Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 6616- 6625

Research Article

Analysis of Various Machine Learning Models for Detecting Depression in Twitter Tweets

Shubham Tyagi[1], Rishabh Solanki[2], Adarsh Tiwari[3], Rohit Ray[4], Dr. Priyanka Paygude[5]
[1,2,3,4] Dept. of Information Technology, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India
[5]Assistant Professor, Dept. of Information Technology, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, India

Abstract

The main objective of this research paper is to help in identifying particular individual as depressive or not. In this research paper we have analysed and compared various machine learning approaches in comparison to our baseline machine learning approach Logistic Regression to identify a tweet as depressive or non-depressive. The tweets sentences are needed to be converted in to the form such as they could be fed to the machine learning models. Compared the two techniques used for this TF-IDF and Count Vectorizer. Other machine learning approaches thathave been analysed are Ridge Classifier, Multinomial Naive Bayes, Complement Naive Bayes, Stochastic Gradient Descent (SGD), Passive Aggressive Algorithms, Support Vector Classification, Voting classifier and Multi-layer Perceptron classifier. Voting classifier performed the best from all of them by giving approximately 80% of accuracy.

Keywords: Depression, Twitter, Mental health, Classification Models, Voting Classifier, Machine Learning.

1. Introduction

Depression in a common mental illness (there are more the 264 million people in the world with anxiety and depression) which has negative influences on person's productivity in work and leads to physical and emotional troubles such as obesity, loss of interest, less sleep and frequent thoughts of suicide. Majority of the people diagnosed with depression resides in the south East Asian region reflecting the large population of countries like India and China. Various studies show that Depression cases are more in female compared to male [5, 2]. Depression is different than common mood fluctuation at its worst depression can lead to suicide. According to WHO more than 800000 deaths happens every year because of suicide. The age group 15 to 29 is most affected by depression as deaths because of suicide is the 2nd leading cause of death in this age group [5].

Quality and effectiveness of mental health treatment have been increased greatly over the past few years, but these treatments would be of no use if they the person with mental illness doesn't seek any treatment. Less than 10% of mentally ill people receive the treatment. It is because people with mental illness are stigmatised. These stigmas are because of less awareness about the disease and due to socially

unacceptable behaviour [10].We can know about how the person is feeling by going through what he has been writing in his tweets. To do that we need to differentiate depressive suggestive tweets from non-depressive suggestive tweets. In past use of social media information to find if the particular social media user is depressed or not has been studied. For example, De Choudhary *et al*. in his research made the use of crowd sourcing to obtain the data from different users and then showed how to apply SVM (support vector machine) classifier to analyse the onset of depression in a particular user [4]. Tsugawa et al in his research of analysing twitter tweets he showed why we should use twitter tweets analysis instead of using questionnaires and how could we improve those analysis techniques by improving our feature. He further proposed of using bag of words with deep learning to improve the accuracy[11].

In this research paper, various machine learning models have been implemented to differentiate tweets into two categories depressive and non-depressive behavioural tweets and their performance has been depicted through figures showing true positive, true negative, false positive and false negative of each in models section. Explanation about how the data was created that was further fed to the different machine leaning models is in the dataset creation section. The methodology used has been depicted in figure 1 in methodology section. There is a baseline model with which we started off initially called logistic regression to which the other entire model's performances were compared and evaluated in the comparative analysis section. The best of them all has been concluded in the conclusion and future work section.

2. Methodology

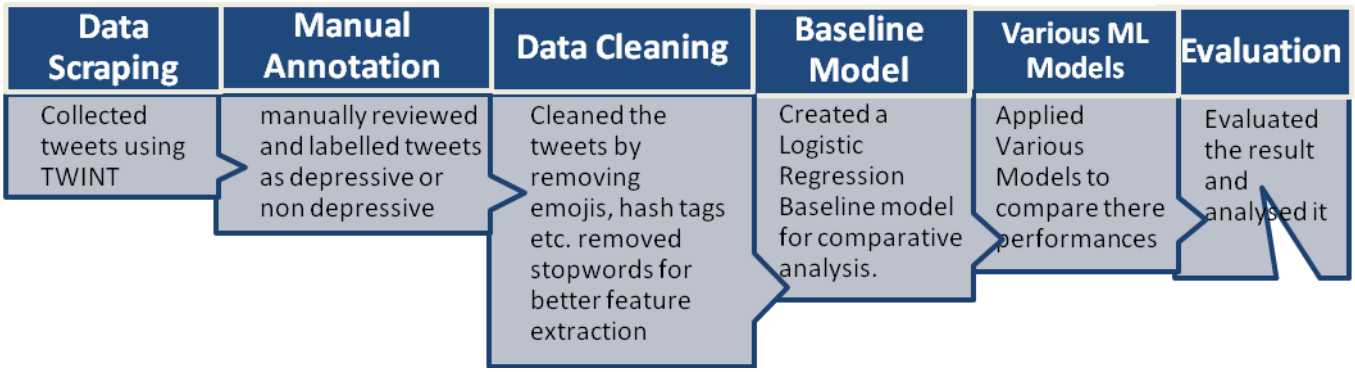

Figure 1 Process used to analyse tweet

3. Dataset Creation

a. Collecting Twitter Data

A Script was written and following steps were applied to identify potentially depressive behaviour in tweets:

i. Various tweets related to depression were collected using the TWINT data scraping library.
ii. Those tweets were which were either promotional or health awareness tweets from the dataset
iii. Removed tweets which contained more than three hash tags, contained @mentions, contained links and URLs.
iv. Those tweets with less than 10 characters, or 2 words.
v. All hash tags were removed from the tweets. This is because hash tags can be perceived as depression indicator by the models.

vi. The resultant data is saved into a file and is further reviewed for better performance.

b. Reviewing dataset

We manually reviewed the dataset which was created by the previous script, which contain filtered tweets that originally contained depressive hash tags. The dataset havea target column set to 1 by default, and we manually set the non-depressive entries to have target of 0, and also removed non-English tweets from the file. In the end we added random tweets which were either non-depressive or neutral to finally create our final dataset. The resulting dataset contain roughly 50-50 split of depressive and non-depressive tweets

4. Baseline

For our baseline model, we will be using Logistic Regression classifier. Before running the model, we need to transform our sentences i.e., Tweets into a representation that can be used as an input for our classifier. We will use 2 approaches to find our word representations.

1. **TF-IDF (term frequency–inverse document frequency)**: It vectorised word by taking into account the frequency of the word in the document and the frequency of those words in other documents [6].
2. **Count Vectorizer**: This approach is called the bag of words model, which means we are representing each sentence or document as a collection of discrete words and ignore grammar or the order in which words appear in a sentence. This will result in word vectors that will be as long as the size of vocabulary.

The baseline model gave better accuracy by using the TF-IDF transformer as compared to count Vectorizer. The model accuracy was further improved by using ngrams equal to 4 in the TF-IDF transformer.

5. Models

5.1. Ridge Classifier

first binary targets are converted to {-1, 1} by this classifier and then it treats the problem as a regression task[8]. It is different from the logistic regression on the basis of the loss function which in this case is l2 penalty as compared to cross-entropy of logistic regression.

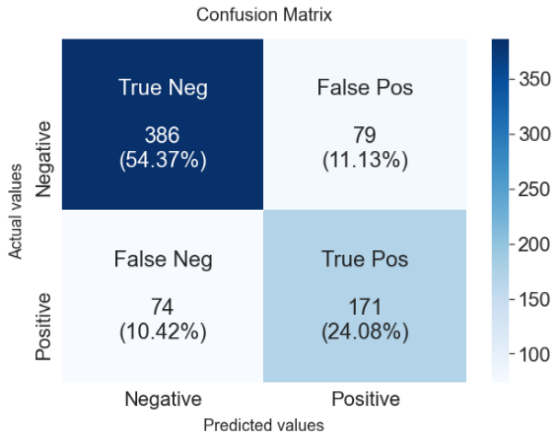

Figure 2 Ridge Classifier

5.2. Naive Bayes

Naïve Bayes classifier is the probabilistic classifier that predicts the class of a data based on the previous data by using probability measure. Between every pair of features with the given value of the class variable conditional independence is assumed by it.

Naïve are generally used in document classification and spam filter as they require less amount of data for predicting the parameters.

5.2.1. Multinomial Naive Bayes

When there is multinomial distributed data we use Multinomial Naïve Bayes. it is one of the two variants that is used for text classification. In this model main focus is on the frequency of the words.

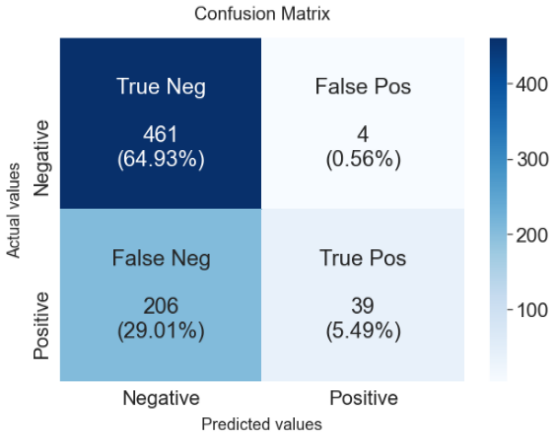

Figure 3 Multinomial Naive Bayes

5.2.2. Complement Naive Bayes

complement naive Bayes (CNB) algorithm is implemented by Complement Naïve Bayes. It is a variation of the standard multinomial naive Bayes (MNB). For imbalance data complement naïve bayes performs well and hence used in comparison to multinomial naïve bayes. CNB provide better result as compared to MNB in text classification problems.

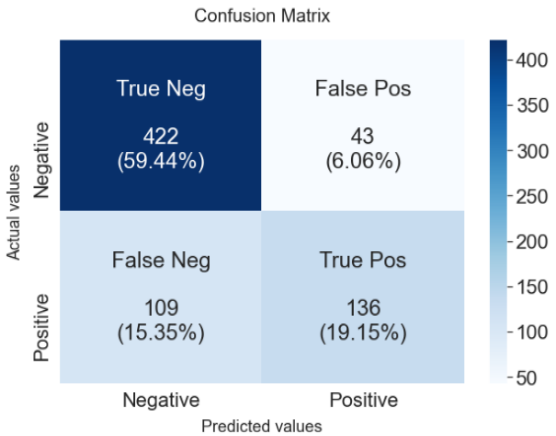

Figure 4: Complement Naive Bayes

5.3. Stochastic Gradient Descent – SGD

This is a optimisation model which is generally used in fine tuning the model parameters for the best fitting the model for predicted and actual results. This estimator implements regularized linear models with stochastic gradient descent (SGD) learning: the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule.[8]

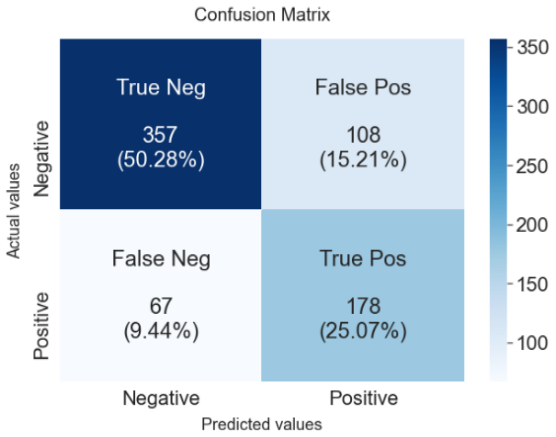

Figure 5: SGD

5.4. Passive Aggressive Algorithms

The passive-aggressive algorithms are a family of algorithms for large-scale learning. This model provides better performance if data comes in sequential order and model is updated step by step rather thanfeeding of data in a bulk.This is similar to the Perceptron.It basically takes a test case learn from that particular test case and then throws it away. So, if the output of that test case is as expected then it becomes passive (does not do anything) but if the outcome is not as expected it becomes aggressive [1,9,14].

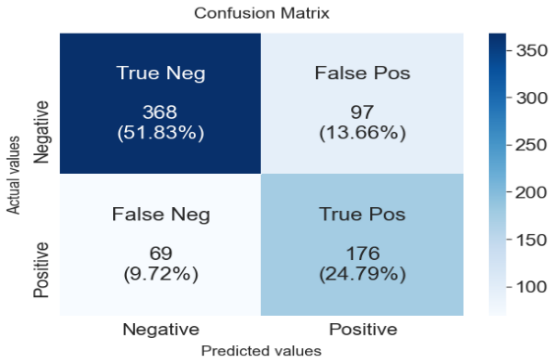

Figure 6: Passive Aggresive Classifier

5.5. Support Vector Classification.

Support vector machines (SVM) are another classification technique commonly used for classifying the data in N-dimensional space. The aim of the classifier is to identify the maximal margin hyperplane that classifies the data [14]. Structural risk minimization (SRM) principal is implemented by it which is basically used to ensure the minimization of the upper bound risk function, related to the generalization performance [13].

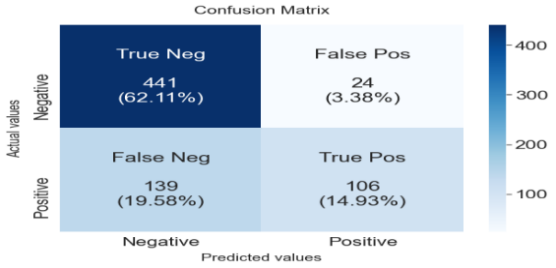

Figure 7: Support Vector Classification

5.6. Multi-layer Perceptron classifier

Three layers are present in a simple neural network model,i.e. input layer, output layer, and hidden layer. Input layers comprises of the n features given as input to the network, the input comprises of the value {x0, x1, x2,…,xm}. Now each input is multiplied by the corresponding weight. The weight determines how important the input is in the classification. All the inputs are provided to the summation function[8].

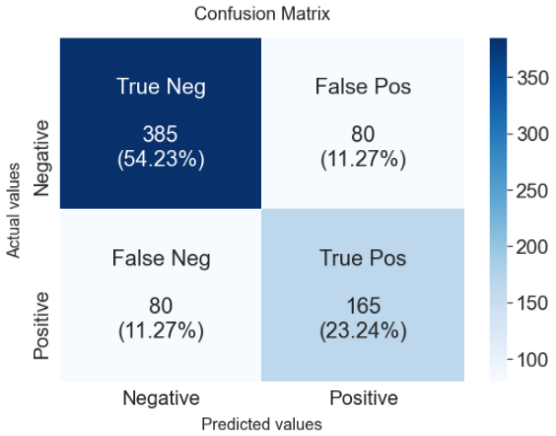

Figure 8: Multilayer Perceptron

5.7. Voting Classifier

In voting classifier rather than using a single model we combine various models to predict the label more accurately. In our model we used the soft voting parameter which rather than taking hard vote for prediction used the concept of average probabilities for predicting the label [12].

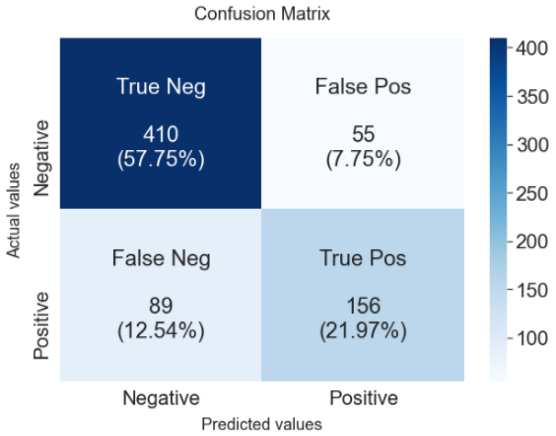

Figure 9: Voting Classifier

6. Evaluation

For each of the models we have calculated True Positive (TP), False Positive (FP), False Negative(FN) and True Negative(TN).

True Positive (TP) — Predicted and actual results are same for positive results.

True Negative (TN) — Predicted and actual results are same for negative results.

False Positive (FP) — Predicted and actual results are different for positive results.

False Negative (FN) — Predicted and actual results are different for negative results.

We have used the following parameters for evaluation,

Precision: It tells what percentage of prediction of the models were actually correct [3].

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

Recall: Out of the total positive, what percentage is predicted positive. It is the same as TPR (true positive rate) [3].

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

F1 Score: It is the HM of precision and recall. It takes both false positive and false negatives into account. Therefore, it performs well on an imbalanced dataset.[3]

$$F1\ score = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} = \frac{2*(Precision*Recall)}{(Precision+Recall)} \quad (3)$$

7. Comparative Analysis

As we can see in figure 10 that our baseline model accuracy was 76.2 % which is less as compared to thesame baseline implementations of other researches because of the explicit noise introduced in our dataset by us. While using the dataset from different researches our baseline model gives accuracy of around 97% because those datasets are top simple and biased. several experiments were conducted and compared performance across different models. As we can see from table 1 Complement Naïve Bayes Classifier gave the result of around 78% accuracy. Various model's accuracies were in the range of 75%-77% since single models were not giving high results we tried to assemble various models to get better result for that we used the voting classifier which can combine various classification model and by using it we got accuracy of 80%.

Table 1: Models Performance

Models	Logistic Regression	Ridge Classifier	Multinomial Naïve Bayes	Complement Naïve Bayes	SVC	Stochastic gradient Descent	Multilayer Perceptron	Passive Aggressive	Voting Classifier
Precision	0.77	0.76	0.80	0.78	0.79	0.76	0.74	0.75	0.78
Recall	0.70	0.76	0.58	0.73	0.69	0.74	0.75	0.75	0.77
F1	0.71	0.76	0.54	0.74	0.70	0.76	0.75	0.75	0.76
Accuracy	76.2	78.1	70.4	78.6	77	75.7	76.6	77.5	80

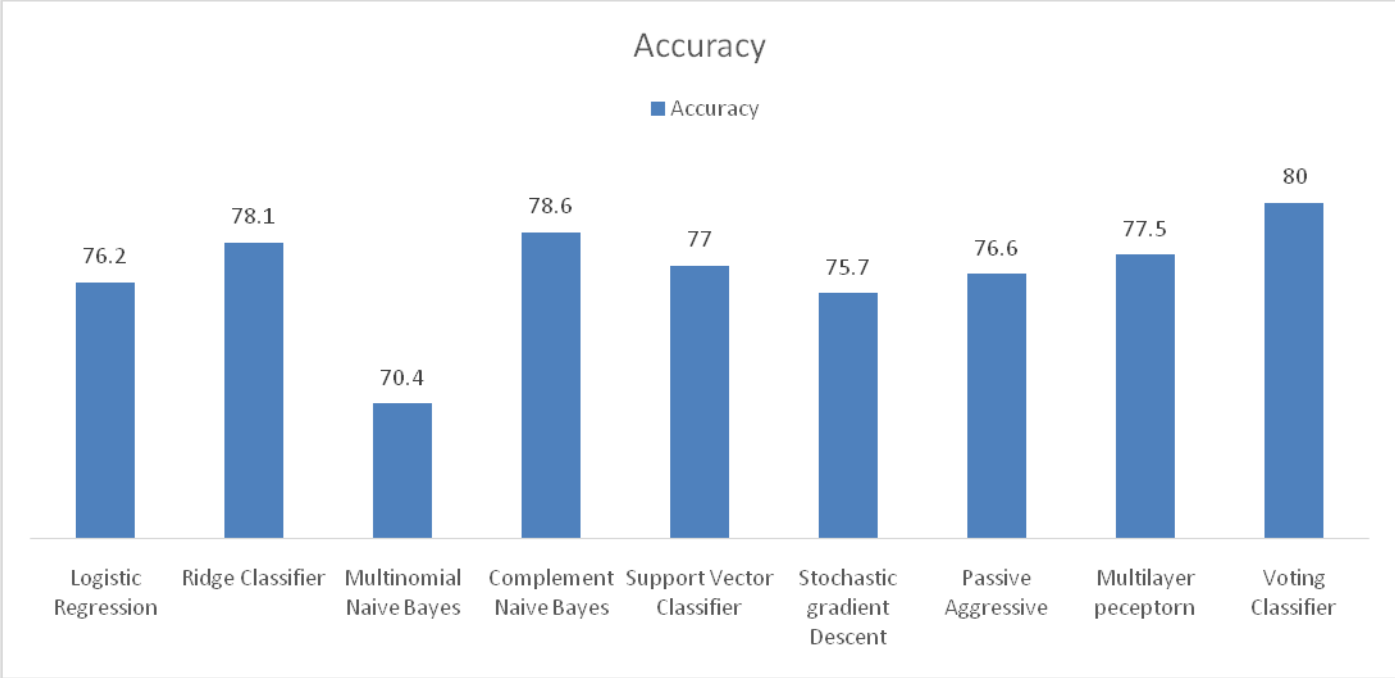

Figure 10 Accuracy Comparison

8. Conclusion and future work

After comparing performance evaluation of various models we can see in the results that while we do not have a very high accuracy score, the score is pretty solid, and it's convincing to say we have a good, working algorithm for detecting whether tweets are suggesting depressive behaviour or not. It was important to ensure that model doesn't learn to classify because of the bias in the dataset. The dataset used has reduced bias and ensures that the model doesn't learn to classify just based on a fixed set of frequent depression related words.

The performance of models can further be improved by using a larger manually annotated dataset if the data is annotated by someone from the mental health domain we can get better results. Having a dataset with all examples correctly labelled would give more confidence in the predictions done by model. Use state of the art liwc software to review the linguistic and emotional content of the tweets, and verify that the labels are correct. We can experiment further on our model development by trying various hyper parameters and see the impact on our model performance. Currently we removed the emojis and images from the dataset but in future emojis and images can also be used for detecting depressive behaviour as use of emojis is quite prominent

9. References

[1] Aayush Ranjan, ― Fake News Detection Using Machine Learning‖, Department Of Computer Science & Engineering Delhi Technological University, July 2018.

[2] A Structured Approach to Detecting and Treating Depression in Primary Care: VitalSign6 Project Manish K. Jha, Bruce D. Grannemann, Joseph M. Trombello, E. Will Clark, Sara Levinson Eidelman, Tiffany Lawson, Tracy L. Greer, A. John Rush, Madhukar H. Trivedi Ann Fam Med. 2019 Jul; 17(4): 326–335. doi: 10.1370/afm.2418 PMCID: PMC6827639

[3] B. P. Salmon, W. Kleynhans, C. P. Schwegmann and J. C. Olivier, "Proper comparison among methods using a confusion matrix," 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2015, pp. 3057-3060, doi: 10.1109/IGARSS.2015.7326461.

[4] De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. Predicting depression via social media. In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM'13) (July 2013), 128–137.

[5] Depression and Other Common Mental Disorders: Global Health Estimates. Geneva: WHO; 2017. Licence: CC BY-NC-SA 3.0 IGO

[6] 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

[7] M. Usman, S. Haris and A. C. M. Fong, "Prediction of Depression using Machine Learning Techniques: A Review of Existing Literature," 2020 IEEE 2nd International Workshop on System Biology and Biomedical Systems (SBBS), 2020, pp. 1-3, doi: 10.1109/SBBS50483.2020.9314940

[8] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011

[9] Shailesh-Dhama,―Detecting-Fake-News-withPython‖, Github, 2019

[10] Stigma of Mental Illness-1: Clinical reflections Amresh Shrivastava, Megan Johnston, Yves Bureau Mens Sana Monogr. 2012 Jan-Dec; 10(1): 70–84. doi: 10.4103/0973-1229.90181 PMCID: PMC3353607

[11] Tsugawa, S., Mogi, Y., Kikuchi, Y., Kishino, F., Fujita, K., Itoh, Y., and Ohsaki, H. On estimating depressive tendency of twitter users from their tweet data. In Proceedings of the 2nd International Workshop on Ambient Information Technologies (AMBIT'12) (Mar. 2013), 29–32.

[12] U. K. Kumar, M. B. S. Nikhil and K. Sumangali, "Prediction of breast cancer using voting classifier technique," 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2017, pp. 108-114, doi: 10.1109/ICSTM.2017.8089135.

[13] Vapnik, V.N. (1995) The Nature of Statistical Learning Theory. Springer, New York.

[14] W. S. Noble, "What is a support vector machine?," Nat. Biotechnol., vol. 24, no. 12, pp. 1565–1567, 2006.

[15]. R. S. Suryawanshi, A. Kadam, and D. R. Anekar, "Software defect prediction: A survey with machine learning approach," Int. J. Adv. Sci. Technol., vol. 29, no. 5, pp. 330–335, 2020.

[16]. A. Kurhade, J. Naveenkumar, and A. K. Kadam, "An experimental on top-k high utility itemset mining by efficient algorithm Tkowithtku," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 8 Special Issue 3, pp. 519–522, 2019.

[17]. Dr.S. D. Joshi, Dr. A. K. Kadam, Pritee

Hulule, “A Survey Novel Approach for Efficient Selection of Test Case Prioritization Techniques,” vol. 3085, no. 12, pp. 999–1001, 2018.

[18]. Dr. A. K. Kadam, Amruta Magdum, prof. Dr. S. D Joshi, “A Survey on Test Case Prioritization with Rate of Fault Detection,” Int. J. Res. Electron. Comput. Eng., vol. 6, no. 4, 2018.

[19]. A. N. Patil, A. Kadam, S. B. Wakurdekar, and S. D. Joshi, “Hybrid Approach of Code Analysis and Efforts Calculation for Software Reliability Growth Measurement and Cost Estimation,” Iioab J., vol. 9, no. 2, SI, pp. 116–120, 2018.

[20]. A. Magdum, S. D. Joshi, A. K. Kadam, and A. Sarda, “Test case ranking with rate of fault finding,” Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 8 Special Issue 3, pp. 462–464, 2019.

[21]. A. K. Kadam, S. D. Joshi, D. Bhattacharyya, and H.-J. Kim, “Software Superiority Achievement through Functional Point and Test Point Analysis,” Int. J. Softw. Eng. Its Appl., vol. 10, no. 11, pp. 181–192, 2016.

[22]. V. E. Pawar, A. K. Kadam, and S. D. Joshi, “Analysis of software reliability using testing time and testing coverage,” Int. J. Adv. Res. Comp. Sci. Manag. Stud, vol. 3, no. 5, pp. 143–148, 2015.