is expected to overtake other disabilities as the primary cause of disability, according to the World Health Organization (WHO). [4] Despite the availability of skilled clinicians, medical, and psychological treatments for melancholy, people or families are hesitant to speak up or contact doctors about this condition for a variety of social reasons. The diagnosis of anxiety and depression involves numerous patient and family interviews, clinical analysis, and surveys, all of which take time and call for clinicians with the necessary training. Automation of stress detection is simple to implement in the current age of machine learning. However, automation ought to be more accessible, use fewer resources, and produce accurate outcomes. In this work, a classification model is trained using acoustic features to classify, the DIAC-WOZ information offered with the AVEC2016 challenge is taken into account. Using the COVAREP toolbox, prosodic, spectral, and voice control characteristics are extracted and fused. To overcome the class imbalance, SMOTE analysis is used, and the SVM method produces a depression classification model with 93% accuracy (DCM). A cureD game for Android The PHQ-8 and DCM questionnaires are used in the self-assessment tool called Deployed on Cloud. Underneath the supervision of an experienced psychiatrist, the application is evaluated using real-time data from 50 subjects, and an accurate of 90% is attained.

In the contemporary generation, depression has grown into a severe issue, as well as the number of those affected by it is rising daily. [5] Nonetheless, some of them are able to admit that they are depressed, while others are unaware of it. Nonetheless, the enormous growth of social networking sites is turning into a "diary" to share their emotional state. Using machine learning algorithms, many types of study have been done to identify sadness from user posts on social media. The researcher can determine whether or not people are experiencing depression by the data that is readily available on social media. Data may be correctly categorized into groups and separated into depressive and non-depressive data using machine learning algorithms. The study project under consideration tries to identify user depression using social media user data. The Nave Bayes classifier and the hybrid model NBTree are then fed the Twitter data in two separate ways. For the purpose of choosing the most effective algorithm to identify depression, the results will be compared based on the highest accuracy value. The outcomes reveal that both algorithms function equally by demonstrating the same level of accuracy.

The emergence of various social networking websites has made it possible for anyone to effortlessly generate, communicate, and transmit their concepts, thoughts, opinions, and emotions to millions of other individuals worldwide. [6] Mini computers and cellphones have entered human pockets as a result of technological advancements, and it is now incredibly simple to share your thoughts on any topic on social media sites like Facebook, Twitter. The use of social networking sites is increasing, and they have been utilized for a wide range of reasons, as a result of the enormous population explosion that has occurred and modern communications over the past ten years. A post-diagnosis analysis of users who have been diagnosed with depression is one such tool that their use may be investigated. In this study, we demonstrate how to use emotion hypotheses, machine learning methods, and text analysis methods on various social media platforms to observe and extract emotions from text to determine a person's degree of depression.

The sickness of depression is common in today's society. [7] Our entire way of thinking, as well as our emotional, cognitive, and regular behavioural habits, are altered and influenced by it. Around 264 million people were affected, and the number is rising daily. It primarily becomes a serious problem or a health matter when it lasts for an extended period of time. It causes the reliable person to falter as well, and in his last act, that person commits suicide. Although there are many factors that contribute to depression, social networking sites like Facebook, Twitter, and others play a significant part in making us feel worse. The majority of people in Asia communicate their feelings on Facebook, Twitter, and other chat platforms. We chose social media for our research project because of this. Although little research has been done, it is extremely rare to find depression in the Bengali community. As a result, there is now a high demand for it. The social media started a study based on tweets, numerous chat app responses, and projected depression postings and opinions. Bengali data was also collected. Several machine learning techniques have been employed to analyze these data, forecast depression, and for algorithmic purposes. The following algorithms have been used: Support vector machine, Random Forest, Decision Tree, K-Nearest Neighbors, Naive Bayes (Multinomial Naive Bayes), and Logistic Regression. The addition of those algorithms will produce the intended outcomes. Moreover, different algorithms send us different results as trends were common, but ultimately the precision was the same for all algorithms applied to our dataset.

Incredibly common mental illnesses affecting millions of people today in the modern age is depression. [8] The signs and symptoms of melancholy vary widely and frequently accompany other illnesses like bipolar disorder, Parkinson's, schizophrenia, etc. It is a severe mental illness that, if untreated, could develop into other health issues. At the moment, diagnosing melancholy in patients is solely dependent on the clinician's experience. Researchers in this area have incorporated various data modalities and computational techniques to help clinicians recognize the traits and categories depressed individuals. The purpose of this research is to provide solutions to some key questions about the publishing trend, data source, algorithms for machine learning, and dataset.

In the modern world, where the usage of the internet and social media is growing at an unprecedented rate, the timely screening or detection of feelings and emotions plays a crucial role. [9] The 300 million persons affected by psychological problems are particularly vulnerable. This is the rationale behind using original research articles to address the research issue. The key to potentially reducing the number of people impacted by this condition is early detection. In this work, a conventional dataset from online social media is analyzed, and detection is possible thanks to a machine learning method. This study suggests a machine-learning system to create an early depression prediction from their mode, which can be shielded from mental illness and suicide circumstances. To offer a high level of accuracy, the support vector machine and Naive Bayes method will be combined. There are numerous cumulative distribution parameters in the classification model that need to be dynamically categorized and identified. The features used for this identification or detection come from the textual, semantic, and writing content. The early prediction is being evaluated using several Deep Learning (DL) techniques. The method's sensitivity and accuracy are offering the crucial circumstances for both early and late detection.

Depression is a psychological disorder that affects over three hundred million humans worldwide. [10] A person who is depressed suffers from anxiety in day-to-day life, which affects that person in the relationship with their family and friends, leading to different diseases and in the worst-case death by suicide. With the growth of the social network, most of the people share their emotion, their feelings, their thoughts in social media. If their depression can be detected early by analyzing their post, then by taking necessary steps, a person can be saved from depression-related diseases or in the best case he can be saved from committing suicide. In this research work, a hybrid model has been proposed that can detect depression by analyzing user's textual posts. Deep learning algorithms were trained using the training data and then performance has been evaluated on the test data of the dataset of reddit which was published for the pilot piece of work, Early Detection of Depression in CLEF eRisk 2017. In particular, Bidirectional Long Short-Term Memory (BiLSTM) with different word embedding techniques and metadata features were proposed which gave good results.

**Description about dataset-**

Datasets are downloaded from Kaggle, we got two datasets

1. Mental health Twitter dataset – it is gathered from twits of twitter where post is labelled with 0 for “not depressed” and 1 for “depressed”.
2. Sentimental tweets – gathered from twitter post where post is labelled with 0 for “not depressed” and 1 for “depressed”.

Splitting dataset –

We used 80 percent data as training data and 20 percent as testing data and ten iterations set to train the model.hgd

**Preprocessing-**

* As such the dataset does not contains any null values, but we created some null values in some rows and follows the steps of cleaning and calculated mean values of columns to fill the missing data.
* Our data columns are in text format which is converted in numeric form.
* converted every uppercase to lowercase letter.
* Used beautiful soup to remove html tags, hyperlink, retweet, because our data was extracted from web so got in html format.
* Tokenize, stemming and lemmatization the text for which CountVerctorizer and TfidfTransformer is used

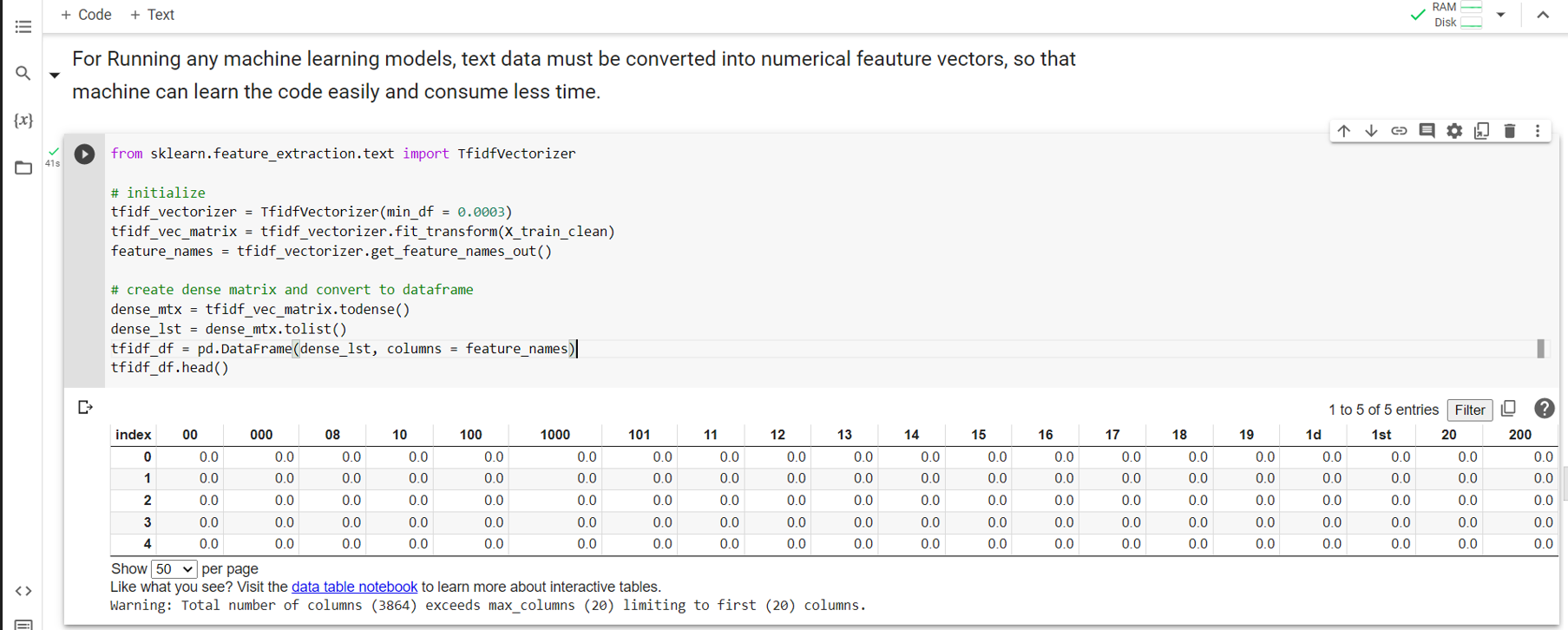


Fig 1. Preprocessing

**Data Exploration –**

While exploring data, we found that our data is labelled data and mental health data contain ten columns where unique id of user, post text, followers of users etc.

**Algorithm –**

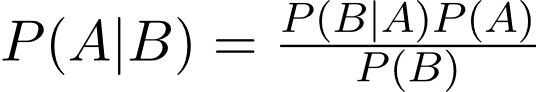
We use three different classification algorithms.

1. **MultiNomial Naïve Bayes** – It is considered one of the best Machine Learning classification algorithm. It is based on bayes theorem and predict the label of each piece of text based on previously labelled words.

The Naïve Bayes classifier is fast for both training the dataset and prediction of result.

Bayes theorem finds the probability of event occurring given the probability of another event that has already occurred.

Mathematical equation –



Code snippet

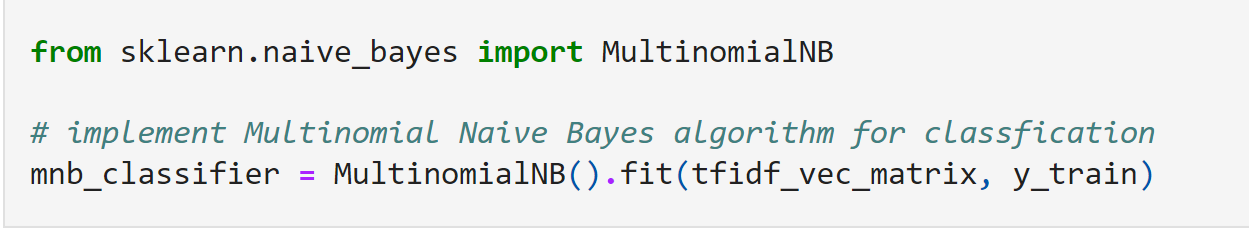


Fig 2. Naïve Bayes

1. **Support vector machine** – Support vector machine (SVM) is used for both classification and regression models. It finds the best hyperplane that can separate data into different classes by maximizing the margins of the closest data points for each class.

The kernel function in SVM classifier have the flexibility

to work with both linear and nonlinear data by transforming the data to become linearly separable.

In our algorithm we used linear kernel for classification.



Fig 3 – graph of SVM

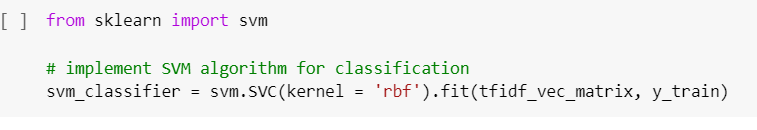


Fig 4 – code SVM

**Logistic Regression**- Logistic regression is used for binary classification and works by using logistic function to output a probability value between 0 and 1 for a data point.

Advantages for this model includes simplicity, speed, and interpretability. However, due to its linear nature it may not perform as well in more complex situations.

Snippet



Fig 5. Code Logistic Regression

We tried all three algorithm and got accuracy about 80%.

1.Multinomial Naïve Bayes Confusion Matrix

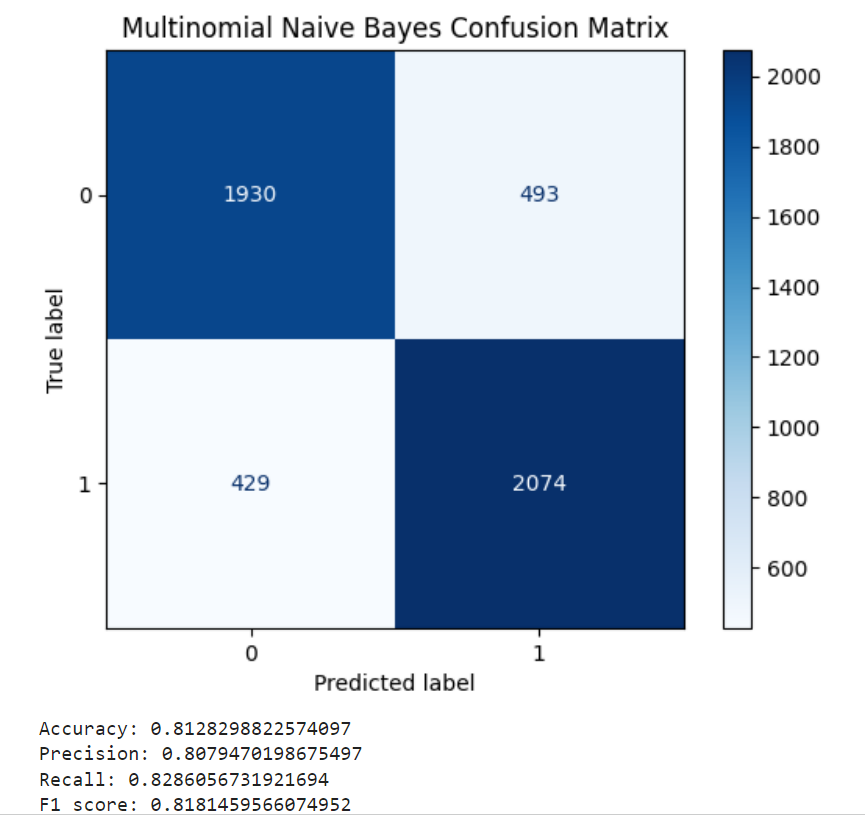


Fig 6. Confusion matrix Naïve Bayes

2.Support vector machine –

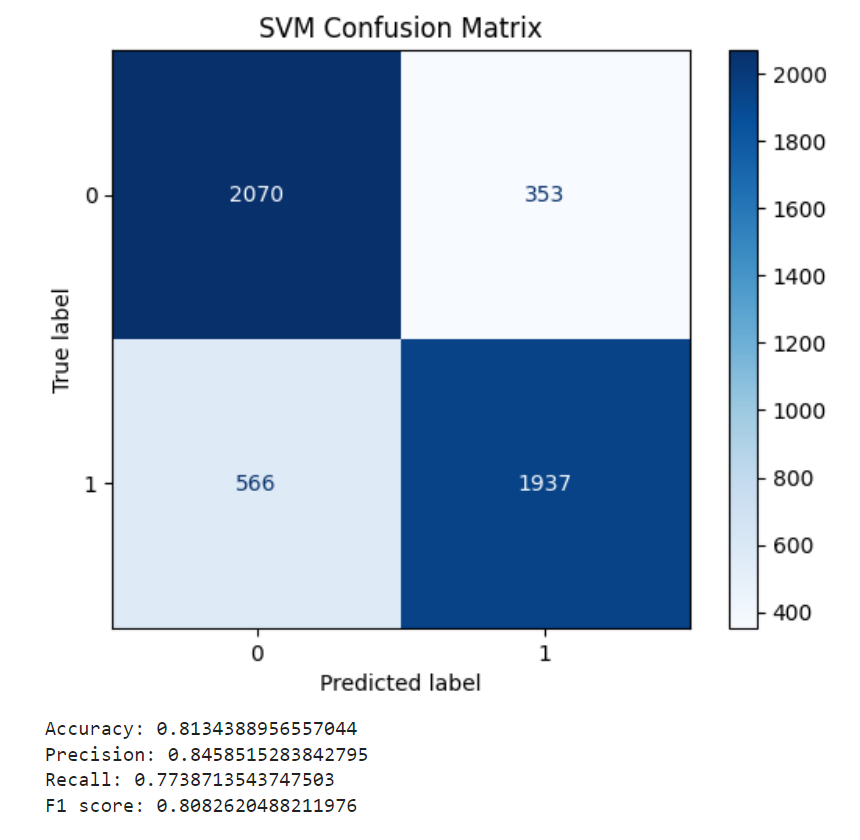


Fig 7. Confusion matrix SVM

3. Logistic Regression-

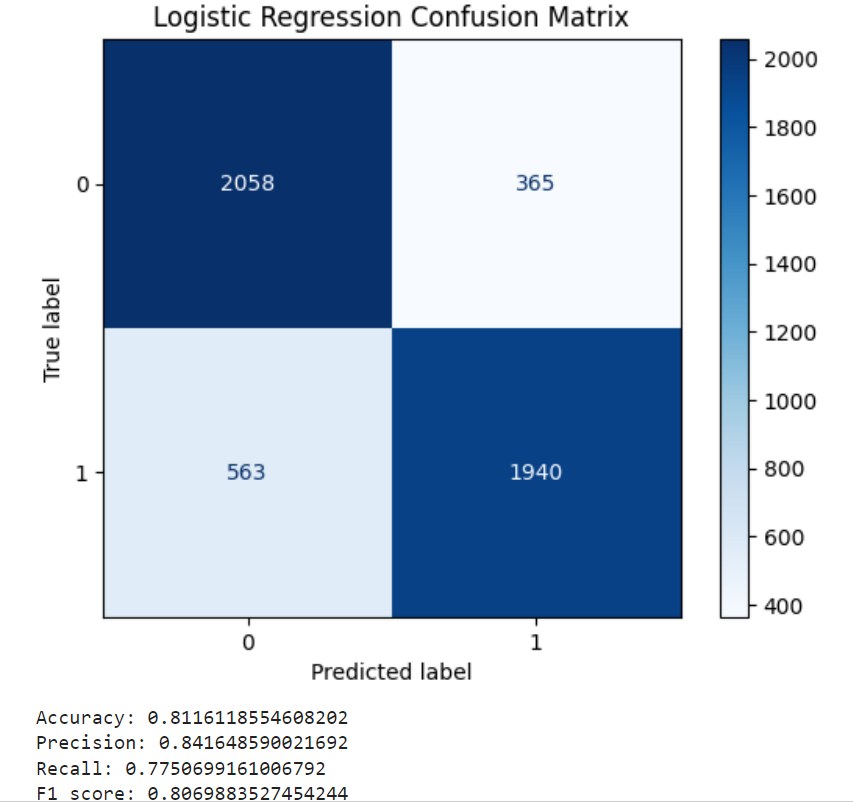


Fig 8. Confusion Matrix Logistic Regression

**Comparison-**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **MultiNomial Naïve Bayes** | **SVM** | **Logistic Regression** |
| **Accuracy** | 0.8072810935173907 | 0.8128298822574097 | 0.8007849506022466 |
| **Precision** | 0.8055774796106288 | 0.8388761467889908 | 0.8298057407944331 |
| **Recall** | 0.8171870829997331 | 0.7808913797704831 | 0.7638110488390712 |
| **F1 Score** | 0.8113407525172229 | 0.8088458880442295 | 0.7954419121734297 |

Hence, from the given chart we can analyze the accuracy, precision, recall, F1 score of all three algorithms, where all algorithms are giving almost equal results.

**Challenges & Success**

While applying multiple algorithm, we are getting accuracy of almost 81%. When we apply KNN algorithm got accuracy about 55 percent, but after changing the optimizer accuracy is improved.

Key findings

* Model is able to recognize 80 percent text correctly.
* SVM accuracy is 81.3 percent.
* Naïve Bayes accuracy is 81.2 percent.
* Logistic Regression accuracy is 81.1 percent.

Possible extension

The system can be extended from text detection to facial expression detection to detect the stress, as well as voice can be used to predict the stress.

Future Scope

The stress detection system can be embedded to company software where management will be able to know the level of their stress and depression, so that they can interact with their employee.

**References**

[1] Chiong, R., Budhi, G. S., Dhakal, S., & Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine*, *135*, 104499.

[2] AlSagri, H. S., & Ykhlef, M. (2020). Machine learning-based approach for depression detection in twitter using content and activity features. *IEICE Transactions on Information and Systems*, *103*(8), 1825-1832.

[3] Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., & Ulhaq, A. (2018). Depression detection from social network data using machine learning techniques. *Health information science and systems*, *6*, 1-12.

[4] Yalamanchili, B., Kota, N. S., Abbaraju, M. S., Nadella, V. S. S., & Alluri, S. V. (2020, February). Real-time acoustic based depression detection using machine learning techniques. In *2020 international conference on emerging trends in information technology and engineering (ic-ETITE)* (pp. 1-6). IEEE.

[5] Govindasamy, K. A., & Palanichamy, N. (2021, May). Depression detection using machine learning techniques on twitter data. In *2021 5th international conference on intelligent computing and control systems (ICICCS)* (pp. 960-966). IEEE.

[6] Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M., & Lee, S. (2017, October). Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. In *2017 international conference on information and communication technology convergence (ICTC)* (pp. 138-140). IEEE.

[7] Victor, D. B., Kawsher, J., Labib, M. S., & Latif, S. (2020, November). Machine learning techniques for depression analysis on social media-case study on bengali community. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1118-1126). IEEE.

[8] Bhadra, S., & Kumar, C. J. (2022). An insight into diagnosis of depression using machine learning techniques: a systematic review. *Current medical research and opinion*, *38*(5), 749-771.

[9] Smys, S., & Raj, J. S. (2021). Analysis of deep learning techniques for early detection of depression on social media network-a comparative study. *Journal of trends in Computer Science and Smart technology (TCSST)*, *3*(01), 24-39.

[10] Shah, F. M., Ahmed, F., Joy, S. K. S., Ahmed, S., Sadek, S., Shil, R., & Kabir, M. H. (2020, June). Early depression detection from social network using deep learning techniques. In *2020 IEEE Region 10 Symposium (TENSYMP)* (pp. 823-826). IEEE.