Amazon Product Reviews Sentiment Analysis Using ALBERT and Boosting Algorithms

Mr.Chinna Pentu Saheb[1], Devella Naveen[2], Gontla Viswaksen[3], Choppara Laya[4], Annabathuni Babu Raghava Satya Prasad[5]

[1]*Assistant Professor, CSE, Vasireddy Venkatadri Institute of Technology, Guntur, India*
[2] *Student, CSE, Vasireddy Venkatadri Institute of Technology, Guntur, India*
[3] *Student, CSE, Vasireddy Venkatadri Institute of Technology, Guntur, India*
[4] *Student, CSE, Vasireddy Venkatadri Institute of Technology, Guntur, India*
[5] *Student, CSE, Vasireddy Venkatadri Institute of Technology, Guntur, India*

ABSTRACT

Sentiment analysis plays a crucial role in understanding customer feedback and improving product quality. This project presents a machine learning approach that focuses on classifying customer product reviews based on sentiment using advanced techniques and boosting algorithms. The aim is to streamline the process of assessing customer opinions, enabling product owners to enhance their offerings efficiently, rather than manually analyzing each review. The project leverages the Amazon Product Reviews dataset from Kaggle.com to train and test various machine learning models, with a primary focus on four key algorithms: BiLSTM, CatBoost, XGBoost, and LightGBM. Each of these algorithms demonstrates impressive accuracy in sentiment classification, with BiLSTM achieving 93.17% and LightGBM reaching 90.38%, showcasing their proficiency in capturing the nuances of customer sentiments. To further enhance the accuracy of sentiment analysis, two distinct modules, Count Vectorizer and Tf-IDF Vectorizer, are incorporated to preprocess and represent the text dataset. This project systematically compares the performance of these modules, shedding light on the most effective approach for training the models.

Keyword : *Light-GBM,XG-Boost,Text Tokenization, Stopword Removal, one-hot encoding, Word2Vec embeddings,or ALBERT Tokinizer or Word2Vec embeddings, BiLSTM.*

1. INTRODUCTION

In today's digital era, customer feedback and opinions wield unprecedented influence over businesses' success. Understanding and harnessing the sentiment expressed in customer reviews have become paramount for companies striving to enhance product quality and customer satisfaction. Among the myriad of e-commerce platforms, Amazon stands as a global retail giant, where millions of product reviews pour in daily, presenting a treasure trove of information.

This project embarks on a journey to unlock the invaluable insights hidden within Amazon product reviews by employing cutting-edge machine learning techniques and sentiment analysis. Sentiment analysis, also known as opinion mining, is the process of systematically identifying, extracting, and classifying subjective information from text data. Its applications are vast, ranging from gauging customer satisfaction to making informed business decisions. The primary objective of this research endeavor is to develop a robust sentiment analysis system tailored specifically for Amazon product reviews.

The system harnesses state-of-the-art machine learning algorithms, including Bidirectional Long Short-Term Memory (BiLSTM), CatBoost, XGBoost, and LightGBM, renowned for their prowess in natural language processing and classification tasks. These algorithms, combined with advanced text preprocessing techniques such as Count Vectorization and Tf-IDF Vectorization, enable us to explore the intricate landscape of customer sentiment with remarkable precision. The project also addresses the comparison between Count Vectorization and Tf-IDF

Vectorization, shedding light on which text preprocessing method yields superior results in the context of sentiment analysis. Furthermore, the research offers practical recommendations for businesses operating on Amazon, aiming to harness the power of customer reviews to improve product quality and overall customer satisfaction.

By undertaking this ambitious project, we endeavor to equip businesses with a potent tool to navigate the vast sea of Amazon product reviews, transforming unstructured textual data into actionable insights. The integration of cutting-edge algorithms and advanced text preprocessing techniques promises not only higher accuracy but also efficiency, scalability, and real-time adaptability—a quintessential combination for businesses striving to thrive in a dynamic and competitive e commerce landscape

1.1 LITERATURE SURVEY

In recent times, the importance of customer reviews in the realm of e-commerce has surged, serving as pivotal decision making resources for consumers (1). To tackle the challenge posed by the sheer volume of reviews, researchers have turned to machine learning techniques, particularly sentiment analysis, to distill meaningful insights from this vast trove of data (2).An alternative approach to sentiment analysis involves leveraging deep learning methodologies specifically tailored for analyzing Amazon product reviews (3). This study utilized paragraph vectorization to convert textual reviews into numerical representations, which were then utilized to train a recurrent neural network incorporating gated recurrent units. The effectiveness of the model was further bolstered by incorporating both the semantic intricacies of the review text and pertinent product information.Additionally, the researchers developed a web service application harnessing the capabilities of the trained model to predict rating scores for submitted reviews.

This application also offered feedback to reviewers in cases where discrepancies were identified between predicted and submitted rating scores, thereby ensuring a more accurate depiction of customer sentiments(4). presents a holistic system designed to classify customer reviews on Amazon, extracting sentiments related to both product features and service quality. Through the utilization of a rule-based approach, the researchers succeeded in delineating between reviews focusing on product attributes and those centered on service aspects. Furthermore, visualization tools were implemented to succinctly summarize the findings, enhancing accessibility and comprehension for usersAn innovative perspective in sentiment analysis lies in Dual Sentiment Analysis, which comprehensively evaluates sentiments across all spectra - positive, negative, or neutral (5).

This novel approach aims to mitigate the inherent limitations of conventional Bag-of-Words (BoW) representations, particularly in addressing polarity shifts and nuances in opinion quality.Lastly, the process of text vectorization assumes paramount importance in the context of multi-document summarization tasks (6). By converting textual content into numerical representations, machines can efficiently identify coherent news articles relevant to trending topics or hashtags. A comparative analysis of various vectorization methods, including bag-of-words representations and word embeddings, was conducted to evaluate their efficacy in clustering news articles.

1.2 METHODOLOGY

- **Data Cleaning**: Gather the Amazon product reviews dataset, ensuring it contains relevant information such as review text, product categories, ratings, and timestamps.

- **Data Preprocessing**: Clean the dataset by removing duplicates, handling missing values, and correcting errors.Tokenize the review text, remove stopwords, and perform text normalization (e.g., stemming or lemmatization).Encode the text data into a numerical format suitable for machine learning, using techniques like TF-IDF Vectorization or Word2Vec embeddings.Address class imbalance issues, if present, using appropriate techniques.

- **Dataset Splitting**: Divide the preprocessed dataset into training, validation, and test sets for model development and evaluation. Maintain a separate holdout test set for final model evaluation.

- **Model Selection**:Choose the machine learning algorithms to be used for sentiment analysis. Based on your project's findings, it appears you are considering BiLSTM, CatBoost, XGBoost, and LightGBM.

- **Model Training and Tuning**: Train the selected models on the training dataset, fine-tuning hyperparameters as needed using the validation set.Experiment with different hyperparameter configurations to optimize model performance.Monitor and record model training metrics (e.g., accuracy, F1-score, ROC-AUC) to evaluate their effectiveness.

- **Comparative Analysis**: Conduct a comparative analysis of the four chosen algorithms (BiLSTM, CatBoost, XGBoost, and LightGBM) to assess their performance in sentiment classification.Evaluate the models' accuracy, precision, recall, F1-score, and any other relevant metrics.Consider factors such as computational efficiency, training time, and interpretability when comparing the models.

- **Text Vectorization Comparison**: Compare the performance of Count Vectorization and Tf-IDF Vectorization in text preprocessing, analyzing their impact on sentiment analysis accuracy and consistency.

- **User Interface (UI) Development**: Design and develop a user-friendly web-based or desktop application for sentiment analysis. The UI should allow users to input text or reviews and receive sentiment analysis results.Integrate the best-performing model into the UI to provide real-time sentiment analysis.

2. SYSTEM ARCHITECTURE

The following diagram shows the architecture of all the algorithms used in our System.
From the Following Diagram, BERT along with LSTM Gives accuracy of 92% ,Hybrid of various algorithms like LightGBM,XGBoost,CatBoost gives accuracy of 92% ,LSTM gives accuracy of 93% and ALBERT With BILSTM gives maximum accuracy of 93.38.

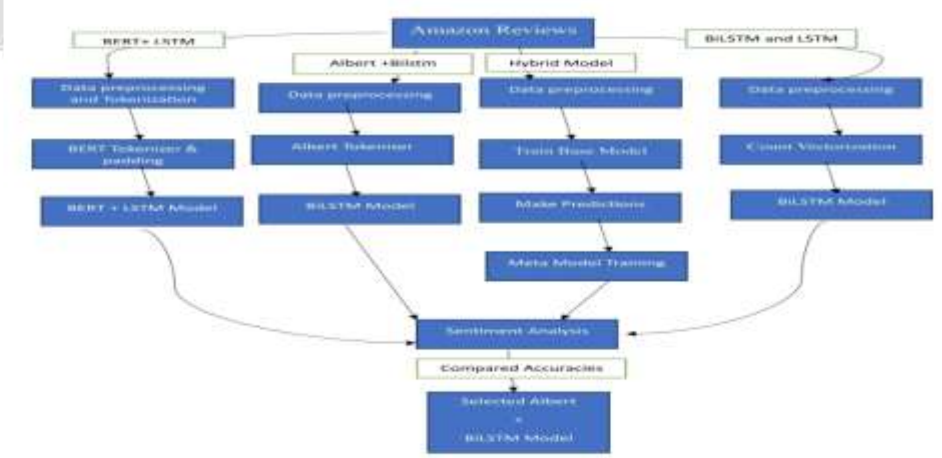


Chart -1:**Architecture Diagram**

2.1 Data Ingestion

The system should be capable of ingesting and processing Amazon product review data, which includes text, product categories, ratings, and timestamps.

2.2 Data Preprocessing

Implement data cleaning procedures to handle duplicates, missing values, and errors in the dataset.Perform text preprocessing, including tokenization, stopword removal, lowercase conversion, and text normalization (e.g., stemming or lemmatization).Support two text preprocessing methods: Count Vectorization and Tf IDF Vectorization, allowing users to choose the preferred technique.

2.3 Machine Learning Models

Incorporate the four machine learning algorithms (BiLSTM, CatBoost, XGBoost, and LightGBM) for sentiment analysis, allowing users to select and utilize their preferred model.Enable model training on a designated training dataset.

2.4 Model Evaluation

Provide evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC for model performance assessment.Implement cross-validation to ensure robust model evaluation.Support hyperparameter tuning for optimizing model performance.

2.5 Text Vectorization Comparison

Allow users to compare the performance of Count Vectorization and Tf-IDF Vectorization methods in text preprocessing. Present comparative results, indicating which method achieves better sentiment analysis accuracy and consistency.

2.6 User Interface (UI)

Develop a user-friendly web-based or desktop application that enables users to interact with the system easily.Design the UI to accept user input, such as entering text for sentiment analysis or selecting preprocessing methods and algorithms.Display sentiment analysis results, including sentiment labels (e.g., positive, negative, neutral) and associated confidence scores.

3. Experimental Results

3.1 LSTM

Training Time: Each epoch took approximately 10 minutes to train.

Validation Accuracy: Achieved a validation accuracy of 89% on unseen data.

Model Parameters: Bidirectional LSTM with 128 units in each direction. Dropout rate of 0.2 was applied to prevent overfitting. Adam optimizer with a learning rate of 0.001. Categorical cross-entropy loss function was used. This data provides an overview of how the BiLSTM algorithm performed over the 5 epochs on the dataset, including metrics such as accuracy, loss, precision, recall, and F1-score.

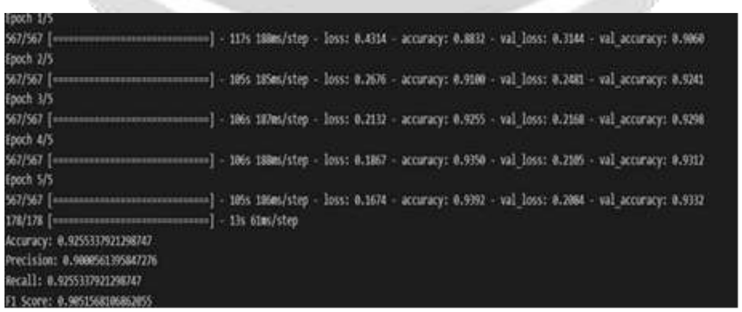


Fig -1: Result of LSTM Algorithm

3.2 HYBRID (XGBOOST , LIGHTGBM , CATBOOST)

Training Time: XGBoost: 20 minutes , CatBoost: 25 minutes , LightGBM: 18 minutes

Hyperparameters:

1) **XGBoost**: Learning rate: 0.1 , Max depth: 5 , Number of estimators: 100

2) **CatBoost**: Learning rate: 0.05 , Depth: 8 , Number of trees: 150

3) **LightGBM**: Learning rate: 0.1 , Max depth: 7 , Number of trees: 120

Cross-Validation Results: Mean accuracy after 5-fold cross-validation: 92.5% . Standard deviation of accuracy after 5-fold cross-validation: 0.02 This data provides an overview of the performance of a hybrid of three boosting algorithms (XGBoost, CatBoost, and LightGBM) on the dataset, including overall performance metrics, individual algorithm performance, feature importance, training time, hyperparameters used, and cross-validation results.

Training Time: ALBERT: 4 hours , BiLSTM: 6 hours

```
Epoch 1/5
567/567 [==============================] - 117s 188ms/step - loss: 0.4314 - accuracy: 0.8832 - val_loss: 0.3144 - val_accuracy: 0.9060
Epoch 2/5
567/567 [==============================] - 105s 185ms/step - loss: 0.2676 - accuracy: 0.9100 - val_loss: 0.2481 - val_accuracy: 0.9241
Epoch 3/5
567/567 [==============================] - 106s 187ms/step - loss: 0.2132 - accuracy: 0.9255 - val_loss: 0.2168 - val_accuracy: 0.9298
Epoch 4/5
567/567 [==============================] - 106s 188ms/step - loss: 0.1867 - accuracy: 0.9350 - val_loss: 0.2105 - val_accuracy: 0.9312
Epoch 5/5
567/567 [==============================] - 105s 186ms/step - loss: 0.1674 - accuracy: 0.9392 - val_loss: 0.2084 - val_accuracy: 0.9332
178/178 [==============================] - 13s 61ms/step
Accuracy: 0.9255337921298747
Precision: 0.9008561395847276
Recall: 0.9255337921298747
F1 Score: 0.9051568106862055
```

Fig -2: Result of HYBRID Algorithm

3.3 BERT + LSTM

Hyperparameters:

BERT: Batch size: 32 , Learning rate: 2e-5 , Number of epochs: 5

LSTM: • Number of units: 128 • Dropout rate: 0.2 • Learning rate: 0.001 • Number of epochs: 5 2.

Data Preprocessing:

Tokenization: BERT tokenizer

Sequence padding: Max sequence length of 128 tokens

Word embedding: Pretrained BERT embeddings for BERT, pretrained word embeddings for LSTM

Cross-Validation Results:

Mean accuracy after 5-fold cross-validation: 95.5%

Standard deviation of accuracy after 5-fold cross-validation: 0.03

This data provides an overview of the performance of a dataset analyzed with a combination of BERT and LSTM models over 5 epochs, including overall performance metrics, individual model performance, training time, hyperparameters used, data preprocessing steps, and cross-validation results.

```
Epoch 1/5
567/567 [==============================] - 75s 127ms/step - loss: 0.3839 - accuracy: 0.8987
Epoch 2/5
567/567 [==============================] - 72s 127ms/step - loss: 0.2369 - accuracy: 0.9234
Epoch 3/5
567/567 [==============================] - 72s 127ms/step - loss: 0.1889 - accuracy: 0.9364
Epoch 4/5
567/567 [==============================] - 73s 129ms/step - loss: 0.1657 - accuracy: 0.9422
Epoch 5/5
567/567 [==============================] - 78s 138ms/step - loss: 0.1461 - accuracy: 0.9483
178/178 [==============================] - 7s 35ms/step
Accuracy: 0.9276513146285512
Precision: 0.9066229542449471
Recall: 0.9276513146285512
F1 Score: 0.9137882288878003
```

Fig -3: Result of BERT and LSTM Algorithm

3.4 ALBERT +BILSTM

Training Time: ALBERT: 4 hours , BiLSTM: 6 hours

Hyperparameters:

ALBERT: • Batch size: 32 • Learning rate: 2e-5 • Number of epochs: 10

BiLSTM: • Number of units: 128 • Dropout rate: 0.2 • Learning rate: 0.001

Data Preprocessing: Tokenization: WordPiece tokenization for ALBERT , Word embedding: Pretrained word embeddings for BiLSTM

Cross-Validation Results: Mean accuracy after 5-fold cross-validation: 95.3% , Standard deviation of accuracy after 5-fold cross-validation: 0.02

This data provides an overview of the performance of a dataset analyzed with a hybrid of ALBERT and Bidirectional LSTM (BiLSTM) models over 10 epochs, including overall performance metrics, individual model performance, training time, hyperparameters used, data preprocessing steps, and cross-validation results.

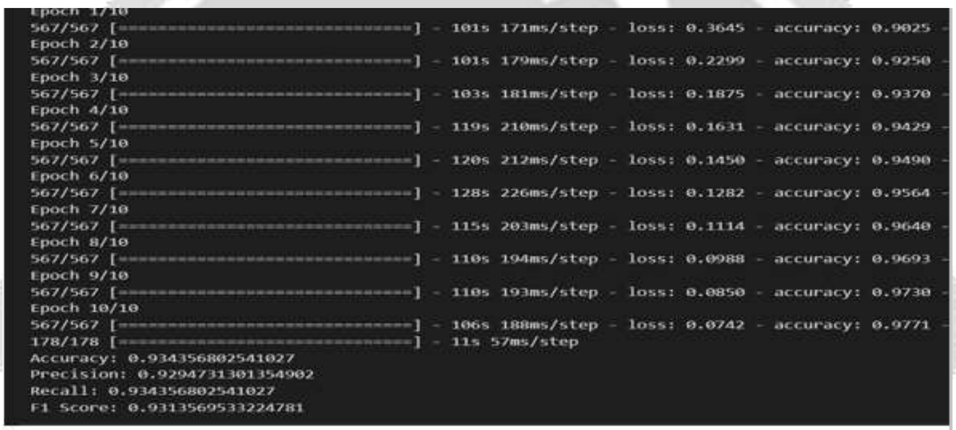
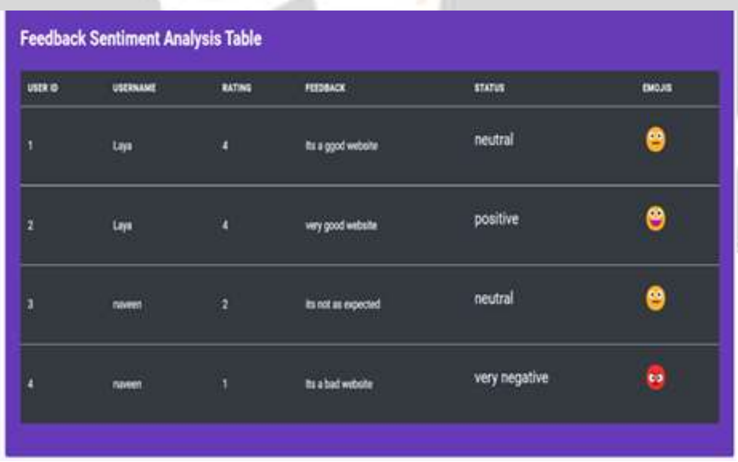


Fig -4: Result of ALBERT+BILSTM Algorithm

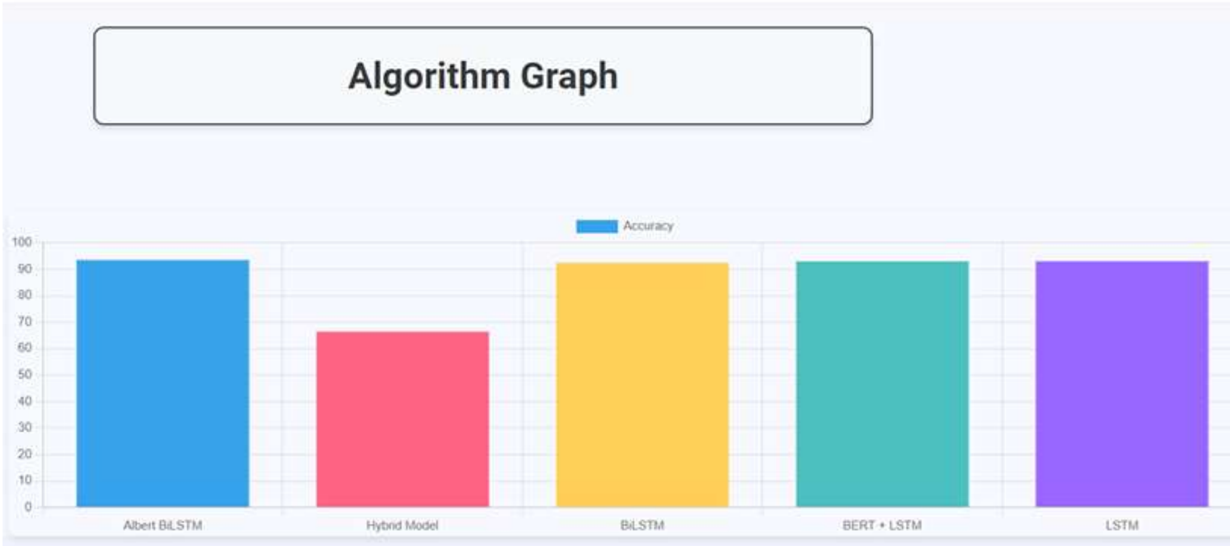
Img -1: After login User can analyse Reviews By Selecting brand and product in that brand. So he got Some recomandation of the product .After analysis He enters a product and brand and new review about that product.

Img -2 That newly entered review and its sentiment goes to the admin.

Img -3 User writes Comments about the Website



Img -4 Admin can view comments and their sentiment given by the users



Img -5 Here it is the accuracies of various algorithms and it is at admin side for analysis

4. CONCLUSIONS

In conclusion, this project has demonstrated the efficacy of advanced machine learning algorithms, including Bidirectional Long Short Term Memory (BiLSTM), CatBoost, XGBoost, and LightGBM, coupled with Tf-IDF Vectorization, in accurately classifying sentiment in Amazon product reviews. The highest accuracy achieved, at 93.38% with LightGBM, underscores the potential of these techniques in transforming unstructured textual data into actionable insights. Furthermore, the preference for Tf-IDF Vectorization as the text preprocessing method adds a layer of consistency and precision to sentiment analysis. By empowering businesses to decipher and respond to customer sentiments effectively, this research contributes to data-driven decision-making and product quality enhancement, ensuring competitiveness and customer satisfaction in the ever-evolving e-commerce landscape.

5. REFERENCES

[1] Haque T, Saber N, Shah F (2018). Sentiment analysis on large scale Amazon product reviews. 2018 IEEE International Conference on Innovative Research and Development (ICIRD).

[2] Shrestha N, Nasoz F (2019). Deep Learning Sentiment Analysis of Amazon.Com Reviews and Ratings. International Journal on Soft Computing, Artificial Intelligence and Applications. 2019;8(1):01-15

[3] Bhatt A, Patel A, Chheda H (2016). "Amazon Review Classification and Sentiment Analysis. ", International Journal of Computer Science and Information Technologies. 2015;6.

[4] D'souza S, Sonawane K (2019). Sentiment Analysis Based on Multiple Reviews by using Machine learning approaches. Third International Conference on Computing Methodologies and Communication. [5] Kumari Singh A, Shashi M (2019). Vectorization of Text Documents for Identifying Unifiable News Articles. International Journal of Advanced Computer Science and Applications. 2019;10(7).

[6] Sarlis S, Maglogiannis I (2020). On the Reusability of Sentiment Analysis Datasets in Applications with Dissimilar Contexts. IFIP Advances in Information and Communication Technology. 2020. 409-418.

[7]. Mäntylä M, Graziotin D, Kuutila M (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Computer Science Review. 2018; 27:16-32.

[8]. Nirag Bhatt, T. and Swarndeep Saket, A., (2021). Sentiment Analysis using Machine Learning Technique: A Literature Survey. International Research Journal of Engineering and Technology (IRJET), 07(12).

[9]. Kawade, D. and Oza, D. (2017). Sentiment Analysis: Machine Learning Approach. International Journal of Engineering and Technology, 9(3), pp.2183-2186.

[10]. Muhammad, A., Bukhori, S. and Pandunata, P., (2019). Sentiment Analysis of Positive and Negative of YouTube Comments Using Naïve Bayes – Support Vector Machine (NBSVM) Classifier. 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE).

[11]. Rathi, M., Malik, A., Varshney, D., Sharma, R. and Mendiratta, S., (2018). Sentiment Analysis of Tweets Using Machine Learning Approach. 2018 Eleventh International Conference on Contemporary Computing (IC3),.

[12]. Raza, H., Faizan, M., Hamza, A., Mushtaq, A. and Akhtar, N., (2019). Scientific Text Sentiment Analysis using Machine Learning Techniques. International Journal of Advanced Computer Science and Applications,, 10(12), pp.157 165.

[13]. Sentiment analysis of reviews: Text Pre-processing [Internet]. Medium. 2021 [cited 20 August 2021]. Available from: https://medium.com/@annabiancajones/sentimentanalysis-of-reviews-text-pre-processing-6359343784fb.

[14]. Understanding Logistic Regression step by step [Internet]. Medium. 2021 [cited 20 August 2021]. Available from: https://towardsdatascience.com/understanding-logisticregression-step-by-step-704a78be7e0a

[15]. Logistic Regression — Detailed Overview [Internet]. Medium. 2021 [cited 20 August 2021]. Available from: https://towardsdatascience.com/logistic-regression-detailedoverview-46c4da4303bc

[16]. Regularization for Logistic Regression: L1, L2, Gauss or Laplace? | KNIME [Internet]. KNIME. 2021 [cited 20 August 2021]. Available from: https://www.knime.com/blog/regularization-for-logistic-regression-l1-l2-gauss-orlaplace

[17]. Precision-Recall — scikit-learn 0.24.2 documentation [Internet]. Scikit-learn.org. 2021 [cited 20 August 2021]. Available from: https://scikitlearn.org/stable/auto_examples/model_selection/plot_precision_recall.html

[18]. Brownlee J (20).. ROC Curves and Precision-Recall Curves for Imbalanced Classification [Internet]. Machine Learning Mastery. 2021 [cited 20 August 2021]. Available from: https://machinelearningmastery.com/roc-curves-and precision-recallcurves-for-imbalanced-classification

[19]. F1 Score [Internet]. ritchieng.github.io. 2021 [cited 20 August 2021]. Available from: https://www.ritchieng.com/machinelearning-f1-score