**ENSEMBLE MACHINE LEARNING MODELS BASED PREDICTIONS FOR** [**SENTIMENTAL ANALYSIS**](https://scholar.google.com/citations?view_op=view_citation&hl=en&user=eugyjMQAAAAJ&citation_for_view=eugyjMQAAAAJ:YsMSGLbcyi4C) **ON TWITTER DATA**

In current days web content comes from social media, multiple companies, different types of events, online products and personal data. This sentiment analysis predicts findings with the help of different methodologies. In this process input is so simple but deriving this information is too difficult. Internet data usage is increasing throughout the world, using this data for feedback purposes. Such type of data classification and organize was most difficult for sentiments. This feedback is most important for improving the business, gaining more profit and understanding the customer’s interest. Finally, from our research, Logistic regression accuracy is 92%, XGBoost accuracy is 90%, Decision trees predict 90% accuracy and Random forests predict 95.5% accuracy. Compared to the ensemble learning model, the Random Forest Tree model performs a higher accuracy rate than ensemble models.

**Keywords:** Ensemble Machine Learning Models, Predictions, [Sentimental Analysis](https://scholar.google.com/citations?view_op=view_citation&hl=en&user=eugyjMQAAAAJ&citation_for_view=eugyjMQAAAAJ:YsMSGLbcyi4C)

1. **INTRODUCTION**

Nowadays technology rules the world for any issue handled within a short time. Machine learning models are used for different problems. They also handle the sentiment analysis context. These models trained the data using examples [3]. After the training test, the data finally predicts the result based on input. Social networks and e-commerce generate huge amounts of data and it consist of valuable data. Maintenance of data with different methods like extraction, cleaning, preprocessing, training, testing and prediction [4].



**Figure 1:** Sample Diagram for sentiment analysis [1]

The robust features provide better results on sentiment analysis type of data. machine learning techniques provide better results compared to other technologies. Supervised, unsupervised and reinforcement learning and NLP techniques used for analysis [5]. The media like newspapers, magazines, social media, digital media and blogs generate content for the views of customers. This data provides the interests of end users and based on these views takes the right decisions for business development. Such type of data is available in multiple domains and useful for their ability to share ideas to overcome the problems. Identify the negative and positive content from the available data to find the sentiments of people quickly [6].

The succeeding paper, section 2 discusses the proposed system and architecture. Section 3 discusses the methodology. Section 4 states the outcome and analysis. Ending with the conclusion of the paper.

1. **PROPOSED SYSTEM AND ARCHITECTURE**

This sentiment analysis predicts findings with the help of different methodologies. In this process input is so simple but deriving this information is too difficult. Internet data usage is increasing throughout the world, using this data for feedback purposes. Such type of data classification and organize was most difficult for sentiments [7]. This feedback is most important for improving the business, gaining more profit and understanding the customer’s interest. We also directed wide-ranging tests on actual world openly obtainable datasets. The significances authenticate the improved presentation of our proposed ensemble learning techniques using the general performance metrics. We pronounce our projected framework followed by the explanation of used processes, datasets, and presentation assessment metrics.

**2.1 Performance Metrics [8]**

###### **2.1.1 Accuracy**

Accuracy is measured on behalf of the percentage of acceptably predicted observations. Accuracy **= (**1)

Generally, accuracy is a good sign of optimal prediction but in all cases, it is not possible. False positive and True negative consequences are also generated in our predictions. Along with accuracy, some other metrics are also used to find wrongly classified observations. The following metrics are precision, recall, and F1-score.

**2.1.2 Recall**

Recall signifies the true class, the entire number of confident orderings.

Accuracy **=**  (2)

**2.1.3 Precision**

Precision signifies the percentage of right positives to complete actions projected as true.

Accuracy **=** (3)

**2.1.4 F1-Score**

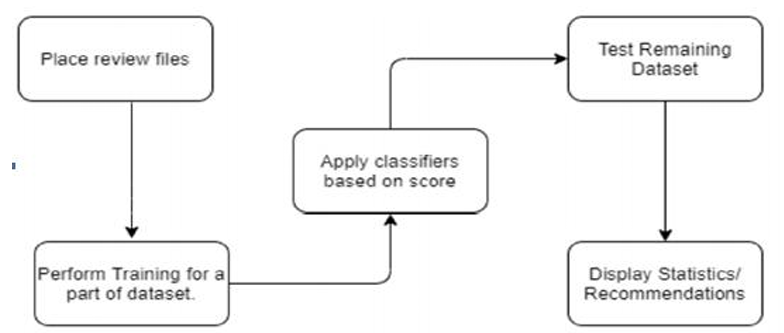
F1-score signifies between precision and recall. It computes the mean between each of the two. F1-score =2  (4)

* 1. **Architecture**

The proposed architecture discusses the multiple phases for predicting accurate results. The following phases are

1. Dataset (files)
2. Training dataset
3. Classifiers
4. Test Training Data
5. Statistical recommendations

The procedure of data starting from pre-processing, after pre-processing then train the data. After completion of training, apply classifiers and then test the data. Finally statistical recommendations displayed.



**Figure 2:** Block diagram of proposed system

**III. RESULTS AND ANALYSIS**

Twitter sentiment analysis using ensemble models, an ensemble regression modelling of the sentiment analysis of a database of tweets related to specific companies. This is structured as follows:

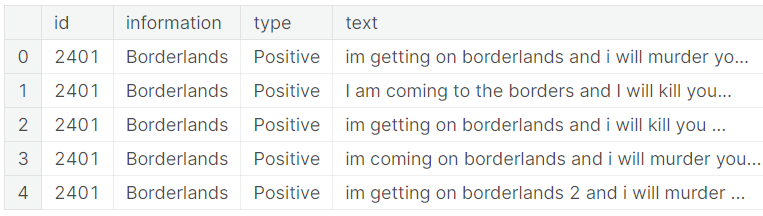
1. Initial data transformation
2. Plotting features
3. Text analysis
4. Decision Tree & Random Forest
5. Logistic Regression & XGBoost
6. Final Remarks

# **3.1 Initial Data Transformation**

As an initial approach, all the main libraries and functions were summarized in the following cell, focusing on data visualization, text analysis, text vectorization, and model building. Link code Then, the validation and train datasets were saved on two variables by using the function of read\_csv from pandas, where both didn't have a data header.

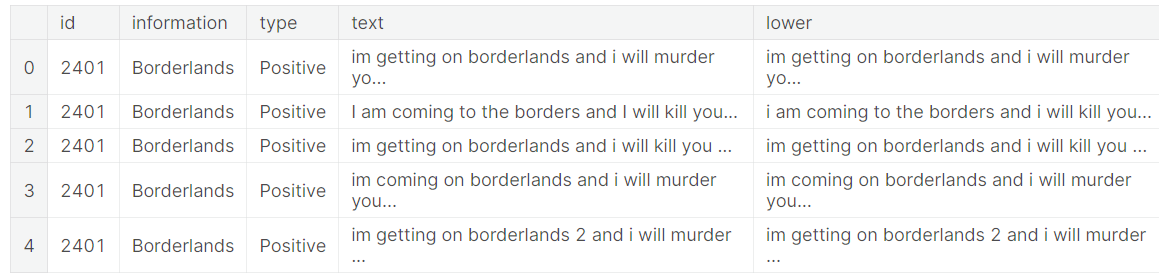
Later, the columns were renamed to represent the given data of tweets. But, with the first 5 rows analysis, it was recognized that positive sentiment was assigned to a "kill" thread related to a video game. Even with this in consideration, the modelling, in this case, will be the same as a traditional NLP project.

**Table 1:** Sample Tweets data [2]



Then, with the validation data, the information of the first 5 rows didn't show any unusual labelling. To prepare the data for the text analysis an additional row was created using the method of str.lower. However, as there were some texts with only numerical values (such as one that only had a 2 as the tweet) an additional function was used for transforming all the data to a string. Then, a regex expression erased the special characters as it is common to have digitisation problems on Twitter. The differences between the two text columns are presented in the next table.

**Table 2:** Differences between the two text columns



# **3. 2 Plotting features**

To identify the main words that were used per label, a word cloud was used to see which are the most important words on the train data. For example, on the positive label words such as love and game were mostly used alongside a wide variety of words classified as "good sentiments".

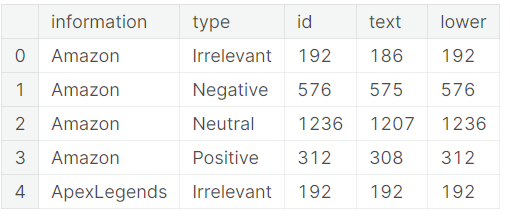


**Figure 3:** a wide variety of words

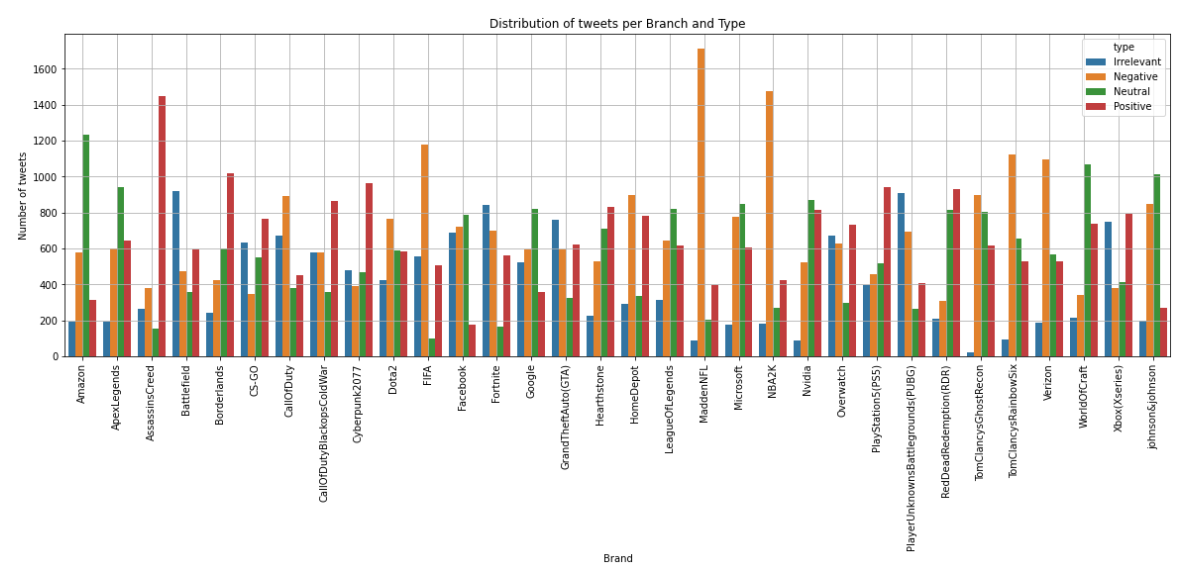
As for the negative tweets, some curse words were the most important while the names of some games and industries were also very used, such as face book and Instagram. The irrelevant tweets show a similar trend as negative ones, something that will impact the overall prediction performance.

Then, on the neutral side, there are almost no curse words, and the most important ones are different from the other 3 categories. Finally, in this section, the information was grouped by the brand (or the column information) to make a bar plot that shows the number of tweets for each one.

**Table 3:** The information was grouped by the brand



As an interesting fact, the number of modified texts coincides with the id. For this reason, as the ID is unique, the following bar plot shows that for games such as Madden NFL and NBA2K, the number of negative tweets is the highest, while on the other brands, the trend is different.



# **Figure 4:** For Madden NFL and NBA2K, the number of negative tweets is the highest, while on the other brands

# **3.3 Text analysis**

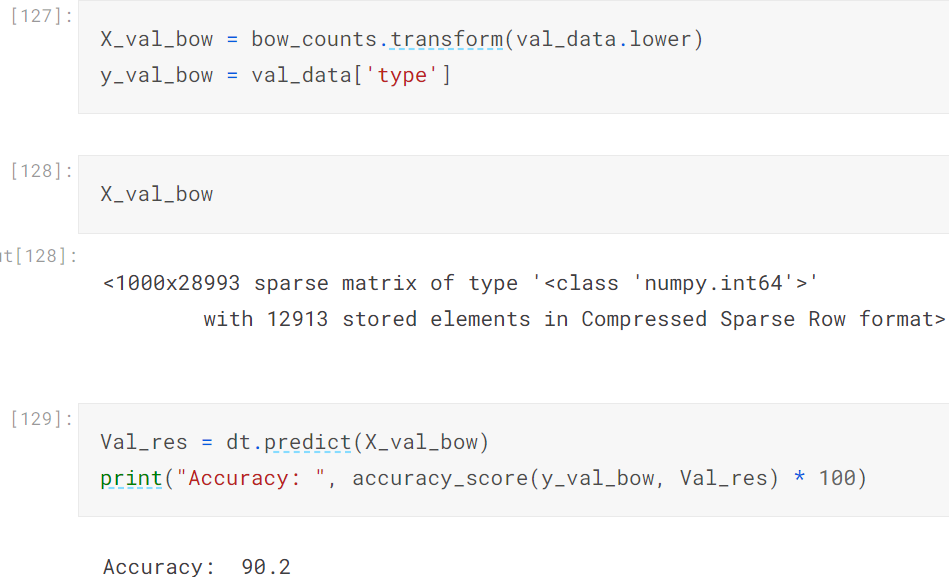
With the clean text, the initial number of unique tokens was counted to identify the model complexity. As presented, there are more than 30 thousand unique words.

# **3.4. Decision Tree & Random Forest**

In this section, we will be combining a decision tree and a random forest model to form an ensemble model. And calculate the accuracy of the model [13].

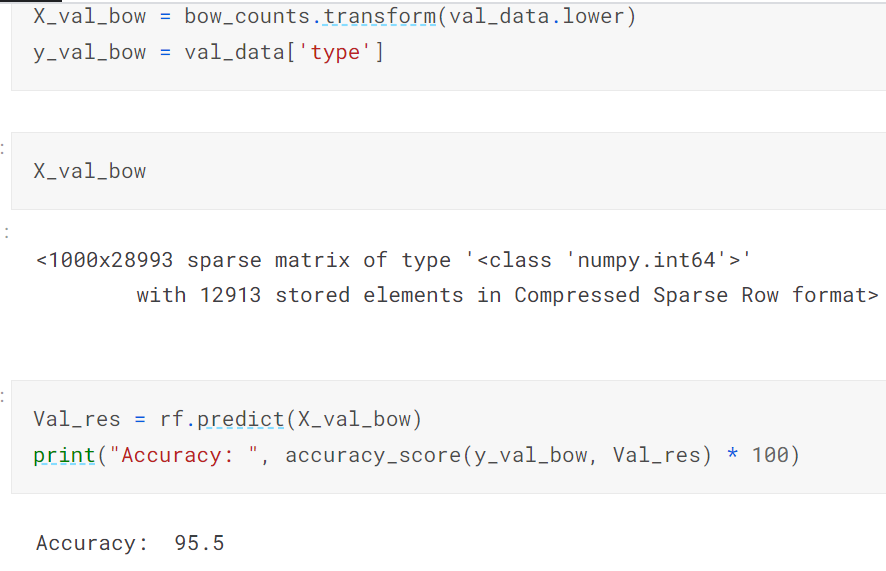
Then, the main data was split into train and test datasets alongside the encoding of the words by using the training dataset as a reference: The total number of tweets for each category shows that negative and positive are the most registered, while the irrelevant is the lowest.

### **3.4.1 Decision Tree Model**



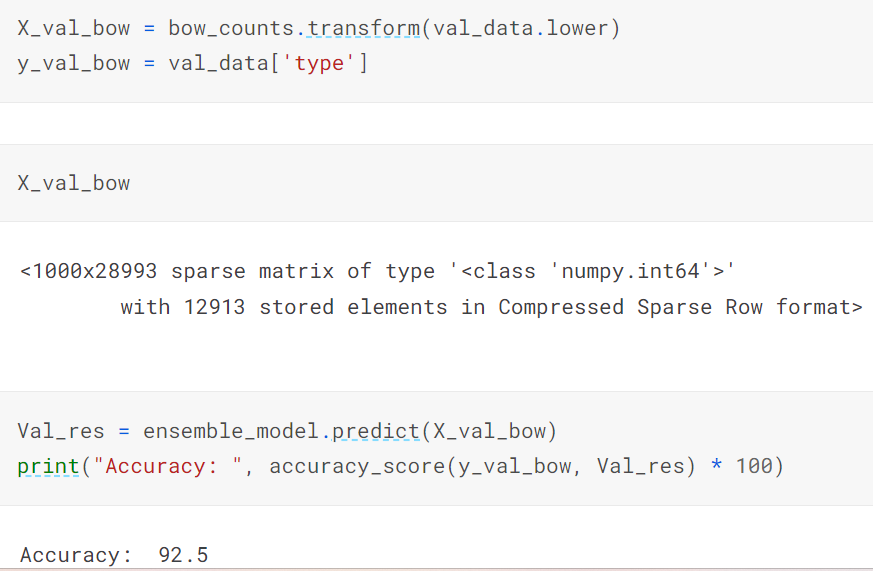
The decision trees can be used for sentiment analysis, achieving accuracies comparable to other machine learning methods like SVM and Random Forest. While decision trees might not always be the highest-performing algorithm, they are known for their interpretability and ease of understanding. The decision tree model's predicted accuracy is 90.2%

### **3.4.2 Random Forest Model**



Random Forest models can achieve high accuracy in sentiment analysis, often exceeding 80% accuracy. This ensemble learning method combines multiple decision trees, each trained on a random subset of the data, to improve performance and reduce overfitting The Random Forest tree model's predicted accuracy is 95.5%.

## **3.4.3 Ensemble model (Decision Tree + Random Forest)**



Ensemble learning, particularly using decision trees and Random Forest, significantly enhances sentiment analysis accuracy by leveraging the collective wisdom of multiple models. Random Forest, a powerful ensemble method, combines numerous decision trees, each trained on a different subset of the data and features, to make predictions. This approach mitigates overfitting and improves generalisation, leading to more reliable sentiment classification. The ensemble model (Hybrid Model) of decision tree and random forest tree predicted accuracy is 92.5%.

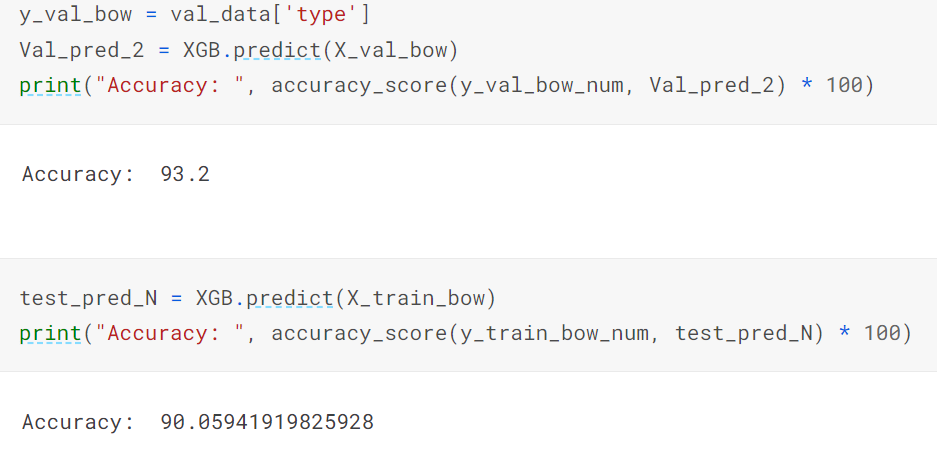
# **5.5 Logistic Regression & XGBoost**

### **3.5.1 Logistic Regression**

# 

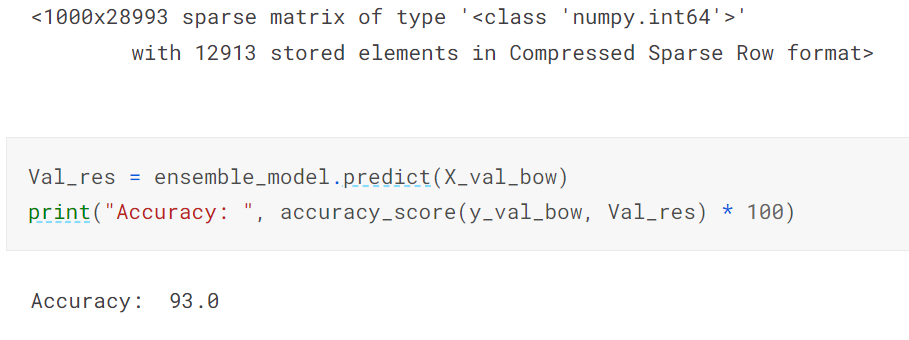
Logistic regression is a commonly used and often accurate machine learning algorithm for sentiment analysis, achieving results comparable to or exceeding other models like SVM. It excels in binary classification, making it well-suited for distinguishing between positive and negative sentiments. Logistic regression has been shown to perform well in various sentiment analysis tasks, including analysing airline tweets and general sentiment classification. The logistic Regression model predicted accuracy is 92%.

### **3.5.2 XGBoost Algorithm**



XGBoost, an implementation of Gradient Boosting, can achieve high accuracy in sentiment analysis. It's known for its high predictive accuracy and ability to handle complex relationships between features. XGBoost can perform well even with imbalanced data. The XGBoost model predicted accuracy is 90%.

## **3.5.3 Ensemble model (Logistic Regression + XGBoost)**



It was found that the accuracy is the lowest of all the modelling, so parameter tuning is required to improve the overall performance. we have used two combinations of machine learning algorithms for sentiment analysis of Twitter data. We can observe that the ensemble model (Decision Tree + Random Forest) yields 92% accuracy. And whereas the ensemble model (Logistic Regression + XGBoost) yields 93% accuracy.

Finally, from our research, Logistic regression accuracy is 92%, XGBoost accuracy is 90%, Decision Tree predict 90% accuracy, and Random Forest predicts 95.5% accuracy. Compared to the ensemble learning model, the Random Forest Tree model performs higher accuracy rate than the ensemble models. The following table gives details.

**Table 4:** Accuracy rate of the classifier

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Name of the Classifier** | **Accuracy Rate** |
| 1 | Logistic Regression | 92% |
| 2 | XGBoost Algorithm | 90% |
| 3 | Decision Tree | 90% |
| 4 | **Random Forest Tree** | **95.55** |
| 5 | Logistic Regression and XGBoost | 93% |
| 6 | Decision Tree & Random Forest | 92% |

**IV.CONCLUSION**

This sentiment analysis predicts findings with the help of different methodologies. In this process input is so simple, but deriving this information is too difficult. Internet data usage is increasing throughout the world, using this data is used for feedback purposes. Such a type of data classification and organize was most difficult for sentiments. This feedback is most important for improving the business, gaining more profit and understanding the customer’s interest. Finally, from our research, Logistic regression accuracy is 92%, XGBoost accuracy is 90%, Decision trees predict 90% accuracy, and Random forests predict 95.5% accuracy. Compared to the ensemble learning model, the Random Forest Tree model achieves a higher accuracy rate than the ensemble models.

**COMPETING INTERESTS DISCLAIMER:**

Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

**REFERENCES**

1. <https://monkeylearn.com/blog/sentiment-analysis-machine-learning/>
2. https://www.kaggle.com/bittlingmayer/amazon reviews#train.ft.txt.bz2
3. Subhani Shaik and Dr. Uppu Ravibabu “Classification of EMG Signal Analysis based on Curvelet Transform and Random Forest tree Method” Paper selected for Journal of Theoretical and Applied Information Technology (JATIT), Vol. 95, December 2017.
4. KP Surya Teja, Vigneswara Reddy and Subhani Shaik,” Flight Delay Prediction Using Machine Learning Algorithm XGBoost”, Journal of Advanced Research in Dynamical & Control Systems, Vol. 11, No. 5, 2019.
5. Subhani Shaik and Dr. Uppu Ravibabu, "Detection and Classification of Power Quality Disturbances Using curvelet Transform and Support Vector Machines", in the 5th IEEE International Conference on Information Communication and Embedded System (ICICES-2016) at S.A. Engineering College, Chennai, India on 25th -26th February 2016.
6. P. Pranathi, V. Revathi, P. Varshitha, Subhani Shaik and Sunil Bhutada,” Logistic Regression Based Cyber Harassment Identification”, Journal of Advances in Mathematics and Computer Science, Volume 38, Issue 8, Page 76-85, June 2023.
7. Neeraja, Anupam, Sriram, Subhani Shaik and V. Kakulapati,” Fraud Detection of AD Clicks Using Machine Learning Techniques”, Journal of Scientific Research and Reports, Volume 29, Issue 7, Page 84-89, June 2023.
8. Ms. Mamatha, Srinivasa Datta and Subhani Shaik,” Fake Profile Identification using Machine Learning Algorithms”, International Journal of Engineering Research and Applications (IJERA), Vol.. 11, Series-2, July 2021.
9. Mr. Sujan Reddy, Ms. Renu Sri andSubhani Shaik,” Sentimental Analysis using Logistic Regression”, International Journal of Engineering Research and Applications (IJERA), Vol.. 11, Series-2, July 2021.
10. Dr. R. Vijaya Kumar Reddy, Dr. Shaik Subhani, Dr. G. Rajesh Chandra, Dr. B. Srinivasa Rao,” Breast Cancer Prediction using Classification Techniques”, International Journal of Emerging Trends in Engineering Research, Vol. 8, No.9,2020.
11. Dr. Sunil Bhutada and Subhani Shaik, “IPL Match Prediction using Machine Learning”, IJAST, Vol.. 29, Issue 5, April 2020.
12. Shiva Keertan J and Subhani Shaik,” Machine Learning Algorithms for Oil Price Prediction”, International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-8, 2019.
13. Ch. Shravya, Pravallika and Subhani Shaik,” Breast cancer prediction using machine learning Techniques”, International Journal of Innovative Technology and Exploring Engineering, Vol. 8, Issue 6, 2019.