# Enhancing Product Recommendation Systems through Text Classification of Customer Reviews

Abdullah Al Fahim
*Department of CSE*
*Brac University*
Dhaka, Bangladesh
abdullah.al.fahim1@g.bracu.ac.bd

Golam Rabbi
*Department of CSE*
*Brac University*
Dhaka, Bangladesh
golam.rabbi@g.bracu.ac.bd

Ananna Acharjee
*Department of CSE*
*Brac University*
Dhaka, Bangladesh
ananna.acharjee@g.bracu.ac.bd

Arannya Dhar Deb
*Department of CSE*
*Brac University)*
Dhaka, Bangladesh
email address or ORCID

*Abstract*—This study focuses on the urgent requirement to improve product recommendation systems by utilizing sophisticated text categorization methods applied to user evaluations. Recognizing the importance of user-generated material, namely in the form of reviews, we provide a system that combines advanced natural language processing and machine learning methods to extract relevant insights. The goal is to enhance the accuracy and customization of the recommendation system by collecting the attitudes and preferences of users that are expressed in textual data. An assessment of the framework will be carried out using a wide range of datasets, guaranteeing its adaptability across different product areas. The expected results have the potential to revolutionize tailored product suggestions in e-commerce, providing platforms with a tool to better understand and meet user requirements. This research enhances customer pleasure and engagement, leading to a more informed and user-centric online buying experience in the growing field of e-commerce.

## I. INTRODUCTION

In the dynamic realm of online commerce, the vast array of items requires advanced recommendation algorithms to assist consumers in making well-informed decisions. In the midst of this digital marketplace, customer evaluations are seen as a useful collection of user experiences and opinions. They provide a wealth of information that may greatly improve the effectiveness of recommendation systems. This study aims to connect unorganized textual data, particularly customer reviews, with the enhancement of product suggestions using sophisticated text categorization methods. Current recommendation systems, however useful, sometimes encounter difficulties in precisely capturing the subtleties of human preferences and moods. This study aims to provide a complete framework that smoothly combines text categorization models with recommendation systems by utilizing natural language processing and machine learning. The purpose of this integration is to analyze the hidden information in customer reviews with the goal of improving the accuracy, relevance, and customization of product suggestions.Our objective in this research is to analyze the complexities involved and pinpoint the shortcomings of existing systems. Additionally, we aim to introduce a fresh strategy that enables e-commerce platforms to offer personalized and user-focused suggestions. This project aims to analyze the complex user attitudes included in textual data. Its findings contribute to the continuing discussion on improving recommendation systems, ultimately leading to a more user-friendly and enjoyable online shopping experience.

## II. RELATED WORKS

Natural language processing (NLP) is being used in e-commerce to automate product evaluation and sentiment analysis. This study examines sentiment prediction using machine learning and deep learning models like BERT, Glove, Elmo, and Fast Text word embedding. The research highlights how natural language processing (NLP) transforms market analysis, staff well-being, and consumer satisfaction forecasts. It also shows how NLP can manage massive textual data and derive valuable insights. It also acknowledges the difficulties of analyzing consumer feedback and the changing NLP landscape across several sectors.

The author evaluates numerous e-commerce recommender systems algorithms and highlights the drawbacks of using simply user ratings, such as sparsity and unreliability. The suggested method generates ideas using user evaluations, sentiment analysis, and item-based collaborative filtering. It acknowledges limits such as a heavy reliance on sentiment analysis, limited comparability with other review-based systems, and a limited Amazon dataset assessment scope. The work also ignores cold-start, data sparsity, and scalability issues. It also does not analyze their approach's statistical significance or computing cost. To further comprehend the review-based recommendation system's efficacy and applicability, the book advises more extensive examinations and comparisons with comparable systems.

The article presents DeepCoNN, a dual neural network model that incorporates natural language processing (NLP) techniques to enhance Yelp recommendations. DeepCoNN achieves higher accuracy compared to baseline models and sentiment-analysis approaches, as measured on Yelp's dataset using RMSE, MAE, and FCP measures. The identified limits include the usage of fixed-length vector representation, the absence of temporal data, and the necessity for measurement of user enjoyment and business impact. These constraints indicate areas where improvements can be made.

The authors present a pioneering algorithm that significantly advances recommender systems. By advocating for the exploration of multi-category fuzzy preferences, the model

transcends binary preferences, offering a nuanced understanding of user preferences and enhancing similarity calculations. Additionally, the introduction of weighted label attributes provides a sophisticated means to represent item characteristics, promising more accurate rating predictions by considering the varying importance of different attributes.The paper demonstrates a balanced perspective by highlighting both the strengths and limitations of the proposed algorithm, fostering transparency. The identified future research directions, including the expansion of fuzzy preferences and refined attribute weighting, showcase a forward-thinking approach. The paper emerges as a commendable contribution, combining innovation, transparency, and strategic foresight to advance the field of recommender systems.

## III. DataSet

The Mi(Xiaomi) mobiles dataset, sourced from Kaggle.com, provides a comprehensive collection of mobile devices manufactured by MI. The dataset, which includes 549 rows and 8 columns, includes model names, price, ratings, image URLs, storage capacity, RAM configuration, os processor, network technologies, and battery specifications. The data is analyzed using textual data, which includes user ratings and reviews, price, storage capacity, RAM configuration, os processor, network technologies, and battery specifications. The dataset file in CSV format contains textual data pertaining to mobile devices, along with ratings for price and overall ratings. The text classification analysis classifies the dataset into three distinct classes: neutral, positive, and negative. The data is essential for conducting performance analysis and understanding the diverse range of mobile devices manufactured by MI.

### A. Data Classfication

The data set consists of product information like the name,rate,price,image.Along with hardware specifications such as internal storage,RAM,memory card slot type.And other categories such as operating system.network connectivity and batteries information.

## IV. Methodology

### A. Data Preprocessing

To achieve efficient data preprocessing of our dataset containing 549 rows and 8 columns, prioritize addressing missing values, standardizing numerical features such as price and ratings, encoding categorical variables like model names and OS processors, and normalizing text data like picture URLs. Rectify anomalies, guarantee standardized formats for storage capacity and RAM arrangement, and verify coherence in network technologies and battery specifications. It is advisable to apply feature scaling to numerical features and utilize techniques such as one-hot encoding for categorical attributes. By preparing the dataset, it ensures that the analysis and modeling process is robust, leading to more accurate and reliable outcomes in machine learning.

### B. Training Models

1) **Naive Bayes Classifier:** Naive Bayes is a classification technique that uses Bayes' theorem and assumes that the features describing an observation are independent of each other, given the class label. Text categorization is frequently employed for spam filtering and sentiment analysis.The formula of it:

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)} \qquad (1)$$

In Naive Bayes, the "naive" assumption is that the features are conditionally independent given the class. This means that the presence or lack of one feature does not have any influence on the presence or absence of any other feature.

2) **BERT (Bidirectional Encoder Representations from Transformers):** BERT, which stands for Bidirectional Encoder Representations from Transformers, is a pre-trained natural language processing paradigm that Google introduced in 2018. The model employs the Transformer architecture and has been purposefully designed to understand bidirectional context in text. This capability allows it to discern the meaning of individual words in the context of the entire sentence. BERT is subjected to a comprehensive pre-training process using vast quantities of textual data. Following this, the model can be refined to execute various natural language processing tasks, including but not limited to text classification, named entity recognition, and question answering. The attention score $\alpha_{ij}$ in BERT is computed using the following formula:

$$\alpha_{ij} = \frac{e^{(W_Q q_i)^T W_K k_j}}{\sqrt{d_k}} \qquad (2)$$

Here:
- $q_i$ and $k_j$ are the query and key vectors, respectively, for the $i$-th and $j$-th words in the input sequence.
- $W_Q$ and $W_K$ are the weight matrices for the query and key projections.
- $d_k$ is the dimensionality of the key vectors.

The attention scores are subsequently used to calculate a weighted summation of the values V (value vectors) in order to obtain the output of the attention mechanism. By using the self-attention method, BERT is better able to understand the contextual meaning of words because it can take into account the entire context of a word when making predictions. BERT's pre-training entails masked language modelling, a process in which random phrases are obscured, and the model is trained to anticipate these masked words by utilising the contextual information provided by the surrounding words.

## References

[1] Add later

[2] Add later
[3] Dr. N. Srinivasan, Anandaraj Selvaraj,Mobile Based Data Retrieval using RDF and NLP in an Efficient Approach ,January 2018
[4] Zhan Su, Haochuan Yang, Jun Ai, Enhancing recommender systems with fuzzy preference, vector similarity, and user community for rating prediction, August 28, 2023.