



The Lumbini Journal of
Business and Economics
Peer Reviewed

Machine Learning Segment Customers Based on Their Reviews in E-Commerce Portals: A Fresh Inquiry

Sourav Das¹
Biswajit Biswas²

Abstract

Article Info

Received:

24 December 2024

Revised:

15 February 2025

Accepted:

18 February 2025

Purpose: Customer segmentation is crucial yet challenging, especially in differential pricing. This study highlights the growing role of machine learning (ML) in analyzing vast data to uncover hidden patterns.

Methods: Online reviews including text and pictorial forms are explored in this study as the data set to newly developed machine learning model. With the proliferation of online reviews, authors will get a wealth of data that can be harnessed to perform customer segmentation using ML techniques. This research focuses on how ML techniques can be used to analyze reviews, including text classification, sentiment analysis and clustering algorithms.

Results: Sentiment analysis categorizes reviews as positive, negative, or neutral, while text classification sorts them by topics (e.g., product features, pricing). Clustering identifies distinct customer segments. ML automates large-scale review analysis, revealing customer preferences that manual methods might miss. This study also compares ML techniques to assess their suitability for marketing segmentation.

Conclusion: Leveraging ML for customer segmentation provides a competitive edge. Further research is needed to refine ML models, enabling businesses to optimize products, services, and marketing strategies for targeted customer needs.

Keywords: Customer segmentation, machine learning, clustering, review analysis, sentiment analysis.

I. Introduction

In this dynamic environment everyone is moving forward towards technology to make their tasks and activities easier. As the environment has become digital nowadays so people are depending upon the online reviews to determine their choices. So, in today's digital era, online review has become a powerful resource of information for businesses to understand customer preferences and opinions. With the increasing volume and variety of online reviews, leveraging ML for customer segmentation based on review analysis has gained significant

¹ Sourav Das is a Marketing Manager Punjab National Bank Circle Office North 24 Parganas comtia. sourav@gmail.com

² Biswajit Biswas is an Assistant Professor Department of Business Administration University of Kalyani, India, bbmbaku17@klyuniv.ac.in

attention among marketers and researchers alike. By extracting insights from these reviews using machine learning, businesses can identify distinct customer segments with unique characteristics, allowing for targeted marketing strategies and personalized customer experiences. Customer segmentation is marketing approaches that include dividing a large customer base into smaller groups based on similar characteristics, behaviors, or preferences. Traditionally, customer segmentation has relied on demographical data, such as income, age, gender, and location. However, with the rise of online reviews as a valuable source of customer feedback, businesses now have the opportunity to perform customer segmentation based on the sentiment, topics, and opinions expressed in reviews. Machine learning (ML) techniques offer powerful tools to evaluate large volumes of unstructured review data and uncover meaningful insight. Sentiment analysis, a common machine learning technique, can automatically classify reviews as per positive, negative, or neutral based on the sentiments expressed in the text.

Text classification algorithms can categorize reviews into topics, such as product features, service quality, or pricing, allowing businesses to understand the key themes that customers discuss in their reviews. Clustering algorithms can group similar reviews together based on similarity in language, sentiment, or topics, enabling businesses to identify distinct customer segments with similar preferences or needs. The use of ML for customer segmentation based on review analysis offers several advantages. Firstly, it allows businesses to automatically process and analyze large volumes of reviews, saving time and effort compared to manual analysis. Secondly, machine learning techniques can uncover insights and patterns that may not be apparent through traditional methods, providing a deeper understanding of customer preferences and opinions. Thirdly, customer segmentation based on review analysis enables businesses to tailor their marketing strategies, products, and services to specific customer segments, leading to improved customer targeting and enhanced customer satisfaction and loyalty.

In this paper, authors explored the application of ML techniques for customer segmentation based on review analysis. Authors have discussed various machine learning methods, such as sentiment analysis, text classification, and clustering algorithms, and their potential benefits for customer segmentation. The authors also highlighted the need for further research and development in this area to optimize the use of ML techniques for customer segmentation and to drive businesses success in the dynamic and competitive digital landscape.

Word Cloud

A word cloud is a form of visual representation that illustrates text data, with the size of each word reflecting its frequency or significance. The more frequently a word appears in the text, the larger it appears in the word cloud. Word clouds are a useful tool for rapidly identifying the prominent themes, concepts, or ideas in a given text. They can be generated using a variety of software and tools, which offer users the ability to customize the layout, font, and color scheme of the word cloud. Word clouds are frequently employed in marketing, social media analysis, and data visualization, as they offer a visually appealing and succinct way to convey patterns within extensive amounts of text data.

Correlation Map

A correlation map is a graphical representation that visualizes the correlation between pairs of variables in a given dataset. The correlation between two variables is typically measured using a correlation coefficient, which quantifies the strength and direction of the linear relationship between the variables. Correlation maps are commonly presented in the form of a matrix, where each variable is plotted against all other variables in the dataset. The intensity and color of each cell within the matrix indicate the magnitude and direction of the correlation coefficient between the two variables being compared.

Sentimental Score

Sentiment analysis is a process of using machine learning and natural language processing algorithms to identify and extract the emotional tone conveyed in a text, such as a review or social media post, and assign it a numerical sentiment score. The score usually ranges from -1 to +1, where a value of -1 represents a highly negative sentiment, +1 represents a highly positive sentiment, and 0 indicates a neutral sentiment. Sentiment analysis has numerous applications, including monitoring brand reputation, analyzing customer feedback, and tracking changes in public opinion. However, it's important to keep in mind that the accuracy of sentiment analysis may be impacted by factors such as sarcasm, irony, or subtle nuances in language that can be challenging to detect. Sentiment scores are often used in marketing, social media analysis, and customer feedback analysis to gain understandings into customer opinions, preferences, and attitudes.

In this study, the authors attempted to use natural language processing (NLP) techniques to analyse customer feedback and reviews, and identify common themes and sentiment and use this information to inform product development, customer service, and marketing campaigns.

II. Reviews

Digital social media plays an important part in the industry of hospitality and tourism. The usage of deep learning and machine has become very much useful in the segmentation of market as well as the preferences of customers prediction by social huge data analysis. The clustering of the traveler's online rating and reviews was done by self-organizing map. This method was developed on the basis of unsupervised and robust supervised machine learning techniques (Nilashi et al., 2021). The behavior of the consumers with respect to the hotels which are eco-friendly those are targeted.

And studied the staying behavior of the tourists and their green knowledge, the intention of the customers to stay in a eco-friendly hotel, their satisfaction and intention to return and the green practices of the customers. Authors try to provide a literature review of the study on the point of views of the travelers of green hotels, MCDM, the use of such technique in the industry of hotels and the implication of clustering method in the industry of tourism (Yadegaridehkordi et al., 2021). The customer consumes larger volumes of data because of the functions in which they involved on their devices like mobile; the service of mobile data revenues for mobile providers has also increased the customer segmentation by utilizing the rules of decision tree classification. Here three types of variables has been used to develop the two machine learning models that can be modified accordingly with the customer's needs and there wiliness to pay.

The usage of machine learning techniques with a set of data built upon the data of the customers to generate from the Telecommunication industry because they made it feasible to test upon various classifications for categorizing a group of a particular age of customers and their VIP status (Dullaghan et al., 2017). The development of a new product, earlier segmentation methods which are based upon psychographic, demographic and the information of purchase behavior cannot determine a targeted segment of customers with their unsatisfied needs. Here, IML- based approach has been developed for getting customer segmentation from the reviews of online purchase (Joung et al., 2023). Mehrbakhsh et al. were conceived that the hybrid ML method can be improve the analysis quality of the segmentation. Segmentation of restaurants by using customers' online reviews and their perspectives (Nilashi et al., 2021).

It focused on the ratings of the travelers and the online reviews of the customers on the hotels which are eco-friendly, along with the services like spa and the ordinary hotels. The integration of these techniques of segmentation of hotels which are environmental effective has developed for the very first time on the basis of hospitality and tourism (Nilashi et al., 2019). Comparative exploration of various techniques that is presented in the online retails data. Such ML clustering models which are unsupervised, which are based on frequency, regency and monetary clustering are examined on the basis of the insights that each and

every model provides (Turkmen, 2022). The era of automation technological revolution and information science and the age of machines have gained great attention from the researchers and developers of artificial calculation along with the ML techniques with precision and practical implications have become an immortal disclosure in each discipline (Behena et al. 2021). Customer segmentation by ratings and online reviews can help various types of hotels, including the hotels that have spa services, to improve their information expenditures of marketing. Findings have confirmed that the methods of hybrid machine learning can also be applied as a progressive recommending agent for resorts /spa hotels and can do segmentation by properly using the huge amount of data collected from several online social media platforms (Ali et al., 2019).

ML is mainly a part of Artificial Intelligence. That machine learning has been taken as a main component of the solutions of digitalization which has captured a great attention of the digital era. The proper learning algorithms to achieve the particular requirements of the application (Ray, 2019). The actions of the Telecom Company's customers can create a vast amount of data which can be help to get significant results regarding to customers' needs and behaviors. The primary characteristics of this kind of data are huge number of traits and highly deficiency which bring challenges on every step of the analytics (Alkhsyrs et al. 2020). Customer segmentation is taken immense amount of attention and also it has broadly been used in sector of strategic marketing, majority is using it to develop different techniques (Hiziroglu, 2013). Artificial Intelligence (AI) based modeling has leading consideration to create effective systems and automated which are smart to satisfy our today's wants and needs. Several companies got benefit by using these modeling techniques to improve their marketing strategies (Wang, 2022). Customer segmentation is a method of categorizing the customers into several groups of customers with harmonious behavior that helps the company to get an idea about their customers.

At first the data will be thoroughly analyzed and then prepared for the model training (Zadoo et al., 2022). Customer segmentation techniques are being implemented by the owners everywhere to develop new dimensions in every industry across the globe. Customer analytics perform a significant part in learning and gaining the trust of the customers by analyzing their behaviors. Author should take care of such reasons so that we can sustain our customers in future for and also we have to monitor the dynamic trends and flow of the customer's needs (Ranjam et al., 2022). In this technology-based age every organization try to spread their business wings in e-commerce platforms (Biswas & Sanyal, 2019). In the fastest-growing e-commerce platforms feedback analysis takes an important role in online sales, customer retention, potential buyer, etc. In this work, the authors take into their consideration the three largest e-commerce sites for feedback analysis.

It used the feed-forward ANN supervised learning to develop the model (Biswas et al., 2021). Nowadays e-commerce based web-sites are examining the OCRs to segment their customers. When customers purchase long lasting products, customers and their reviews given only one time for a long period, that is why the score of the review of the product can significantly influenced by the factors of service or can gently be evaluated (Jiang et al., 2015).

From the literature review, authors found that the lack of standardization in review data collection and analysis. Different businesses and platforms use different methods for collecting and analyzing review data, which can make it difficult to compare results across different studies. This analysis can provide insights into customer preferences and sentiments, it may not fully capture the complex behaviors and decision-making processes that drive customer purchase and loyalty. Authors also found that machine learning has the potential to provide valuable insights into customer behavior and preferences, there are still several research gaps that need to be addressed to fully realize it's potential.

III. Methodology

In this research work authors performed the statistical analysis by taking the parameters as Brand, Model, Color, Memory, Storage, Rating, Original Price and Selling Price which gives an idea about the awareness of review that how much it is affecting in online shopping. It is being performed in Jupyter Notebook with the help of machine learning techniques. The details of this analysis are described below. The first step is to collect review data; the data was extracted from Amazon.com using previously developed model. Once the review data is collected, the data was preprocessed to clean and prepare it for analysis. This may involve tasks such as data cleaning, text normalization, removing irrelevant information, and handling missing values or outliers. Data preprocessing is crucial to ensure the quality and reliability of the data used for customer segmentation. After completing the data processing feature extraction was done. In this stage relevant features or attributes were selected from the review data to represent the customer feedback. It included extracting keywords, phrases, or other contextual information from the reviews, such as product features, customer experiences. Feature extraction was important to diminish the dimensionality of the data and improves the efficiency level of the ML algorithms. After the review data was relevant and preprocessed features are extracted, we have trained the ML with this data. Next, the authors used sentiment analysis as opinion mining for customer segmentation based on review analysis. It included using NLP techniques to analyze the emotional or sentiment tone of the reviews, such as (positive, negative, or neutral). After completing all step we used clustering techniques to define the segment of customer.

IV. Results and Discussion

The python packages used in the model is imported first for data analysis and prediction. In Fig.1 have taken numpy package for mathematical operations, pandas for data analysis, matplotlib.pyplot package for plotting histogram and scattering plot, seaborn package is taken for visualization of data.

Figure 1

Package Import

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns #visualize
from textblob import TextBlob
import nltk
from nltk import word_tokenize, FreqDist
from nltk.corpus import stopwords
import re
import wordcloud
from PIL import Image
```

In Figure 2 the dataset is imported and viewing the shape of dataset whether the data is imported properly or not for analysis.

Figure 2

Data Import

```
In [2]: dataset = pd.read_csv('Flipkart_Mobiles_2020.csv')
```

```
In [3]: dataset.shape
```

```
Out[3]: (3114, 8)
```

```
In [4]: dataset.head(5)
```

```
Out[4]:
```

	Brand	Model	Color	Memory	Storage	Rating	Selling Price	Original Price
0	OPPO	A53	Moonlight Black	4 GB	64 GB	4.5	11990	15990
1	OPPO	A53	Mint Cream	4 GB	64 GB	4.5	11990	15990
2	OPPO	A53	Moonlight Black	6 GB	128 GB	4.3	13990	17990
3	OPPO	A53	Mint Cream	6 GB	128 GB	4.3	13990	17990
4	OPPO	A53	Electric Black	4 GB	64 GB	4.5	11990	15990

In figure 3 authors have used is null() command to detect the null value from the entire dataset, when the null values get detected we drop out the null value from the given set of data by using dropna() command and display the remaining data below.

Figure 3

Drop Null Values and Create New Dataset

```
In [7]: dataset.isnull().sum()
```

```
Out[7]: Brand      0
Model      0
Color      0
Memory     43
Storage    39
Rating     144
Selling Price 0
Original Price 0
dtype: int64
```

```
In [8]: new_dataset = dataset.dropna()
```

```
In [9]: print (new_dataset)
```

```

      Brand  Model  Color  Memory  Storage  Rating  Selling Price  \
0      OPPO  A53  Moonlight Black  4 GB   64 GB   4.5         11990
1      OPPO  A53    Mint Cream  4 GB   64 GB   4.5         11990
2      OPPO  A53  Moonlight Black  6 GB  128 GB   4.3         13990
3      OPPO  A53    Mint Cream  6 GB  128 GB   4.3         13990
4      OPPO  A53  Electric Black  4 GB   64 GB   4.5         11990
...      ...    ...      ...      ...      ...      ...      ...
3109  SAMSUNG  M52 5G  Blazing Black  6 GB  128 GB   4.3         25990
3110  SAMSUNG  M52 5G    Icy Blue  6 GB  128 GB   4.3         25489
3111  SAMSUNG  M52 5G    Icy Blue  8 GB  128 GB   4.3         27239
3112  SAMSUNG  M52 5G  Slate Black  8 GB  128 GB   4.2         22989
3113  SAMSUNG  M52 5G    Sky Blue  8 GB  128 GB   4.2         20350

```

In Figure 4 the data is described and displayed a statistical summary of the data frame. If the data sets are containing numerical values then data. Describe () command is used for graphical representation which measures description of the data in form of descriptive statistics.

Figure 4

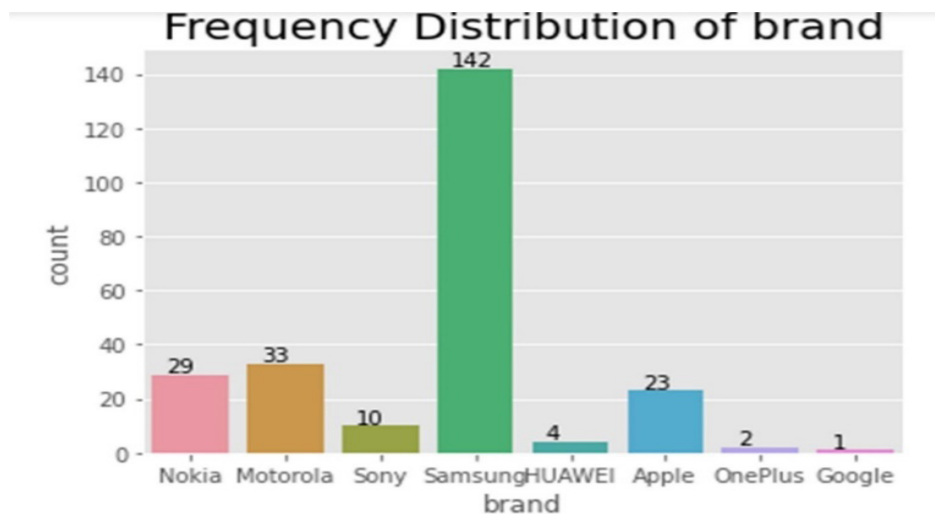
*Data Description*In [6]: `dataset.describe()`

Out[6]:

	Rating	Selling Price	Original Price
count	2970.000000	3114.000000	3114.000000
mean	4.243098	26436.625562	28333.473025
std	0.271991	30066.892622	31525.599889
min	2.300000	1000.000000	1000.000000
25%	4.100000	9990.000000	10030.250000
50%	4.300000	15000.000000	16889.500000
75%	4.400000	28999.000000	31500.000000
max	5.000000	179900.000000	189999.000000

In figure 5 we are expressing the frequency level of selling of mobile phones from different brands. Where among all mobile brand Samsung has the higher frequency level than other brands. Which mean the mobile phones from Samsung brand were sold more than other?

Figure 5

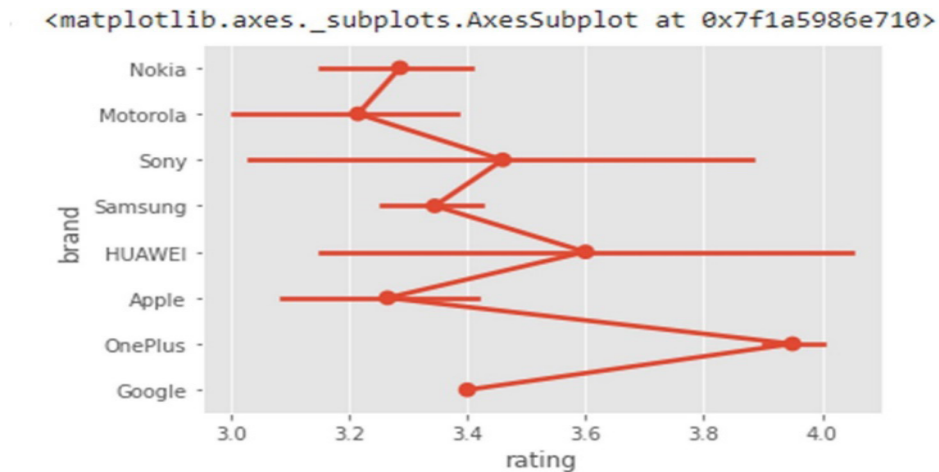
Frequency Distribution of Brand

In figure 6 we are displaying the customer ratings of different brands. Where the rating of Google is 3.4 and one plus has the rating within 3.9-4.0, Apple has a rating between 3.1-4.5, Huawei has a rating between 3.1-4.1, Samsung has a rating between 3.3-3.5, Sony has a rating between 3.1-3.9, Motorola has a rating between 3.0-3.4 and Nokia has 3.1-3.4 rating. Which indicates that Motorola has the lowest ratings among all the brands, where Huawei has

the highest ratings than others.

Figure 6

Rating VS Brand



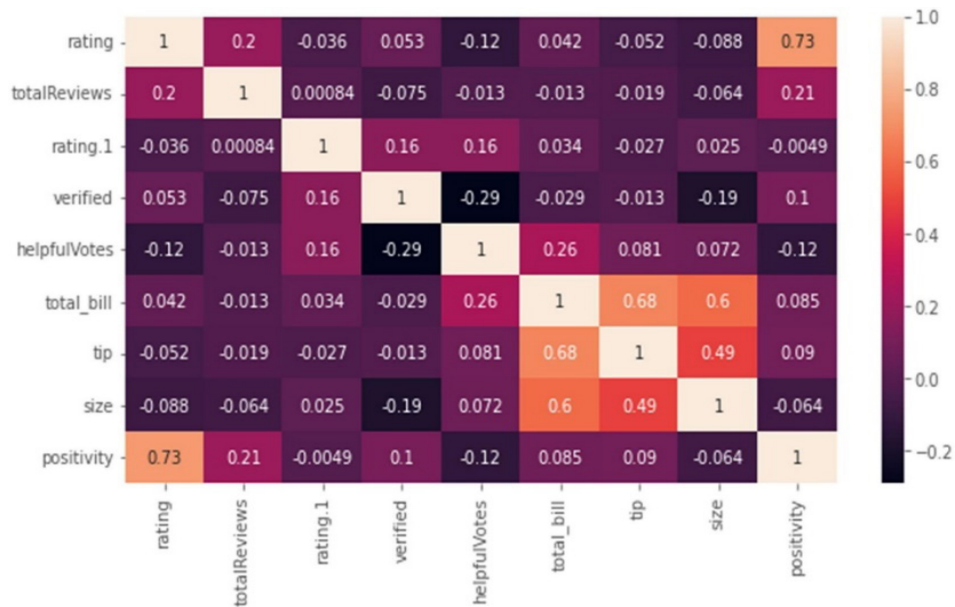
In fig. 7 is visualising the words by making those words more bold and highlighted accordingly. It is done by analyzing the words which are used by the customers to expressed their experience about the product.

Figure 7

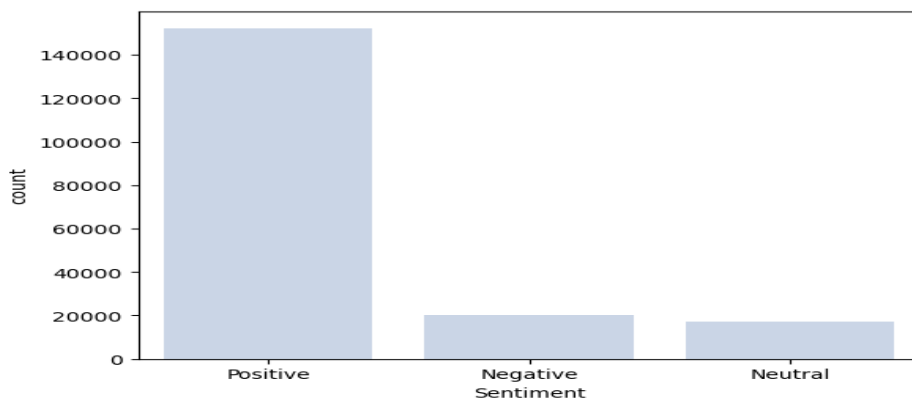
Word Cloud



By the figure 8, we are trying to show the correlations and coefficients among all the possible pair of variables. Each cell is displaying the correlation between two variables. This matrix is basically used to summarize the dataset and the value of this matrix always lies between -1 to +1.

Figure 8*Correlation Map*

The figure 9 is a language processing method which displays the character of the data provided by the customers by giving feedback on a particular product. Where positive means the customers are very much satisfied, negative signifies dissatisfaction and neutral means not much satisfied & not much dissatisfied. As we can see in the given figure more than 1400000 customers are highly satisfied where less than 200000 customers are dissatisfied and neutral

Figure 9*Sentimental Score*

V. Conclusions and Implication

Customer segmentation based on review analysis using machine learning presents a promising approach for businesses to gain valuable insights from online reviews and tailor their marketing strategies to specific customer segments. Through sentiment analysis, text classification, and clustering algorithms, machine learning can automatically process and analyze large volumes of unstructured review data, uncover hidden patterns, and identify distinct customer segments with unique characteristics. The use of ML techniques for customer segmentation based on review analysis offers several advantages, including time and effort savings, deeper understanding of customer preferences, and improved customer targeting. By leveraging these insights, businesses can develop targeted marketing strategies, personalized customer experiences, and enhanced customer satisfaction and loyalty.

In this research work, authors built a model using machine learning algorithms that businesses can use to gain a great appreciation of their customers' preferences, behaviors, and sentiments. By identifying and analyzing customer segments based on their feedback, businesses can tailor their marketing strategies to better meet the needs and expectations of different customer groups. This developed model also enables businesses to develop targeted marketing campaigns and promotions. By tailoring marketing messages to specific customer segments, businesses can improve the relevance and effectiveness of their marketing efforts, which can lead to increased customer engagement, loyalty, and sales.

Accuracy score and effectiveness of this developed model using ML depend heavily on the quantity and quality of review data available. Limited or biased review data may result in inaccurate or incomplete segmentation results, leading to suboptimal marketing strategies. Also, reviews are subjective and context-dependent, and the same sentiment expressed in different reviews may have different implications. Machine learning models may struggle to accurately capture the nuances of sentiment analysis in reviews, resulting in potential misinterpretations or misclassifications.

In the future, this model using machine learning lies in the development of more advanced personalization techniques. This may include leveraging other data sources, such as customer transaction data, browsing behavior, or social media data, to create more detailed and accurate customer segments, enabling businesses to deliver highly personalized marketing strategies. Authors try to include real-time customer segmentation based on review analysis using machine learning that can enable businesses to adapt their marketing strategies in real-time based on changing customer sentiments and preferences. Dynamic segmentation techniques that continuously update and refine customer segments (CS) based on real-time review data can provide businesses with more relevant and timely insights for marketing decision-making.

Reference

- Ahani, A., Nilashi, M., Ibrahim, O., Sanzogni, L., & Weaven, S. (2019). Market segmentation and travel choice prediction in Spa hotels through TripAdvisor's online reviews. *International Journal of Hospitality Management*, 80, 52-77.
- Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. *Journal of Big Data*, 7(1), 9.
- Behera, L., Nanda, P., Mohanta, B., Behera, R., & Patnaik, S. (2021). Machine learning for customer segmentation through bibliometric approach. In *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019* (pp. 189-206). Springer Singapore.
- Biswas, B., Sanyal, M. K., & Mukherjee, T. (2021). Feedback analysis for digital marketing in India: Empirical study on Amazon, Flipkart, and Snapdeal. *International Journal of Online Marketing (IJOM)*, 11(1), 78-88.

- Biswas, B., & Sanyal, M. K. (2019, January). Soft intelligence approaches for selecting products in online market. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 432-437). IEEE.
- Dullaghan, C., & Rozaki, E. (2017). Integration of machine learning techniques to evaluate dynamic customer segmentation analysis for mobile customers. arXiv preprint arXiv:1702.02215.
- Hiziroglu, A. (2013). Soft computing applications in customer segmentation: State-of-art review and critique. *Expert Systems with Applications*, 40(16), 6491-6507.
- Joung, J., & Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management*, 70, 102641.
- Nilashi, M., Samad, S., Minaei-Bidgoli, B., Ghabban, F., & Supriyanto, E. (2021). Online reviews analysis for customer segmentation through dimensionality reduction and deep learning techniques. *Arabian Journal for Science and Engineering*, 46(9), 8697-8709.
- Nilashi, M., Ahmadi, H., Arji, G., Alsalem, K. O., Samad, S., Ghabban, F., & Alarood, A. A. (2021). Big social data and customer decision making in vegetarian restaurants: A combined machine learning method. *Journal of Retailing and Consumer Services*, 62, 102630.
- Nilashi, M., Mardani, A., Liao, H., Ahmadi, H., Manaf, A. A., & Almukadi, W. (2019). A hybrid method with TOPSIS and machine learning techniques for sustainable development of green hotels considering online reviews. *Sustainability*, 11(21), 6013.
- Ray, S. (2019, February). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 35-39). IEEE.
- Ranjan, A., & Srivastava, S. (2022, November). Customer segmentation using machine learning: A literature review. In *AIP Conference Proceedings* (Vol. 2481, No. 1, p. 020036). AIP Publishing LLC.
- Turkmen, B. (2022). Customer Segmentation with Machine Learning for Online Retail Industry. *The European Journal of Social & Behavioural Sciences*.
- Wang, C. (2022). Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach. *Information Processing & Management*, 59(6), 103085.
- Yadegaridehkordi, E., Nilashi, M., Nasir, M. H. N. B. M., Momtazi, S., Samad, S., Supriyanto, E., & Ghabban, F. (2021). Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques. *Technology in Society*, 65, 101528.
- Zadoo, A., Jagtap, T., Khule, N., Kedari, A., & Khedkar, S. (2022, May). A review on churn prediction and customer segmentation using machine learning. In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)* (Vol. 1, pp. 174-178). IEEE.