Indiana Journal of Multidisciplinary Research
Abbriviate Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.12671257

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Review Article

Volume-04|Issue-03|2024

Sentimental Analysis for E-Commerce Reviews

Prathveesh Shetty*[1], Revanth T N[2], Vishnupriya B[3], Vishruth M S[4], Malini M Patil[5]

[1,2,3,4,5] Department. of Computer Science Engineering, RV Institute of Technology and Management, Bengaluru, Karnataka, India.

Article History
Received: 20.05.2024
Accepted: 05.06.2024
Published: 30.06.2024

Citation
Shetty, P., Revanth, T. N., Vishnupriya, B., Vishruth, M. S., Patil, M. M. (2024). Sentimental Analysis for E-Commerce Reviews. *Indiana Journal of Multidisciplinary Research*, *4*(3), 75-81.

Abstract: The exponential growth of e-commerce has led to an influx of customer reviews, which are a rich source of user sentiment and preferences. However, extracting meaningful insights from these reviews poses significant challenges, including data preparation, feature extraction, model optimization, addressing class imbalances, enabling real-time analysis, and determining suitable evaluation metrics. This paper proposes a novel approach to sentiment analysis for e-commerce reviews, leveraging the strengths of Support Vector Machine (SVM), Logistic Regression classifier and Random Forest classifier. While these algorithms individually offer valuable insights, their combined power through an ensemble learning technique, specifically a voting mechanism, is harnessed to enhance the accuracy of sentiment classification. This ensemble method not only improves the precision of product recommendations but also significantly enhances the overall user experience in the e-commerce platform. The proposed approach provides a robust solution to the challenges encountered in developing a precise sentiment analysis system, paving the way for more sophisticated, real-time analysis of e-commerce reviews.

Keywords: Data preparation, Support vector machines, Logistic regression classifier, Random Forest classifier, Ensemble learning, Sentiment classification, etc.

INTRODUCTION

The advent of e-commerce has revolutionized the way businesses operate, leading to an explosion of customer-generated data in the form of reviews. These reviews are a rich source of information, providing insights into customer opinions and sentiments about products and services. However, extracting meaningful information from these reviews is a challenging task due to the unstructured nature of the data. This necessitates the development of robust sentiment analysis systems capable of classifying sentiments from customer reviews.

Sentiment analysis, also known as opinion mining, involves the use of natural language processing, text analysis, and computational linguistics to identify and extract subjective information from source materials. In the context of e- commerce, sentiment analysis can help businesses understand customer opinions, improve their products and services, and refine their business strategies. Despite its potential, the development of an accurate sentiment analysis system presents several challenges. These include data preparation, feature extraction, model optimization, handling class imbalances, enabling real-time analysis, and selecting appropriate evaluation metrics. To address these challenges, this study proposes an ensemble learning model that combines Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) classifiers. The performance of this ensemble model is evaluated using precision, recall, f1- score, and accuracy metrics.

The remainder of this paper is organized as follows: Section 2 presents the methodology used in this study, including sentiment analysis, word cloud visualization, sentiment distribution, keyword extraction, preprocessing and data splitting, individual classifier training, ensemble classifier creation, and model evaluation. Section 3 discusses the results and provides valuable insights into the effectiveness of ensemble learning techniques in sentiment analysis for e-commerce. Finally, Section 4 concludes the paper and suggests future research directions.

LITERATURE SURVEY

The authors of [1] use hybridization techniques to categorize twitter data streaming based on sentiment analysis. The categorization of sentiment analysis uses the genetic algorithm, particle swarm optimization, and decision tree algorithm. The sentiment polarity categorization procedure is used by the authors of [2] to depict a sentiment analysis system for product reviews. Three phases make up the entire process. Naive Bayesian, support vector machine, and random forest are the categorization techniques chosen. According to the approach suggested by the author in [3], a reviewer's credibility is determined by how closely or professionally he is connected to the product category under evaluation. In [4] demonstrates how we rated aspects according to their value. Our system accepts free text reviews as input. Free reviews are analyzed using natural language processing (NLP), which detects the product's features. We utilized the supervised classifier SVM to categories attitudes, and then we used the probabilistic aspect ranking method to rank natural aspects. It has the

***Corresponding Author:** Prathveesh Shetty

advantage of swiftly classifying. In [5] authors used the natural language processing field has recently become interested in fine-grained sentiment analysis of text reviews. The majority of previous work has focused on developing effective feature representations of text reviews for categorization. The author in [7] proposedthe TF-IDF and FPCDA phrase FE methodology for the sentiment analysis of product reviews. By considering the various lengths of the product reviews, the local patterns of the feature vectors were discovered by employing the OPSMbi-clustering algorithm. The Prefix Span was created to recognize frequent phrases and pseudo-consecutive phrases with a high level of discrimination and word-order information. Authors in [8] used feature-specific sentiment analysis to investigate the product review. To determine the relationship between the features and the opinions they are linked with, a dependency parsing technique is applied. They created a system that gathers opinion expressions defining various prospective characteristics from reviews and extracts them.

In [9], authors created lexical integrated two-channel CNN-LSTM (Convolutional Neural Network Long Short-Term Memory) family models for sentiment analysis, which are deep learning-centered models for sentiment analysis. To create a sample of input data that had a reliable size and to develop the proportion of sentiment data in each review, the sentiment padding methodology was used. Sentiment padding solved the gradient disappearing issue that might arise when using '0' padding between the inputs layer and the first hidden layer. Premium lexicon components were developedfor sentiment analysis to be used in the operation of sentiment padding. From paper [10], the realm of e-commerce, sentiment analysis plays a pivotal role in understanding consumer opinions. This review paper provides an overview ofsentiment analysis techniques, including lexicon-based and machine learning methods. It emphasizes the application ofsentiment analysis within the retail industry, particularly e-commerce, where operators seek actionable insights from vast text databases. [11] presents a comprehensive study on the application of Natural Language Processing (NLP) and Machine Learning (ML) in analyzing customer reviews in the e-commerce sector. They underscore the challenges in manually evaluating these reviews and propose the use of NLP and ML for automation1. The paper examines the Amazon dataset using various combinations of voice components and deep learning, with a focus on identifying sentiments as ‘Positive’, ‘Neutral’, ‘Negative’, or 'Indifferent'. In [12] “Sentiment Analysis Application in E-Commerce: Current Models and Future Directions” by Huang Huang, Adeleh Asemi, and Mumtaz Begum Mustafa presents a comprehensive review of sentiment analysis (SA) in the context of e-commerce1. They investigated the methods used to address the SA problem in e-commerce, the most commonly used e-commerce platforms for data collection, and the future direction of research in this area1. The authors suggested several future directions to improve the current SA models, including addressing the limitations of existing models and exploring new approaches1.

PROBLEM STATEMENT

In the realm of e-commerce, sentiment analysis is pivotal for comprehending customer opinions and refining business strategies. However, the development of an accurate sentiment analysis system, capable of classifying sentiments from customer reviews, presents several challenges. These include data preparation, feature extraction, model optimization, handling class imbalances, enabling real-time analysis, and selecting appropriate evaluation metrics. This study addresses these challenges by implementing an ensemble learning model, combining Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF) classifiers. The performance of this ensemble model is evaluated using precision, recall, f1-score, and accuracy metrics, providing insights into the effectiveness of ensemble learning techniques in sentiment analysis for e-commerce.

MATERIALS AND METHODOLOGY

The proposed approach for sentiment analysis of e-commerce reviews involves a comprehensive methodology that begins with sentiment analysis using the TextBlob library in Python to classify reviews as 'positive', 'negative', or 'neutral'. This is followed by the generation of word clouds using the WordCloud library in Python to visually representthe most frequently occurring words in each sentiment category. The distribution of sentiments in the reviews is then visualized using a bar graph created with Matplotlib. Keyword extraction is performed using the CountVectorizer function from the sklearn.feature_extraction.text library in Python to identify the most frequently occurring words in each sentiment category. The text data is preprocessed using the Term Frequency-Inverse Document Frequency (TF-IDF) technique and split into training and testing sets. The sentiment analysis system is trained using several machine learning models, including Support Vector Machine (SVC) Classifier with random classifier, Logistic Regression classifier, Random Forest classifier, and Voting Classifier- Ensemble meta-classifier. An ensemble classifier is createdusing the VotingClassifier class from sklearn.ensemble, which combines the predictions of the three previously definedmodels. The performance of the ensemble classifier is evaluated using precision, recall, f1-score, and accuracy metrics,and a confusion matrix is plotted using seaborn’s heatmap function. This approach provides a robust solution to the challenges encountered in developing a precise sentiment analysis system, paving the way for more sophisticated, real-time analysis of e-commerce reviews.

***Corresponding Author:** Prathveesh Shetty

METHODOLOGY

Data Collection: The data collection process involves gathering customer reviews from e-commerce platforms. These reviews are a rich source of user sentiment and preferences, providing valuable insights into customer opinions and sentiments about products and services. The collected data is typically unstructured, which poses significant challenges for analysis.

Sentiment Analysis: The sentiment of the reviews was determined using the TextBlob library in Python. The sentiment polarity of each review was calculated, and based on the polarity score, the sentiment was classified as 'positive', 'negative', or 'neutral'. The outcome of this step is a new column in the dataframe, 'Sentiment', which categorizes each review as 'positive', 'negative', or 'neutral'. This categorization is crucial for further analysis andvisualization of the data.

Word Cloud Visualization: To visually represent the most frequently occurring words in the positive, negative, and neutral reviews, word clouds were generated using the WordCloud library in Python.The outcome of this stepis a set of word clouds that provide a visual representation of the most common words in each sentiment category. These word clouds can help in understanding the key themes in the reviews and can guide further analysis.

Sentiment Distribution: To understand the distribution of sentiments in the reviews, a bar graph was created using Matplotlib, a popular data visualization library in Python. The sentiment counts were normalized to percentages to provide a more intuitive understanding of the distribution.

Keyword Extraction: To identify words that appear frequently in positive, negative, and neutral analysis, keyword extraction is used by the CountVectorizer function in the sklearn.feature_extraction.text library in Python. This function converts a collection of text data into a matrix of symbol numbers. The top 10 keywords were determinedfor each category. These points provide insight into the main themes or concepts in the analysis for each set of ideas.

Preprocessing and Data Splitting: Use time-frequency inverse document frequency (TF-IDF) technology topre-process text into images representing machine learning algorithms. The data is divided into training and testing, and 80% of the data is used for training and 20% for testing.

Individual Classifiers: Sentiment analysis systems are trained using a variety of machine learning models, including Support Vector Machine (SVC) classifiers with stochastic classifiers, logistic regression classifiers, random forest classifiers, and voting Classifier with meta classifiers. The logistic regression classifier was set to a maximum of 1000 iterations to ensure convergence. The random forest classifier was initialized using 100 decision trees. These models were chosen for their ability to handle high loads and their resistance to extreme fits

Integrated classifier: The integrated classifier is built using the VotingClassifier class from sklearn.ensemble. This classification includes predictions from the three previously mentioned models. The voting system is set to "soft"; This means that the task force predicts the class list based on the argmax of the number of predicted outcomes recommended for good combination products.

Ensemble Classifier: An ensemble classifier is created using the VotingClassifier class from sklearn.ensemble. This classifier combines the predictions of the three previously defined models. The voting parameter is set to 'soft',which means that the ensemble classifier predicts the class label based on the argmax of the sums of the predicted probabilities, which is recommended for an ensemble of well-calibrated classifiers.

Model evaluation: Use fitting of training data (X_train, y_train) to train the candidate set. Use the prediction method to predict the test data (X_test). The performance of the combination is evaluated as follows: Write a classification map that includes accuracy, recall, f1 score, and support for each class. Use Seaborn's heat map function to plot the confusion matrix. This provides a visual representation of the classifier's performance. Calculatethe accuracy of the classification and print it.

WORKFLOW

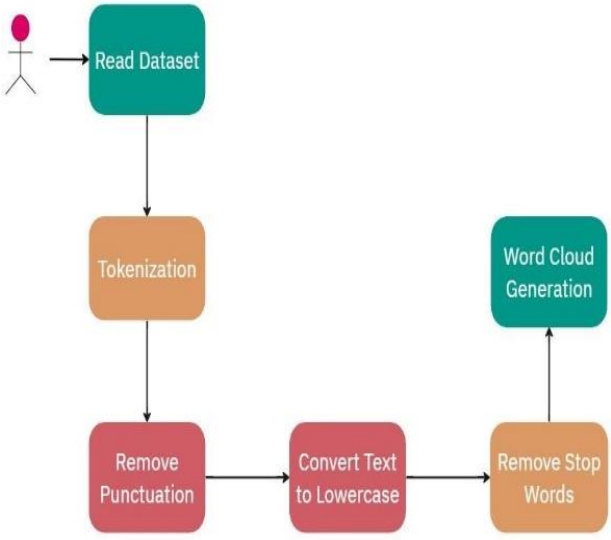

Figure 1: Data Preprocessing Steps

This Figure illustrates a flowchart for the process of generating a word cloud from a dataset. The process starts with reading the dataset, followed by tokenization, removal of punctuation, conversion of text to lowercase, and removal ofstop words. Finally, a word cloud is generated from the processed text, visually

***Corresponding Author:** Prathveesh Shetty

representing the most frequently occurring words in the dataset.

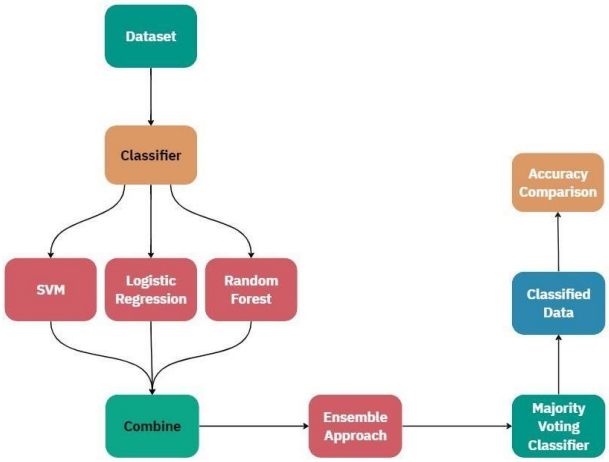

Figure 2: Sentiment Classification Phase

This Figure illustrates a flowchart for the ensemble learning process in a sentiment analysis project. It starts with a dataset that is processed by classifiers including SVM, Logistic Regression, and Random Forest. These classifiers are then combined using an ensemble approach, specifically a majority voting classifier, to improve accuracy. The results are then compared for effectiveness, providing insights into the performance of the ensemble learning model.

MATHEMATICAL MODEL

To find an image that is the same as the input image, both histograms are compared and the image corresponding to the closest histogram is returned. Different approaches are used to calculate the distance between two histograms. Here we use the Euclidean distance based on the formula (1).

$$D = \sqrt{\sum_{i=1}^{n} (hist1_i - hist2_i)^2} \quad \text{- (1)}$$

SVM Classifier: Support Vector Machine (SVM) is a powerful learning machine used for arbitrary or random classification, propagation and even anomaly detection.

The SVM classifier maps the input v_i features, maps them to a higher space, and finds the hyperplane that best divides the points into category C. The dimension function of SVM is given by formula (2).

$$f_{SVM}(v_i) = sign(\sum^{n} \quad j-1$$

$$a_j y_j K\ (v_i, v_j) + b) \text{ - (2)}$$

Logistic Regression Classifier: Logistic Regression Classifier: Logistic regression is a statistical problem algorithm used for binary classification. Predicts the outcome of categorical dependent variable 1. The result must be a categoricalor discrete value. 0 or 1, True or False, etc. it could be.

Logical function, also known as sigmoid function. The sigmoid function is an S-shaped curve that can take a real valueand represent it between 0 and 1, but cannot reach these limits exactly 2. The logistic function is shown by the formulabelow (3)

$$\sigma(z) = \frac{1}{1+e^{-z}} \text{ - (3)}$$

Random Forest Classifier: Random Forest is a classifier that works by creating multiple decision trees during trainingand outputs the clusters. This is the formula used for the distribution or mean estimated by a regression tree. The basic principle behind the design model is the process of combining weak learners to create strong learners

The random forest classifier outputs a class during training by creating a multidimensional trees T = {t1, t2, t3,…., tm}, which is the class (classification) of each tree model. The decision function can be expressed by formula (4).

$$f_{RF}(v_i) = \frac{1}{m}\sum_{k-1}^{m} f_{t_k}(v_i) \text{- (4)}$$

Ensemble Model: The posterior distribution $f ensemble(v_i)$ is determined by combining SVM, logistic regression and random forest model, and simple voting, weighted average or other aggregation methods can be used to determine. lastset of views. Therefore, the joint decision can be modeled as shown in Equation (5).

$$f_{ensemble}(v_i) = Aggregate(f_{SVM}(v_i), \sigma(z), f_{RF}(v_i)) \text{ -(5)}$$

RESULTS

Table1. Summary of Classification of Results

Datasets	Accuracy	Positive Precision	Neutral Precision	Negative Precision
Amazon Reviews	85 %	0.85	1.00	0.75
Flipkart Reviews	91 %	0.88	0.91	0.91
EBAY Reviews	80 %	0.83	0.70	0.75

The table in the Figure compares the accuracy and precision of sentiment analysis on reviews from Amazon, Flipkart, and EBAY. Flipkart Reviews have the highest accuracy at 91% and the highest negative precision at 0.91. Amazon Reviews achieve perfect neutral precision at 1.00. EBAY Reviews have the lowest accuracy at 80% and the lowest neutral precision at 0.70.

***Corresponding Author:** Prathveesh Shetty

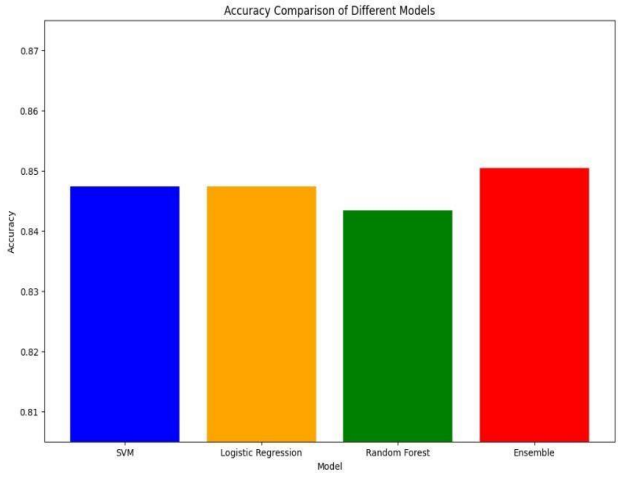

Figure 3: Accuracy comparison for SVM, LR. RF and Ensemble Models for Amazon Reviews

The Figure is a bar graph comparing the accuracy of four different machine learning models: SVM, Logistic Regression, Random Forest, and Ensemble. The Ensemble model achieves the highest accuracy, close to 0.852, while the SVM and Logistic Regression models have similar accuracies just above 0.848. The Random Forest model has the lowest accuracy, slightly above 0.842.

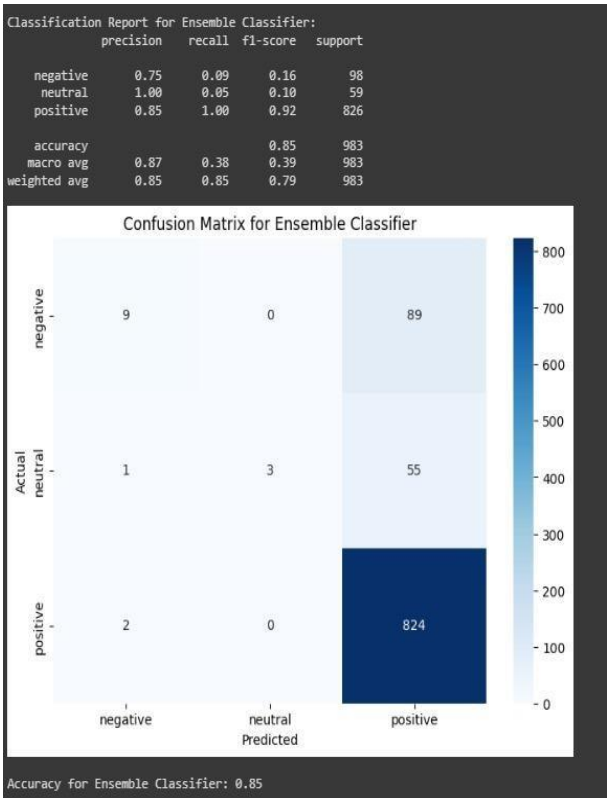

Figure 4: Model Evaluation for Amazon Reviews

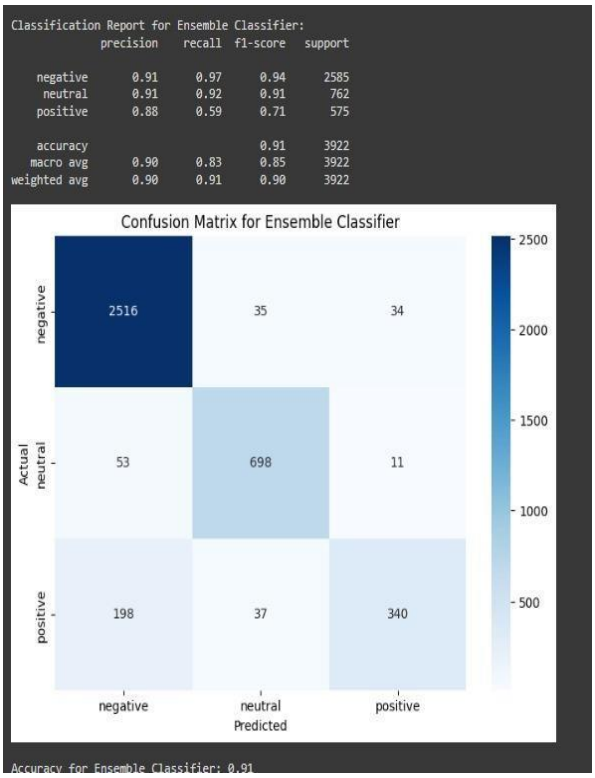

Figure 5: Model Evaluation for Flipkart Reviews

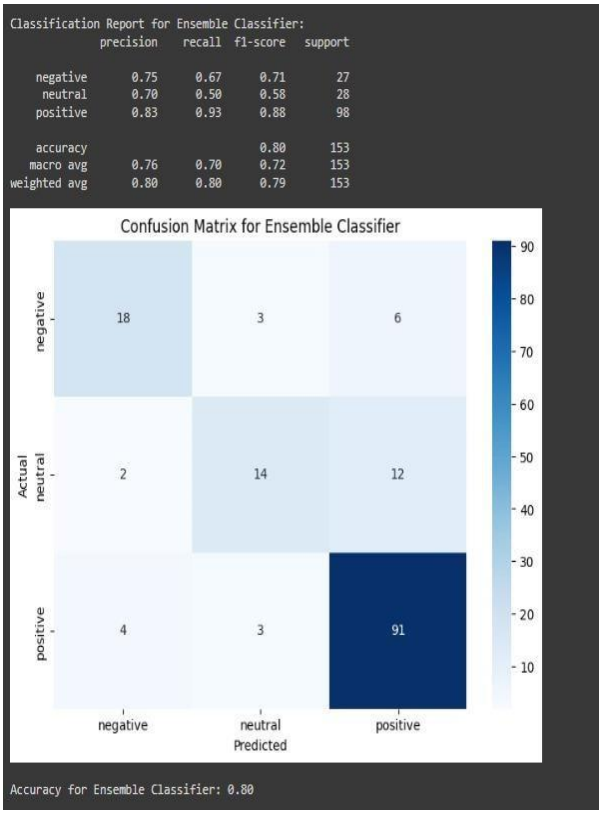

Figure 6: Model Evaluation for EBAY Reviews

***Corresponding Author:** Prathveesh Shetty

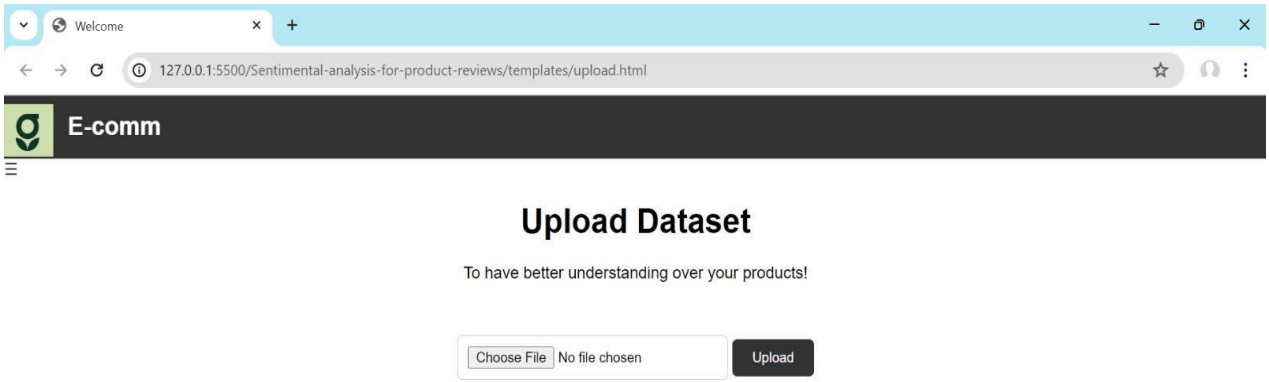

Figure 7: Interface for uploading dataset

CONCLUSION

The implementation of ensemble learning utilizing Support Vector Machines (SVM), Logistic Regression, and Random Forest algorithms for sentiment analysis on an e-commerce website has proven highly effective. Each algorithm brings distinct strengths that enhance the model's overall performance. SVM excels in handling high-dimensional data, crucial for analyzing the nuanced sentiments embedded in the textual data of customer reviews. Its ability to effectively separate classesin non-linear feature spaces allows it to capture subtle expressions of sentiment that might be missed by less sophisticated models.

Logistic Regression adds value through its probabilistic approach to binary classification, making it ideal for determining the likelihood that a review expresses a positive or negative sentiment. Its simplicity and efficiency in processing large datasets make it a practical choice for the voluminous data typically found on e-commerce platforms. The model's interpretability is especially beneficial, providing clear insights into which features most influence the sentiment classification.

Random Forest complements these approaches by leveraging multiple decision trees to improve prediction accuracy androbustness, effectively handling complex feature interactions and reducing the risk of overfitting. This is particularly useful in e-commerce, where user-generated content can vary widely, necessitating a flexible and adaptive model capable of generalizing well to new data.

Together, these algorithms form a robust ensemble that maximizes accuracy and reliability in sentiment analysis. By combining their individual advantages, they address the complexities of e-commerce data, resulting in a more nuanced and comprehensive understanding of customer sentiments. This ensemble approach not only enhances prediction accuracy but also offers scalable and adaptable solutions suitable for the dynamic nature of online retail environments.

REFERENCES

1. Nagarajan, S. M., & Gandhi, U. D. (2019). Classifying streaming of Twitter data based on sentiment analysis using hybridization. *Neural Computing and Applications*, *31*, 1425-1433.
2. Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big data*, *2*, 1-14.
3. Sindhu, C. and Mukherjee, D. (2021, Plaub Hlis Ntuj). A unified content model for electronic product analysis. *2021 5th International Conference on Computational Methods and Communications (ICCMC)* (pp. 574-578). IEEE.
4. Patil, H. and Mane, P.M. (2016). The product reviews survey needs to be analyzed by rank. *International Journal ofScience and Research*, 5(9), 749–752.
5. Wadbude, R., Gupta, V., Mekala, D., Jindal, J. and Karnick, H. (2016). User bias in identifying positive emotions is eliminated. *Association for Computational Linguistics, European Section*, arXiv:1612.06821v1 [cs.CL] Retrieved 20 December 2016.
6. Mehta, P., & Pandya, S. (2020). A review on sentiment analysis methodologies, practices and applications. *International Journal of Scientific and Technology Research*, *9*(2), 601-609.
7. Chen, X., Xue, Y., Zhao, H., Lu, X., Hu, X., & Ma, Z. (2019). A novel feature extraction methodology for sentiment analysis of product reviews. *Neural Computing and Applications*, *31*, 6625-6642.
8. Mukherjee, S., & Bhattacharyya, P. (March 2012). Be specific with product reviews. *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 475-487). Springer, Berlin, Heidelberg.
9. Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE access*, *8*, 23522-23530.
10. Marong, M., Batcha, N. K., & Mafas, R. (2020). Sentiment Analysis in E-Commerce: A Review on The Techniques and Algorithms. *Journal of Applied Technology and Innovation (e-ISSN: 2600-7304)*, *4*(1), 6.
11. Singh, U., Saraswat, A., Azad, H. K., Abhishek, K., & Shitharth, S. (2022). Towards improving e-commerce customer review analysis for sentiment

***Corresponding Author:** Prathveesh Shetty

detection. *Scientific Reports*, *12*(1), 21983.

12. Huang, H., Asemi, A., & Mustafa, M. B. (2023, July). Sentiment Analysis Application in E-Commerce: Current Models and Future Directions. In *International Conference on Electronic Government and the Information Systems Perspective* (pp. 67-72). Cham: Springer Nature Switzerland.

***Corresponding Author:** Prathveesh Shetty