Sentiment Classification on E-Commerce Women Review Data to Predict Recommendation or Non-Recommendation of Products (2023)

Humna Naeem Zaidi,
Faculty of Engineering, Environment and Computing,
Coventry University
Coventry, England
zaidih3@uni.coventry.ac.uk

[1] ***Abstract*—Businesses often struggle with understanding the voice of the customer, but this challenge can be tackled by leveraging sentiment analysis to extract valuable information from customer reviews. Customer feedback offers a valuable platform for uncovering a wide range of reactions initiated by customers regarding the products they have purchased. By employing text analytics, businesses can obtain a more comprehensive understanding of customer satisfaction or dissatisfaction. Relying solely on customers' ratings is insufficient in capturing their experiences, as reviews can provide crucial feedback that wouldn't be captured through a simple Likert scale. Even a five-star review may contain important suggestions for improving delivery time or customer support. By delving deeper into customer reviews, businesses can gain valuable insights to respond strategically and effectively and consider making necessary adjustments to their existing systems and approaches to better serve their customers.**

***Index Terms*—machine learning; python; class imbalance; feature extraction; online shoppers' intention, adaptive boosting,, logistic regression, naive bayes , stochastic gradient.**

I. STUDY IS RELEVANT?

The use of machine learning algorithms has increased dramatically in various fields. For the purposes of this work, we use a classification algorithm to classify a large number of reviews into helpful and unhelpful reviews.

II. OTHERS

(Markus et al.; 2019) designed a probit A model based on the Nagelkerke pseudo-R-squared scale for describing total stars evaluation. The results showed that the Probit model performed better on star rating. Again and again, The model was easy to interpret and helped us analyze customer reviews. This model effectively addressed existing methods of describing overall star ratings as generic. It does not address methodological issues related to these star ratings and ignores comments Texts containing valuable information about customer ratings on various topics The aspect of the item that was rated.

III. INTRODUCTION

Customer reviews hold immense influence in the realm of e-commerce, significantly shaping consumer purchasing decisions and brand success. To gain valuable insights from these reviews and comprehend the underlying sentiment, businesses are continuously searching for effective methodologies. Within this context, sentiment classification becomes particularly relevant, especially in the domain of women's products. Sentiment classification involves automatically analyzing and categorizing textual data, such as customer reviews, to determine the sentiment expressed toward a specific product or service. In the e-commerce landscape, sentiment classification proves invaluable for predicting whether customers will recommend or not recommend a product to others. The primary objective of this study is to apply sentiment classification techniques to a dataset comprising women's product reviews in the e-commerce sector. By leveraging machine learning and natural language processing (NLP) algorithms, The author's aim to develop a model capable of accurately predicting product recommendations or non-recommendations based on the sentiments conveyed in the reviews. This research holds substantial significance for e-commerce businesses as it provides a means to extract actionable insights from customer feedback. By accurately discerning sentiment within reviews, businesses can gain a deeper understanding of customer preferences, satisfaction levels, and areas for improvement. This knowledge, in turn, can be utilized to enhance product development, refine marketing strategies, and improve customer support, ultimately leading to enhanced customer experiences and heightened customer loyalty. Throughout this study, The author will focus on exploring supervised machine learning algorithms specifically tailored for sentiment classification. Furthermore, The author will evaluate their performance using appropriate metrics, conducting extensive experiments to identify the most effective approach in predicting product recommendations or non-recommendations. In essence, this research aims to bridge the gap between customer reviews and actionable business insights through the implementation of sentiment classification. By accurately predicting sentiments expressed

[1]

in women's product reviews, businesses can make informed decisions that drive customer satisfaction, bolster product success, and ultimately foster business growth.

IV. PROBLEM AND DATASET

The objective of this study is to address a binary classification problem in which I aim to predict whether customers will recommend a product based on the text of their reviews. The target variable in this problem is binary, taking on values of 0 or 1, indicating whether the customer appreciates the product enough to recommend it or not. By analyzing the textual content of the reviews, The author's goal is to accurately classify whether a customer's sentiment leans towards recommendation or non-recommendation.

According to the previous study the authors usually discard the missing value in sentimental classification but in this experiment the author kept the record constant by using The "SimpleImputer" with constant strategy. See the **Fig.1** for flow of steps to accomplish the objective

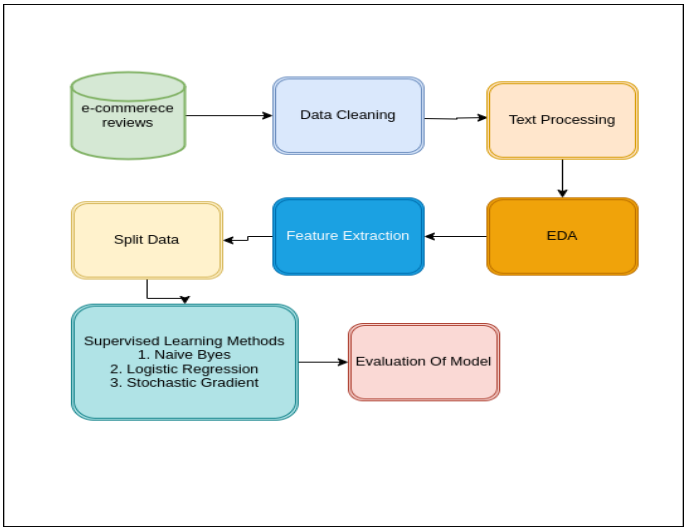


Fig. 1. Sentimental Classification Problem solution steps.

A. DATASET

- **Clothing ID:** A unique identifier for the specific product being reviewed.
- **Age:** The age of the reviewer.
- **Title:** The title of the review given by the customer.
- **Review Text:** The main body of the review written by the customer.
- **Rating:** The score given by the customer to rate the product on a scale of 1 (worst) to 5 (best).
- **Recommended IND:** A binary variable indicating whether the customer recommends the product. 1 typically represents a recommendation, while 0 indicates that the customer does not recommend it.
- **Positive Feedback Count:** The number of other customers who found this review positive or helpful.
- **Division Name:** The categorical high-level division or category of the clothing product.
- **Department Name:** The categorical department name of the clothing product.
- **Class Name:** The categorical class name of the clothing product.

These variables provide information about the product, the customer's opinion, and some demographic information about the reviewer. They can be used for various analyses, such as sentiment analysis, recommendation prediction, and understanding customer preferences within different categories and departments.

B. PRIVACY AND DATA PROTECTION

Sentiment classification often involves analyzing user-generated content, which may contain personal or sensitive information, But the dataset the author is using is totally anonymous.

V. CLASSIFICATION TECHNIQUES

A. NAIVE BAYES

The Naive Bayes classifiers are a popular family of probabilistic machine learning models utilized for classification purposes. These models are built upon Bayes' theorem and make a "naive" feature independence assumption. Despite this simplification, Naive Bayes classifiers have demonstrated strong performance in diverse applications, particularly in tasks involving text classification.

These classifiers estimate the probability of an instance belonging to a specific class by combining the prior probability of the class with the conditional probabilities of the features given that class. The estimation of these probabilities is derived from the training data. The "naive" assumption implies that the features are treated as conditionally independent, implying that the presence or absence of one feature does not impact the presence or absence of another.

Naive Bayes classifiers offer computational efficiency and require relatively small amounts of training data. They are well-suited for tasks involving high-dimensional feature spaces and are particularly effective for categorical or discrete features. These classifiers find extensive application in various domains, including document classification, spam filtering, sentiment analysis, and recommendation systems. In summary, Naive Bayes classifiers present a straightforward yet powerful approach for classification tasks. They leverage probabilistic reasoning and the assumption of feature independence to make predictions, offering a practical and effective solution.

B. LOGISTIC REGRESSION

Logistic regression is a widely used classification algorithm in text classification. Specifically designed for binary classification problems, it predicts whether a given text belongs to a particular class or not. In text classification, logistic regression estimates the probability of a text belonging to a specific class using a logistic function.

It considers the relationship between the input features, such as word frequencies or bag-of-words representations, and the probability of belonging to a certain class. By fitting a linear

model to the input features and applying a logistic (sigmoid) function, logistic regression ensures that the predicted probabilities fall within the range of 0 to 1, enabling easy interpretation.

During training, logistic regression optimizes the model's parameters using techniques like maximum likelihood estimation. It adjusts the weights assigned to different features to minimize the difference between predicted probabilities and actual class labels in the training data. Once trained, the logistic regression model can make predictions on new, unseen texts by computing the probability of belonging to a certain class. Applying a threshold determines the final classification decision.

Logistic regression is valued for its simplicity, efficiency, and interpretability. It can handle high-dimensional feature spaces commonly encountered in text classification. It can also be extended to handle multi-class classification problems using one-vs-rest or multinomial logistic regression.

C. STOCHASTIC GRADIENT

Stochastic Gradient Descent (SGD) is an optimization algorithm commonly used in text classification tasks. It is particularly effective when dealing with large datasets and high-dimensional feature spaces. In text classification, SGD aims to find the optimal set of parameters that minimizes the classification error or maximizes the classification accuracy. It does so by iteratively updating the parameters based on the gradients computed from a randomly selected subset of the training data, known as a mini-batch.

This random sampling makes SGD computationally efficient and suitable for large-scale text classification problems. During each iteration, SGD computes the gradients of the loss function concerning the parameters using the mini-batch. It then adjusts the parameter values in the direction of the negative gradients, scaled by a learning rate. This iterative process continues until convergence or a predefined number of iterations is reached.

The stochastic nature of SGD introduces randomness into the parameter updates, allowing it to escape local optima and potentially find better solutions. However, this randomness can also introduce some instability in the convergence behavior compared to other optimization algorithms. SGD is especially well-suited for text classification because it can handle large feature spaces and high-dimensional data efficiently.

Text data often have numerous features, such as word frequencies or TF-IDF values, making SGD a practical choice for training classifiers on such datasets. Overall, SGD is a powerful optimization algorithm for text classification, providing a computationally efficient approach to finding optimal parameter values. Its ability to handle large-scale datasets and high-dimensional feature spaces makes it a valuable tool in the field of text classification.

VI. EXPERIMENTAL SETUP

In previous work

A. DATA PROCESSING

This research is centered around analyzing the sentiments expressed in women's reviews and examining how they align with the recommendations given.The author has chosen specific features for the author analysis, namely the "Review Text" and the "Recommendation IND."

As “Unnamed” is reflect no information so The author will exclude them for further analysis.

Fig. 2. WordCloud before data processing.

Data Processing is a crucial step before applying any classification algorithm as it can significantly impact the accuracy of the results. The removal of noise plays a vital role in this process. Various preprocessing techniques are employed to achieve this objective. See the **Fig. 3** for reference of data processing steps

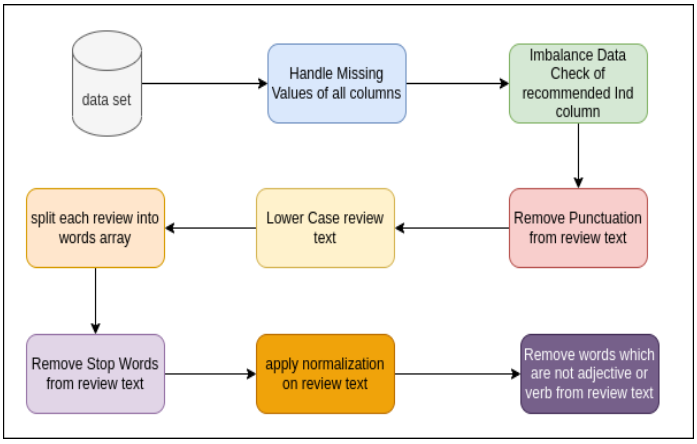


Fig. 3. Data Processing flow of steps.

To handle missing values the author used the SimpleImputer advance technique.

WordCloud **fig. 4** is the result of data processing

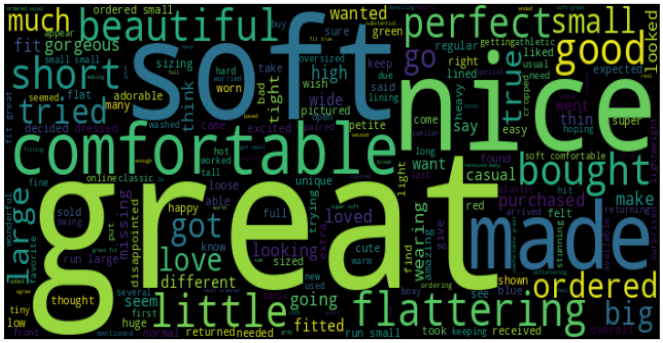


Fig. 4. WordCloud After Data-Processing.

B. EDA

Data visualization techniques are commonly employed in Exploratory Data Analysis (EDA) to analyze and understand datasets by summarizing their key properties. EDA goes beyond the formal modeling or hypothesis testing tasks, aiming to uncover insights that the data may reveal. Additionally, it helps in assessing the appropriateness of statistical methods considered for data analysis. This study will utilize various visualization methods, including line charts, histograms, pie charts, and treemap charts, to facilitate a comprehensive exploration and interpretation of the data.

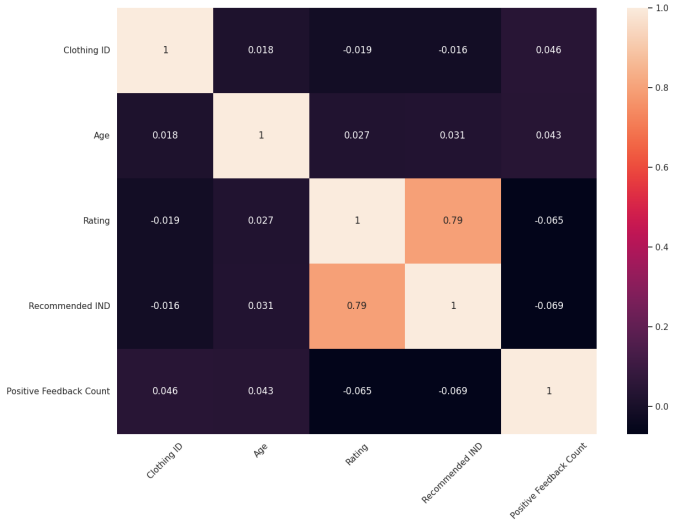


Fig. 5. Correlation Heatmap

The relationship between each column in any data set is determined by the corr() method. Let's look at the correlation between Rating and Recommended IND in f**ig. 5 .** Its value of 0.79 indicates that the two are highly correlated, with a correlation of nearly 1. In layman's terms, people are more likely to recommend clothing with higher ratings.

Some of the data explaining graphs are **fig. 6** , **fig. 7**, **fig. 8 , fig. 9**

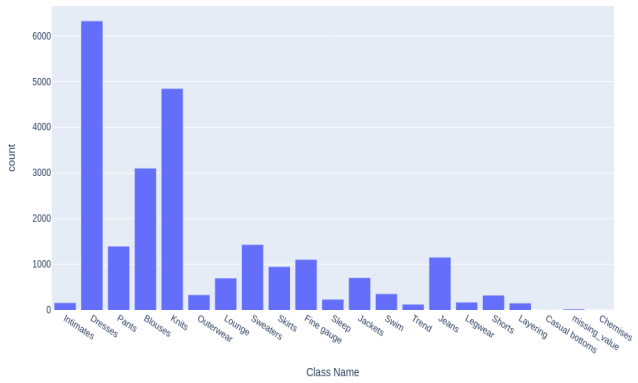


Fig. 6. Distribution of Reviews By Class Name.

Fig. 7. Distribution of Reviews by Rating

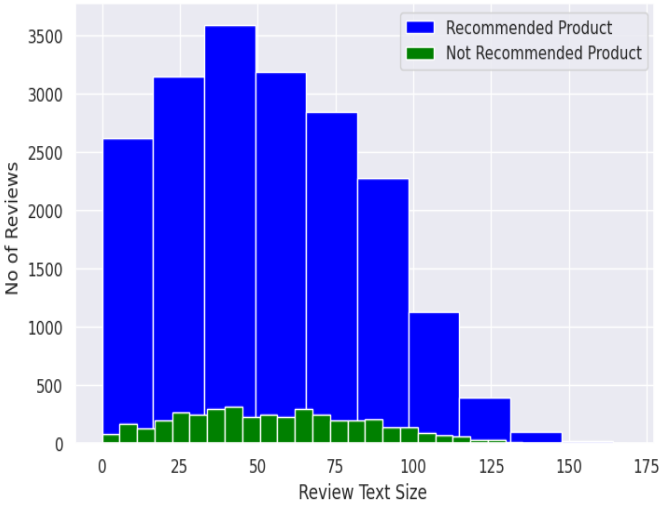


Fig. 8. Distribution of Recommendation by Length of review.

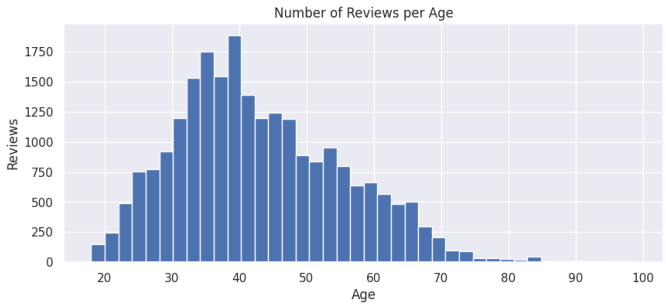


Fig. 9. Distribution of Reviews per Age.

C. Data Pre-Processing

To enhance the accuracy of the training process after EDA analysis pre-processing steps usually performed on processed dataset. Currently the main focus lies on the "Recommended IND " and "Review Text" columns, as they are crucial for solving the author problem. Consequently, The author will remove the other columns from the dataset and proceed to rename the remaining columns. This streamlined approach allows us to concentrate on the essential data required for the author's specific objectives.

Fig. 10. WordCLoud of Recommended Text

D. FEATURE EXTRACTION

To analyze text effectively, it is necessary to convert the text into vector representations. This crucial step, known as "feature extraction" or "vectorization," allows us to represent words as integers or floating-point values. The choice of technique determines the specific encoding method employed. These vector representations are then utilized as inputs for machine learning algorithms, enabling us to leverage text data for analysis and prediction.

In experiment the author is using the TF-IDF technique for feature extraction

E. SPLIT DATASET

To create a model, you need to prepare your data. First, fig. 10 is the distribution of recommendation data according to review count.

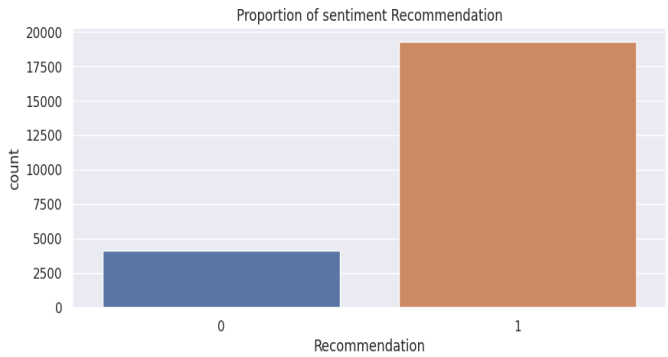


Fig. 11. Imbalance Data of Recommendation

The target class variable exhibits an imbalanced distribution, with a higher prevalence of 'Recommended' values compared to 'Not Recommendation' values

One approach to address the issue of imbalanced class variables is by employing techniques such as oversampling the minority class or undersampling the majority class. To balance the classes, the author utilizes the advanced method SMOTETomek. This technique involves generating synthetic instances of the minority class to augment its representation in the dataset.

The author will split the data set into two parts

- Train Data Set
- Test Data Set

VII. CLASSIFICATION

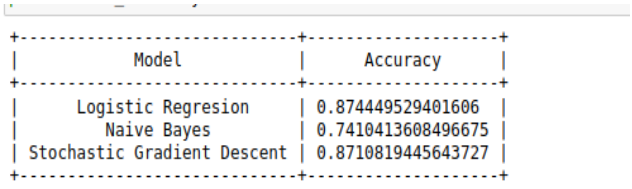
To accomplish the author objective of predicting the recommendation of clothing from reviews, The author will develop three distinct classification models and evaluate their performance to determine the most effective. Following are the models

- Logistic REgression
- Naive Bayes
- Stochastic Gradient Descent

VIII. RESULT

In this project, sentiment classification was utilized to determine whether a product is recommended or not. To enhance the accuracy of predictions, a variety of machine learning algorithms were employed for comparison. The classification algorithms used in the analysis included Logistic Regression, Naive Bayes, and Stochastic Gradient Descent. The dataset used for training and evaluation was sourced from the Woman Clothing Review dataset available on OpenML.

When comparing the models, it becomes challenging to decide on a single model as the top 2 models exhibit very similar scores. However, there is no straightforward answer as to which model is definitely better, as each model performs differently depending on the dataset and conditions. Each modeling algorithm has its own advantages and disadvantages. Therefore, the selection of a specific algorithm should be based on the specific requirements of the analysis, such as accuracy or precision.



Model	Accuracy
Logistic Regresion	0.874449529401606
Naive Bayes	0.7410413608496675
Stochastic Gradient Descent	0.8710819445643727

Fig. 11. Accuracy table of model

From the fig. 11 table of comparison , Logistic Regression and Stochastic Gradient Descent have almost equal percentile of accuracy

IX. DISCUSSION AND CONCLUSION

In conclusion, this study focused on analyzing the Women's E-Commerce Clothing Reviews dataset to predict recommendations using classification techniques. Through the analysis, several important factors were identified that influence women customers' recommendations of clothing items. Quality, appearance, and size emerged as key factors that significantly impact recommendations.

The study reveals that quality, appearance, and size are the key factors that influence these recommendations. In negative feedback, women frequently express disappointment with products that are either too small or too large, poorly fitting, or made from cheap fabric, which gives a negative impression. Conversely, positive comments often highlight the attractive appearance and comfort of the clothes.

Therefore, this anonymous E-commerce platform must prioritize addressing issues related to product size and manufacturing. By doing so, the platform can attract more potential customers and reduce customer churn among women. The utilization of natural language processing and sentiment analysis can provide valuable insights into the challenges faced, not only in women's clothing but also in other real-world sales comments. This approach can be implemented within the E-commerce platform to enhance customer satisfaction and overall improve product recommendation

A. ADDITIONAL WORK

To enhance the area of study the author performed a small experiment “PREDICTION OF RATING FROM REVIEWS”. To predict “Rating” the author is performing following algos

1) Random Forest
2) Support Vector Machines
3) Logistic REgression

B. FUTURE WORK

In this study the author focused on studying the women's attitude toward prediction but in next study the area of womens trend toward online shopping and experience can be explored to make the online shopping more advanced.

X. REFERENCES

Markus, B., Bernd, H., Mathias, K., Andreas, O. and Alexander, S. (2019). Explaining the stars: Aspect-based sentiment analysis of online customer reviews, TwentySeventh European Conference on Information Systems (ECIS2019), StockholmUppsala, Sweden. .