IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

PRODUCT RECOMMENDATION BASED ON SENTIMENT ANALYSIS OF GENUINE REVIEWS

[1]Dr. Shivananda V Seeri, [2]Ajnya V Naik, [3]Bhagyashree Ramachandra, [4]Mridul Rajeev, [5]Prajna G

[1]Professor ,Mangalore Institute of Technology & Engineering, Moodbidre, India
[2]Student,Mangalore Institute of Technology & Engineering, Moodbidre, India
[3]Student,Mangalore Institute of Technology & Engineering, Moodbidre, India
[4]Student, Mangalore Institute of Technology & Engineering, Moodbidre, India
[5]Student,Mangalore Institute of Technology & Engineering, Moodbidre, India

Abstract : In the era of online shopping, product reviews have gained immense significance as a determining factor in the consumer's decision-making process. Customers heavily rely on reviews to assess the quality and dependability of products prior to making a purchase. This study proposes a recommendation system for Amazon electronics products that incorporates ML algorithms, including XGBoost, logistic regression, CatBoost. The dataset used here in this project consists of Amazon electronics reviews obtained from Jmcauley's website, which includes reviews, ratings, product title, product id and reviewer details. The system applies these algorithms to classify reviews as negative or positive and recommends products depending on sentiment score of positive reviews. The algorithms underwent training using a preprocessed dataset, and their performance was assessed by evaluating their accuracy using a separate testing dataset. Findings revealed that XGBoost yielded the highest accuracy, achieving 86.2%, followed by CatBoost which yielded the accuracy of 85.3% and logistic regression gave an accuracy of 84.8%.

Index Terms - **Sentiment Analysis, Product Recommendation, Preprocessing, Machine Learning, Logistic Regression.**

INTRODUCTION

Sentiment analysis involves identifying and extracting opinions, emotions, and attitudes from textual data. Over last few years, analysing the sentiments of product reviews has evolved as a crucial aspect of consumer decision-making process. In context of product reviews, sentiment analysis involves analyzing user-generated content to determine whether a review is negative or positive. This analysis gives valuable information into customer satisfaction, which can be utilized to enhance products and services. Product recommendation systems are specifically designed to aid customers to find products that go along with their requirements and preferences. Machine learning algorithms, like Logistic regression, XGBoost, and CatBoost, is used to divide the reviews as positive, neutral or negative and recommend products accordingly. The primary objective of this system is to improve the customer experience by offering product recommendations that are based on sentiments expressed in customer reviews.

NEED OF THE STUDY

The need for the current system stems from the growing importance of product reviews in the consumer decision-making process. With the widespread availability of internet access and the rise of e-commerce platforms, product reviews are a crucial resource of information for customers looking to make informed purchasing decisions. Online reviews can significantly impact customer purchase behavior and product sales.Yet, the overwhelming quantity of reviews poses a challenge for customers in terms of effectively processing and comprehending this vast amount of information. In addition, traditional product recommendation systems that rely on popularity-based or collaborative filtering approaches might not accurately reflect customer preferences and sentiment. Hence, there is need for more sophisticated recommendation systems that embodies sentiment analysis of the product reviews to provide more personalized recommendations.This system aims to address this need by proposing a system for product recommendation that leverages sentiment analysis of Amazon electronics reviews using ML algorithms. The proposed system can increase accuracy of product recommendations and help customers to make informed purchasing decisions. The findings of this study can be valuable for e-commerce companies and retailers looking to improve their product recommendation systems and enhance customer satisfaction.

LITERATURE REVIEW

[1] Recommending Insurance products by using Users' Sentiments, Rohan Parasrampuria1 , Ayan Ghosh2 , Suchandra Dutta3 and Dhrubasish Sarkar4(*): The paper "Recommending Insurance products by using Users' Sentiments" proposes a new method for recommending insurance products to customers based on their sentiments. To implement their approach, the authors collect customer reviews and ratings of insurance products from online sources, and use NLP(natural language processing) techniques to

analyze the expressed sentiments in the reviews. They then train the ML models to predict the sentiment that corresponds with specific insurance products, and use these predictions to make personalized recommendations to customers.

[2] Sentiment Analysis for Product Recommendation Using Random Forest, Gayatri Khanvilkar1 *, Prof. Deepali Vora2 :The paper presents an approach of ML for sentiment analysis in product recommendations. The authors used Random Forest algorithm to classify reviews as positive or negative, and then used the sentiment scores to recommend products to users. They collected data from an online shopping website and evaluated performance of their model using accuracy, precision, F1 score and recall. The result showed that this approach had high accuracy and could effectively recommend products based upon sentiment analysis.

[3] Product Recommendation System from Users Reviews using Sentiment Analysis,:The paper presents a recommendation system of products that employs sentiment analysis of user reviews. The system processes the reviews using NLP techniques and extracts features related to product characteristics and user sentiments. Then, a recommendation model based on user preferences is developed using collaborative filtering and content-based filtering techniques. The system is evaluated using a dataset of Amazon product reviews and achieves good performance in terms of accuracy & precision. The authors suggest that this proposed system is useful for e-commerce websites to improve their product recommendations and enhance user satisfaction.

[4] Sentiment Analysis of Product Reviews, Devendra Kamalapurkar : The paper proposes a system for sentiment analysis of reviews using ML algorithms. The system preprocesses the reviews by removing stop words, punctuation marks, and converting text to lowercase. It then uses feature extraction techniques to extract important words and applies ML algorithms such as Naive Bayes, Decision Trees, SVM, to classify reviews as positive or neutral or negative, . The authors evaluate the proposed system on dataset of product reviews and achieve good accuracy in classification of sentiment. They conclude by mentioning that the approach can be useful for businesses to extract insights from customer reviews and improve their products and services.

[5] The paper proposes feature-based sentiment analysis approach to analyze product reviews. Proposed method uses a combination of ML algorithms and lexicon-based methods to extract features followed by classifying them as positive, neutral or negative sentiments. The authors opine that given approach outperforms existing methods in terms of accuracy, precision and recall. The paper also provides a detailed explanation of the feature extraction and sentiment classification methods used in their approach.

[6] The paper proposes a novel approach to sentiment analysis that uses sequence model to learn user & product distributed representations. The authors opine that traditional approaches to sentiment analysis fail to capture the nuanced and complex relationships between users, products, and sentiment.The proposed method first encodes user and product information as sequences, and then uses sequence model to learn distributed representations for users & products. The authors evaluate their approach on two datasets, and show that it outperforms several state-of-the-art sentiment analysis methods in case of accuracy and F1 score. The paper presents an approach that incorporates user and product information in a meaningful way, and could have applications in fields such as social media and e-commerce analysis.

[7] The paper speaks about the use of sentiment analysis, a sub-field of text mining, to classify customer reviews into negative and positive categories. It involves the usage of ML classification models, like Naïve Bayes, SVM, and Decision Tree, inorder to extract subjective information from a text. The authors extracted over 4,000,000 unstructured data containing mobile phone reviews from Amazon.com and filtered out noisy data. The reviews were pre-processed so as to evaluate sentiment using supervised learning, and the best one among classification models was determined by 10 Fold Cross Validation. The paper highlights the importance of analysis of sentiment in enhancing global connections among consumers and influence their buying patterns.

[8] This paper explores the role of ML algorithms for sentiment analysis. The authors discuss the challenges of analyzing the massive amount of user-generated data available on social media platforms and micro-blogging websites, where users express and share their opinions on various subjects. The paper categorizes ML algorithms into three approaches: lexicon-based, machine learning-based, and hybrid techniques. Machine learning-based techniques are found to be efficient and reliable for opinion mining and sentiment classification . The paper also provides a comprehensive analysis of different ML techniques and their respective accuracy. The authors review related works in the field and discuss various algorithms used for sentiment categorization, such as Maximum Entropy (MaxEnt), Stochastic Gradient Descent (SGD), and Random Forest. The paper concludes by highlighting importance of these techniques in sentiment analysis & need of further research in this area.

METHODOLOGY

The proposed methodology involves the utilization of ML algorithms, specifically Logistic Regression, Catboost, and XGboost to classify reviews. The dataset that is used to train these models is obtained from Amazon Electronics data available on the Jmcauley website, which consists of two sets - review data and metadata. The data is pre-processed to enhance accessibility and analysis ease and will be used to extract features and implement machine learning algorithms.

Data Collection: Collect a large dataset of product reviews from e-commerce website.

Data Preprocessing: Preprocess the collected data to remove unnecessary information like punctuation marks, and stop words. Also, convert the text to lowercase and perform lemmatization to standardize the text.

Sentiment Detection: Using Sentiment Intensity Analyzer to detect sentiment of each review as positive, negative, or neutral.

Feature Extraction: Extract the relevant features from review text, such as keywords, phrases, and emotions which is used to classify the reviews. This is done using techniques like TF-IDF and Count vectorizer.

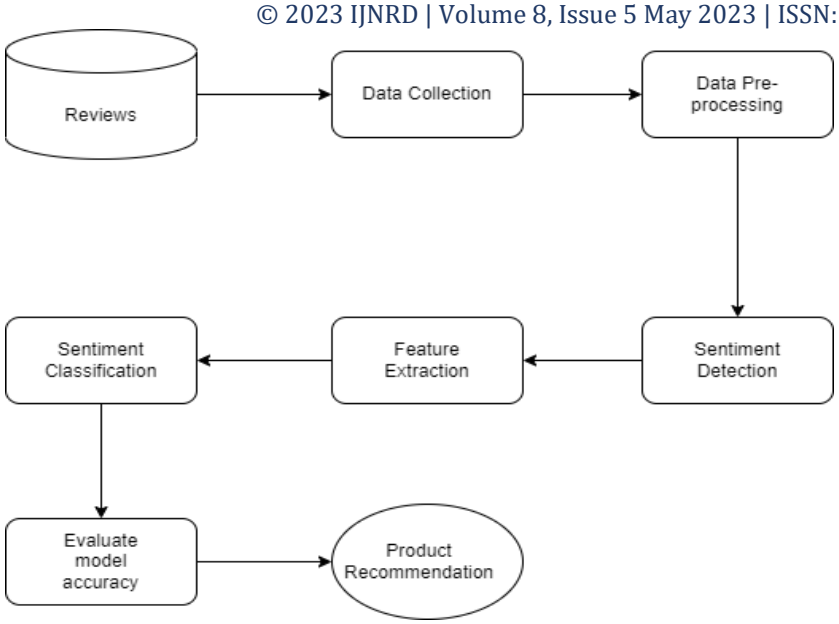


Fig 1.Proposed System Methodology

Sentiment Classification: Use a classification algorithm like Catboost, Logistic Regression and XGboost to classify the reviews based on extracted features and sentiment scores.

Logistic Regression: It is statistical algorithm that is used for binary classification problems. It is a type of supervised learning, meaning it requires labeled data to train model. The aim of using logistic regression is to predict the probability of an input belonging to a particular class, given a set of input features. The model is based on logistic function (also called as the sigmoid function) and can be represented by the following formula:

P(y=1|x) = 1 / (1 + exp(-z)), where:P(y=1|x) is the probability of the input belonging to class 1 (e.g. positive sentiment) given the input features x, exp is the exponential function, z is the linear combination of input features and their corresponding coefficients: z = b0 + b1x1 + b2x2 + ... + bn*xn, b0, b1, b2, ..., bn are the coefficients of logistic regression model, that are learned during training, x1, x2, ..., xn are the input features.In logistic regression, model is trained by minimizing a cost function, such as cross-entropy loss, which measures difference between predicted probabilities and the true labels. Once the model is trained, it is be used to make predictions on new input data by computing the probability of each input belonging to class 1 and using a decision threshold to classify the input as either class 1 or class 0 (e.g. negative sentiment).

CatBoost: The CatBoost algorithm is gradient boosting algorithm that is based on principle of minimizing a loss function using an ensemble of weak models. The algorithm builds a set of decision trees to form an ensemble model, where each tree is trained on a randomly selected subset of data and a randomly selected subset of the features.. The formula for the CatBoost algorithm has following steps:Initialization, Training the decision trees, Regularization, Combining the decision trees.The overall formula for the CatBoost algorithm can be written as:

y = mean(y_train) + sum(f_k(x_i)) for k in 1 to K, where: y is predicted value for a sample x_i,mean(y_train) is the mean value of target variable in the training data,f_k(x_i) is predicted value of k-th decision tree for the sample x_i,K is number of decision trees in the ensemble. Overall, the CatBoost algorithm is powerful and flexible algorithm for building accurate and robust predictive models, especially while handling datasets that have a huge number of categorical features.

XGboost: XGBoost (Extreme Gradient Boosting) is a popular gradient boosting algorithm that is widely used in machine learning for classification, regression, and ranking tasks. It is based on the principle of gradient boosting, where an ensemble of weak models is combined to form a strong model that would make accurate predictions. The general formula for the XGBoost algorithm can be written as follows:

y_hat = f(x) = w0 + sum(w_i * h_i(x)) ,where: y_hat is the predicted value for a given input x,w0 is the bias term, which represents the mean value of target variable, w_i are the weights assigned to every weak model (i.e., each decision tree) ,h_i(x) is the prediction of the i-th weak model (i.e., the i-th decision tree) for the input x.

The goal of XGBoost algorithm is to determine optimal values for weights w_i and decision trees h_i(x) that minimize a given loss function. The loss function typically measures difference between predicted values and actual values of the target variable for a given set of input features.The XGBoost algorithm achieves this by using gradient descent to iteratively update the weights and decision trees. At each iteration, a new decision tree is included in the ensemble to improve accuracy of model. The weights and the decision trees are updated in a way that minimizes the loss function while also incorporating regularization techniques to prevent overfitting.

Evaluate Model Accuracy: Evaluate accuracy of model using various metrics like F1-score, precision and recall. Use techniques like cross-validation to ensure accuracy.

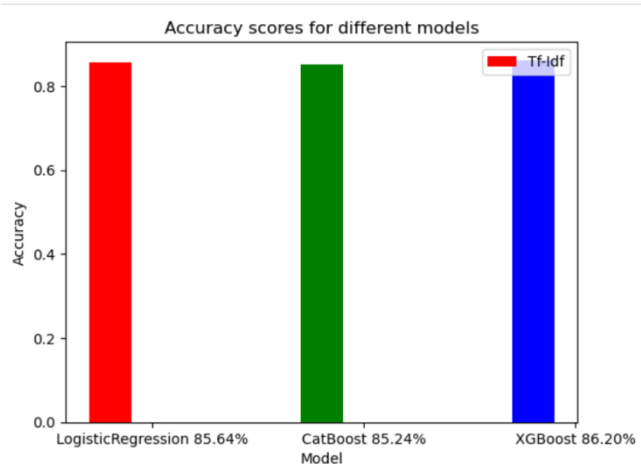
Product Recommendation: Once the reviews have been classified, recommend the products with highest number of positive reviews to customers. This is done by sorting the products based on their average sentiment scores.

RESULTS AND DISCUSSION

The dataset is split into train data and test data with 80% train data and 20% test data .The proposed system was tested and was able to classify accurately , the sentiment of customer reviews with an average accuracy of 86.37%. Moreover, the recommendation engine was capable to suggest relevant products to customers based on sentiment analysis. In our evaluation, the system achieved a good accuracy in recommending products that were positively reviewed.

Table 1. Evaluation parameters for classifiers

Classifiers	Logistic Regression		Catboost		XGboost	
Vectorizers	Count Vectorizer	Tf-Idf	Count Vectorizer	Tf-Idf	Count Vectorizer	Tf-Idf
Accuracy	84.86	85.63	85.33	85.23	86.03	86.42
Precision	82.46	83.09	81.26	80.75	83.07	84.08
Recall	84.86	85.63	85.33	85.23	86.03	86.42
F1 Score	83.40	82.51	81.56	81.80	83.65	84.12



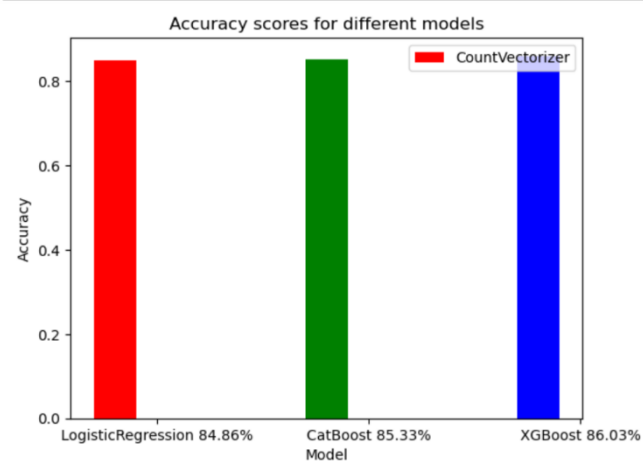


Fig 4. Graph of Accuracy of Classifiers with Tf-idf Vectorizer Fig 5. Graph of Accuracy of Classifiers with CountVectorizer

The metrics used are: Accuracy: It measures the fraction of true predictions among all the predictions made by model. accuracy_score(y_true, y_pred): (TP + TN) / (TP+ FP + TN + FN)

Precision:It calculates ratio of true positive predictions to total number of positive predictions made by model. precision_score = TP / (TP + FP)

Recall: It measures ratio of true positive predictions to all actual positive samples in the test set.recall_score = TP / (TP + FN)

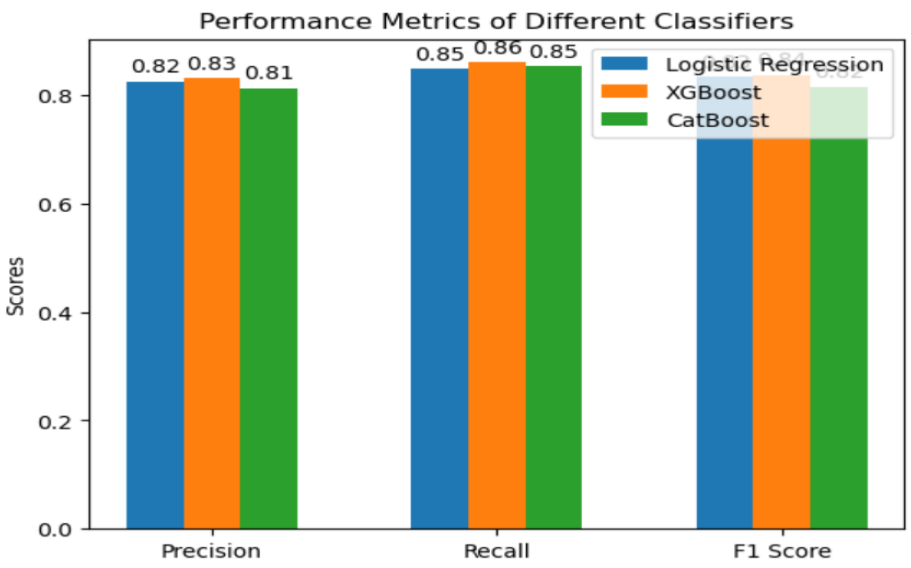


Fig 6. Graph of Precision, Recall, F1 Score of Classifiers

F1 Score: It is the harmonic mean of precision and recall that provides a balance between the two metrics.f1_score(y_true, y_pred, average='weighted'): 2 * (precision * recall) / (precision + recall), where TP (True Positive) is number of correctly predicted positive samples, TN (True Negative) is number of correctly predicted negative samples, FP (False Positive) is number of incorrectly predicted positive samples, and FN (False Negative) is the number of incorrectly predicted negative samples. The average='weighted' parameter

in the precision, recall, and f1_score functions computes the weighted average of the metrics, where each class's contribution is weighted by its number of samples.

CONCLUSION

In conclusion, the product recommendation system leverages sentiment analysis of Amazon electronics reviews, which can provide valuable information into the customers' preferences and opinions. In this regard, the proposed system utilizes machine learning algorithms, including Catboost, Logistic Regression, and XGboost, to classify Amazon electronics reviews as positive or negative and recommend products accordingly. The system is designed and implemented in a way that transforms textual features into numerical features by utilizing vectorization techniques such as tf-idf and count vectorizer. By doing so, the proposed system has the potential to improve accuracy of product recommendations, empowering customers to make more informed purchasing decisions. The insights obtained from this, could be particularly valuable for e-commerce companies and retailers seeking to optimize their product recommendation systems and enhance customer satisfaction.

REFERENCES

[1] Recommending Insurance products by using Users' Sentiments Rohan Parasrampuria1 , Ayan Ghosh2 , Suchandra Dutta3 and Dhrubasish Sarkar4

[2] Sentiment Analysis for Product Recommendation Using Random Forest , Gayatri Khanvilkar1 *, Prof. Deepali Vora2

[3] SENTIMENT ANALYSIS OF PRODUCT REVIEWS, Devendra Kamalapurkar *1, Ninad Bagwe2 , R. Harikrishnan 3 , Salil Shahane 4 , Mrs. Manisha Gahirwal 5

[4] Feature based Sentiment Analysis for Product Reviews, Dr. D. Sivaganesan ,Dhruv Aggarwal ,Sridhar K ,Arunkumar M.

[5] E-Commerce Site's Fake Review Detection and Sentiment Analysis using ML Technique , J Bharatkumar1 , Kartik M2 , Kiran Shetty3 , K Shreyas Pai4, Sunil Kumar S5

[6] Sentiment analysis using product review data ,Xing Fang* and Justin Zhan

[7] SmartTips: Online Products Recommendations System Based on Analyzing Customers Reviews ,Noaman M. Ali 1,* , Abdullah Alshahrani 2 , Ahmed M. Alghamdi 3 and Boris Novikov 4

[8] A Topic based Approach for Sentiment Analysis on Twitter Data , Pierre FICAMOS*, Yan LIU

[9] Sentiment Analysis on Product Reviews Using Machine Learning Techniques ,Rajkumar S Jagdale ,Vishal S. Shirsath ,Sachin Deshmukh

[10] Sentiment Analysis of Customer Reviews Using Deep Learning Techniques, by Kavita Ganesan and ChengXiang Zhai

[11] Comprehensive Review of Opinion Summarization, HYUN DUK KIM , KAVITA GANESAN, PARIKSHIT SONDHI, and CHENGXIANG ZHAI University of Illinois at Urbana-Champaign