IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2024

DOI: 10.17148/IJARCCE.2024.131005

Improving E-commerce Text Detection: A Comparative Study of Hybrid Approaches

Md. Sadiq Iqbal [1], Mohammod Abul Kashem [2], Mohammed Ibrahim Hussain [3], Akash Kumar Pal [4], Md. Rifat-Uz-Zaman [5]

Associate Professor, Department of Computer Science and Engineering, Dhaka University of Engineering & Technology, Gazipur, Bangladesh[1]

Professor, Department of Computer Science and Engineering, Dhaka University of Engineering & Technology, Gazipur, Bangladesh[2]

Assistant Professor, Department of Computer Science and Engineering, Bangladesh University, Dhaka, Bangladesh[3]

Lecturer, Department of Computer Science and Engineering, Bangladesh University, Dhaka, Bangladesh[4]

Lecturer, Department of Computer Science and Engineering, Bangladesh University, Dhaka, Bangladesh[5]

Abstract: Text recognition for E-Commerce boosts search engine performance, facilitates product discovery, enhances user experience, lets you create personalized recommendations, and simplifies inventory control. We've developed a novel Ensemble SVM Multinomial Naive Bayes Approach in our research projectthat is specifically designed to detect e-commerce text. Our dataset had four different classes: books, electronics, household, and clothing and accessories. It contained 50,425 numeric values.Our impressive training accuracy of 99.83% and validation accuracy of 98.35% were attained by applying this state-of- the-art model. The precision of e-commerce text detection has advanced significantly with this accomplishment. We truly believethat our detection technology is capable of what it does. In our opinion, it offers a sound and useful approach to the analysisof text related to online commerce. In addition, we anticipate its integration serving as a cornerstone of future e-commerce sections, promising improved functionality and precision, which excites us about its future potential to refine the e-commerce scene.

Index Terms: component, formatting, style, styling, insert.

I. INTRODUCTION

In an effort to increase computer systems' capacity to produce and comprehend human language, natural language processing, or NLP, has grown in importance over the years as a study area [1] [2]. With the use of machine learning (ML) algorithms and a vast amount of textual data, a big language model that can produce language that is similar to that of a person has been developed as a result of recent advances in this field. For natural language processing, transformers are among the most effective instruments. It may be split into two main sections essentially [3]. Both the decoder and the encoder. With a text sequence as input, the encoder portion generates an encoded representation sequence. Text recognition, extraction, or identification from documents or pictures used in e-commerce contexts is known as e-commerce text detection. Text extraction [4]–[6]from product photos, scanned documents, webpages, and other visual material associated with online sales and shopping is the main goal here. The objective is to employ technology [7] to examine photographs and identify any text that may be contained in them, frequently utilizing computer vision and machine learning methods.

Once the text has been retrieved, it may be utilized for a number of tasks, such as organizing items into catalogs, gathering data for searches, facilitating automatic translation, or enhancing accessibility for those with visual impairments. Text detection [8] may be applied to e-commerce to extract product details from photographs, such as name, brand, specs, or price, therefore facilitating consumers' search and discovery of desired items. To reliably extract and interpret text from visual imagery, e- commerce text detection frequently uses a variety of methods and tools, including deep learning models and optical character recognition (OCR).

To identify E-commerce text in a dataset with four different classes—Electronics, Household, Books, and Clothing and accessories. we created the Ensemble SVM-Multinomial Naive Bayes model. With a validation accuracy of 98.35% and a training accuracy of 99.83%, our model has demonstrated outstanding performance.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2024

DOI: 10.17148/IJARCCE.2024.131005

These remarkable outcomes demonstrate the model's extraordinary capacity to recognize text connected to e-commerce. The key contributions of the study are summed up as follows:

- To detect e-commerce text, we create a set of benchmark trials using machine learning.
- To determine the best result, we put the Ensemble SVM- Multinomia NB model into practice.
- We got the greatest results compared to other research by combining the Machine Learning approach and the SVM-Multinomia NB algorithm

The paper is structured so that the previous approaches to the same problems are presented in Section II, and our suggested approach is presented in Section III. The study's results are contained in Section IV. Section V marks the article's conclusion.

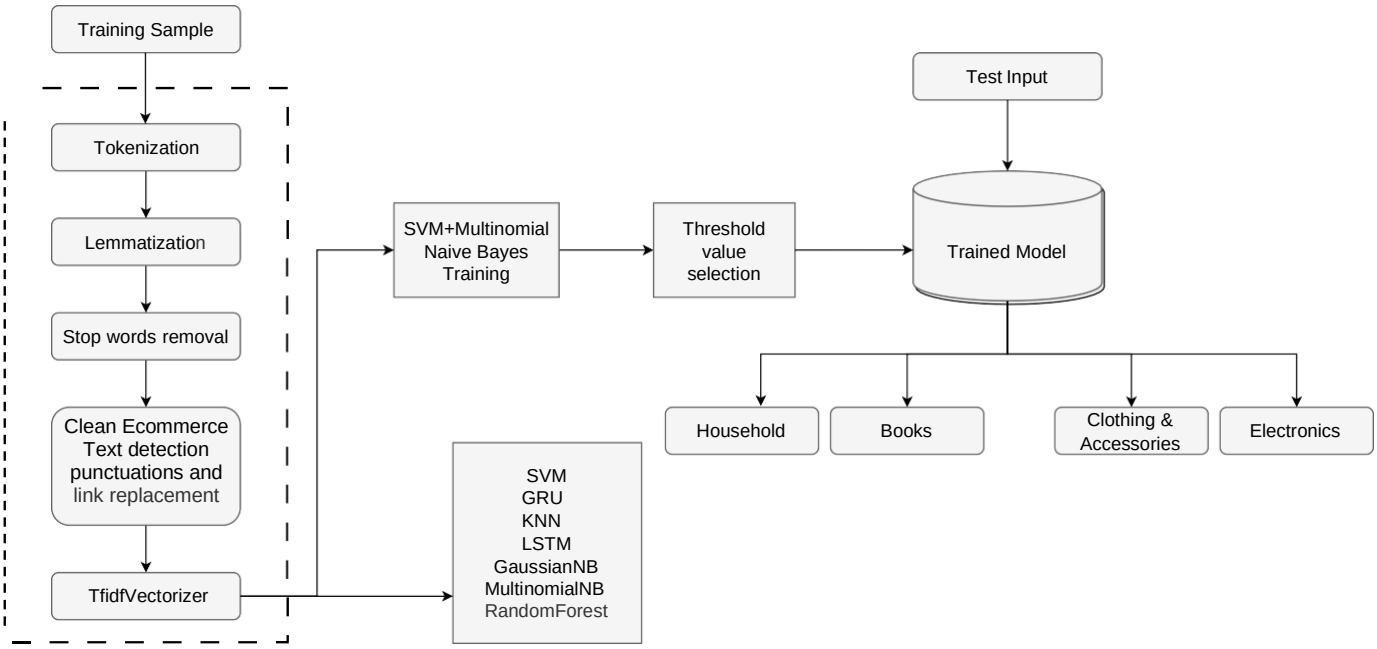

Fig. 1. SVM-Multinomial Naive Bayes Architecture for E-Commerce Text Detection

II. RELATED WORK

Text categorization is essential in natural language processing, especially in commerce. This work involves properly categorizing a statement or document from a preset collection. Text detection research in E-commerce is scarce.

Ganesh et al. [9] Misleading news and product reviews were manufactured by TGMs. They designed detectors to distinguish TGM-generated text from human-written text to pre- vent abuse. Recent NLP and ML advances improved English detection. Despite its relevance, no publication has evaluated the fast-growing literature or introduced novices to primary research issues. Critically examining this material helped me understand the problem. They meticulously studied the state- of-the-art detector's faults and discussed future research in this exciting subject. MADHUMATHI et al. [10] independently gathered E-commerce reviews. Databases held raw web crawler text. NLP cleaned data. It employs computer algorithms. They want human connection. The procedure was mechanized and data was cleaned for efficiency. Natural Language Processing aggregated opinions after data cleaning. NLP methods included Sentiment Analysis, Topic Modeling, and Text Generation.

Souradip et al. [11] addressed the LLM text-human information distinction problem for various applications. They showed it was achievable except when human and machine text distributions were comparable across support. They observed that human-like machine-generated text increased the detection sample size. They allowed multi-sample detector study by restricting AI-generated text detection sample complexity. Better identification was seen in Xsum, Squad, IMDb, and Kaggle FakeNews. OBERTa-Large/Base-Detector and GPTZero vs. GPT-2, GPT-3.5-Turbo, Llama-2-13B-Chat-HF, and Llama-2-70B-Chat Fitting OpenAI's sequence length data theoretically supports these

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2024

DOI: 10.17148/IJARCCE.2024.131005

conclusions. Luciano et al. [12] discussed GPT- 3, a third-generation autoregressive language model that uses deep learning to construct human-like phrases, was honored. GPT-3 failed three math, semantic (Turing Test), and ethics examinations. Industrializing low-cost, automated semantic object generation had several effects. Chang et al. [13] Discussed Reddit, Facebook, Twitter, and Instagram text bot growth. These AI entities quickly spread disinformation and fraud on social media, swaying public opinion on important political, economic, and social matters. They found AI-bot disinformation. They tested a machine-learning system against a Generative Pre-trained Transformer (GPT) to detect AI- and human-generated texts. Their detection classification accuracy was 97%–99% depending on the loss function. Crothers et al. [?] used Text from machines becomes hard to discern from human-written material. They showed these tendencies using ChatGPT, published shortly after their first research. Abuse pathways limited advanced NLG systems. The largest machine-generated text detection study examined modern NLG threat models. They said reliable detection systems must be fair, strong, and accountable.

Lack of E-commerce Effectiveness The main concern is text detection. Additionally, current approaches perform poorly. Our SVM-Multinomial Naive Bayes technique has garnered historical attention for solving this difficulty, attracting scholars to this study field.

III. METHODOLOGY

Our E-commerce text identification method makes use of a complex model that combines Multinomial Naive Bayes and Support Vector Machine. This potent combination yields an impressive detection capability, allowing precise and effective identification on our platform.

A. Dataset Analysis and Discussion

Our dataset, which we obtained from Kaggle, was divided into four categories: books, electronics, household goods, clothing, and accessories. This 36 MB dataset (available in CSV format) contains 50,425 entries of numerical data. We used the industry standard 80-20 split for training and testing while building our model. To be more precise, 40,324 data entries (80% of the dataset) were set aside for training, which helped our model identify trends and connections in the data. The remaining 10,085 entries (20% of the dataset) were then set aside for testing so that we could assess the accuracy and performance of the model. This methodology guarantees resilience and efficiency when addressing diverse text identification assignments in the e-commerce sector. Figure 1 shows the data distribution.

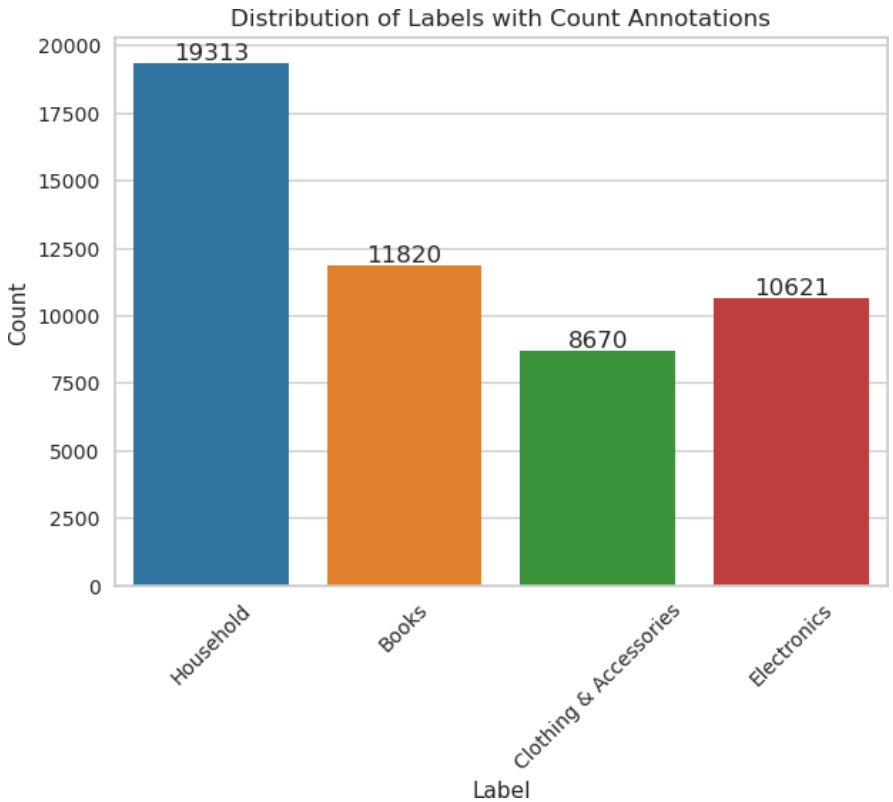

Fig. 2. Data Distribution of four class

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2024

DOI: 10.17148/IJARCCE.2024.131005

We analyse the distribution of text length across different levels. The image below displays a graphical depiction.

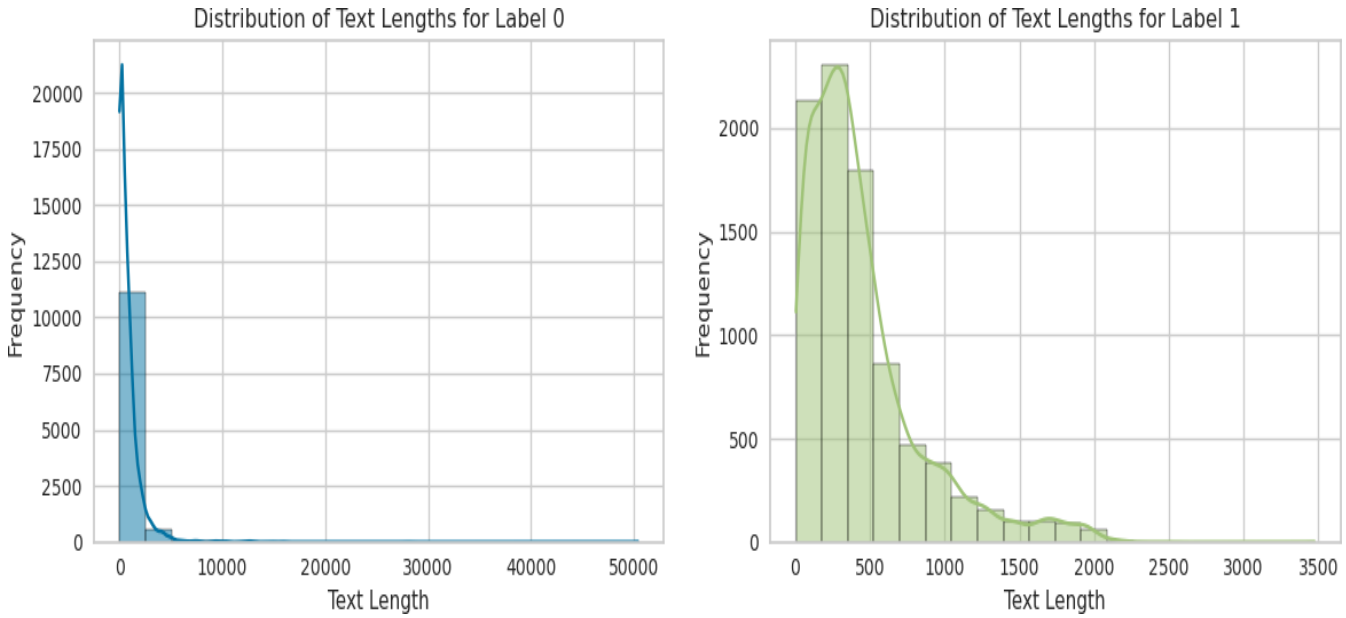

Fig. 3. Distribution Of text Length

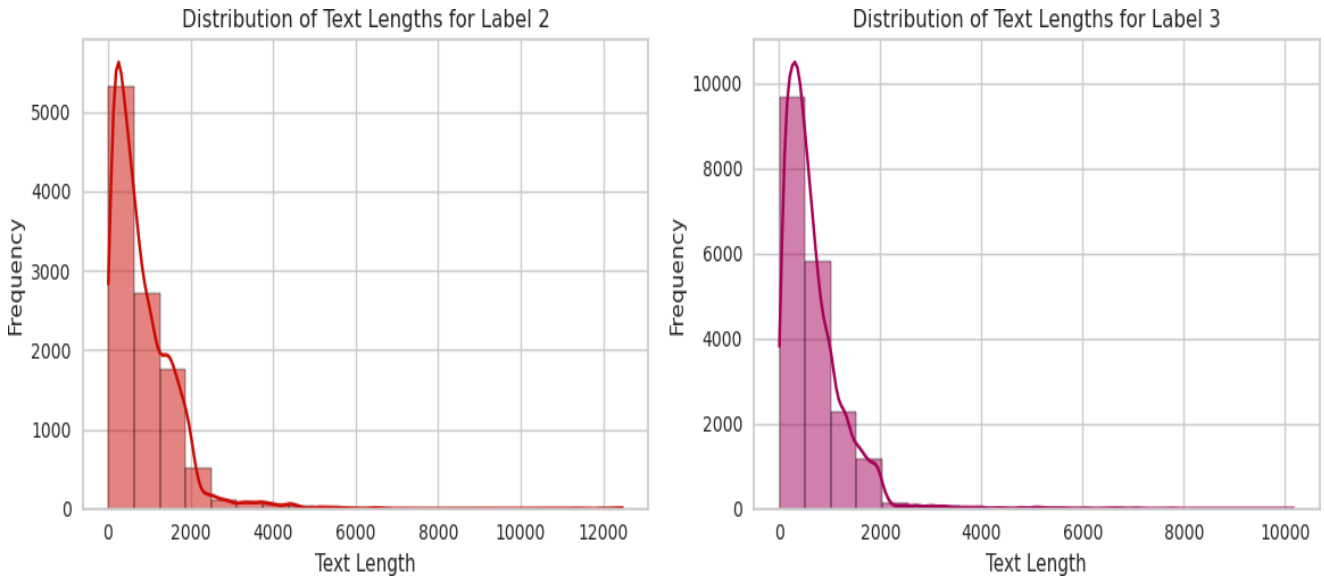

Fig. 4. Distribution Of text Length

B. Data pre-processing

We employed many effective techniques for data preparation, including tokenization, lemmatization, stop word removal, cleaning ecommerce text by detecting punctuations and replacing links, and using Tf-idf Vectorizer. Within our dataset, we've remove duplicate word as top 30 terms used in text processing. The image below shows the graphical depiction of these terms, providing a visually understandable understanding of their frequency and importance. In figure 4 we shows the Graphical Representation.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2024

DOI: 10.17148/IJARCCE.2024.131005

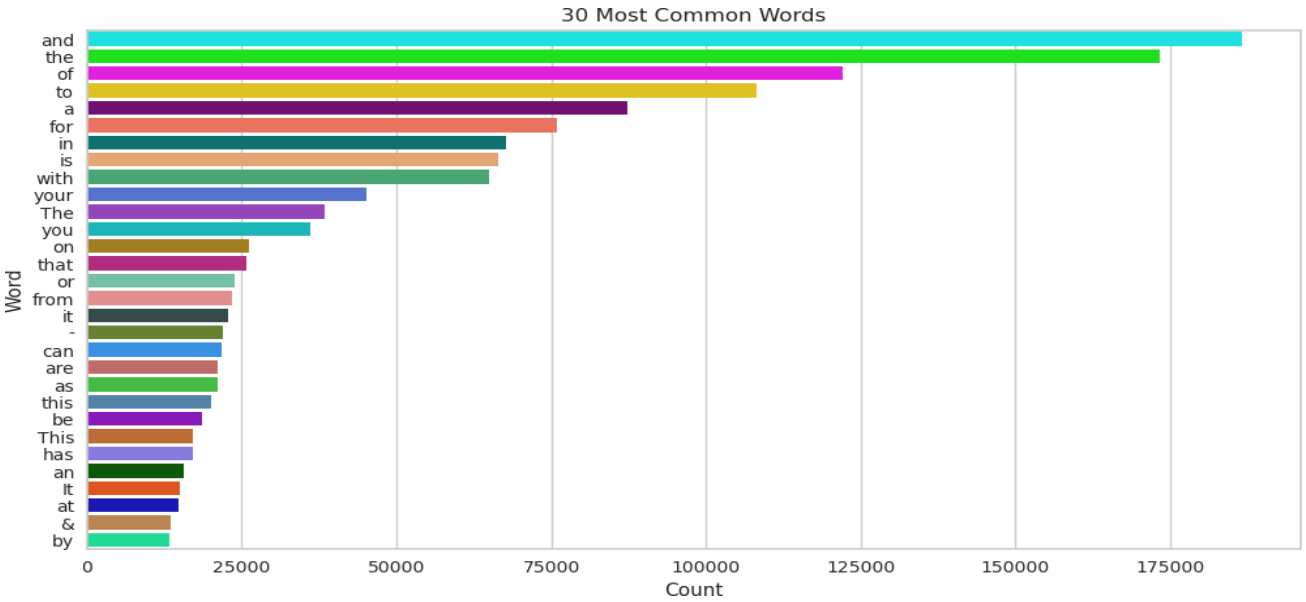

Fig. 5. Common 30 words of Text detection

C. Proposed SVM-Multinomial NB Model

The SVM-Multinomial Naive Bayes model utilizes a distinctive approach that synergistically integrates the capabilities of Support Vector Machines (SVM) and Multinomial Naive Bayes algorithms. Figure 5 shows the Overall Model architecture of our system.

• Feature Extraction: The model initiates by extracting features from the text data, such as word frequency or occurrence, utilizing the Multinomial Naive Bayes technique. This stage facilitates the creation of a visual or symbolic depiction of the textual information.

• Training Phase: The collected characteristics are then learned and classified using Support Vector Machines (SVM), which are renowned for their efficient classification capability. The goal of SVM is to identify the ideal hyper plane in the feature space for dividing various classes.

• Model Fusion: Through the amalgamation of the ad- vantages of both methods, the model attains heightened precision and resilience in text identification assignments. A more complete and dependable detection system is produced when Multinomial Naive Bayes performs well in feature representation and Support Vector Machines (SVM) thrive in classification.

• Prediction: In the prediction stage, the model makes use of the patterns it has learned to reliably identify and classify text in an e-commerce setting. With respect to the traits it has acquired during training, it is capable of distinguishing between different text classes or categories with effectiveness

All things considered, this hybrid method combines the best features of SVM and Multinomial Naive Bayes to produce a strong and accurate text identification system designed for e- commerce applications.

IV. RESULT ANALYSIS

A. Evaluation Criteria

This research examines the effects of the combined SVM- Multinomial NB approach on E-commerce identification. The text is looked over. To evaluate the efficacy of an algorithm or strategy, transferable and comparative performance metrics are needed. The process of choosing and using training and testing sets is a significant obstacle for evaluating a strategy's efficacy since it might lead to inconsistent model performance. True-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) values make up the confusion matrix, which serves as the foundation for building most performance indicators. Depending on how the performance assessment is carried out, the importance of these factors may change.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2024

DOI: 10.17148/IJARCCE.2024.131005

The number of accurate estimates that an identification system can predict from all of its estimations is what determines how accurate it is. The precision is evaluated by means of,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

To evaluate the accuracy of our system's performance on the evaluation dataset, we also calculate precision and recall values.

Precision and recall are important metrics for evaluating binary classification models, which generally include two classifications: "positive" and "negative". These metrics reveal a model's prediction accuracy and event detection. Precision is measured by dividing the model's positive predictions by the number of correctly predicted positive instances. It measures how effectively the model predicts success.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall, also known as sensitivity or true positive rate, is calculated by dividing true positive predictions by genuine positive cases in the dataset. The model's capacity to discover and record all positive findings is measured by this statistic.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

B. Results

We calculated Accuracy, Validation Accuracy, Loss, and Validation Loss during model training. We use these measures to highlight our model's performance and learning in Figure 6. This visual tool shows how effectively the model learns from training data and generalizes to unseen or validation data for E-commerce text recognition.

TABLE I
PRECISION,RECALL AND F1 SCORE SHOWING

Precision	Recall	F1 Score
0.99	0.92	0.98
0.99	0.99	0.99
0.96	0.98	0.97
0.98	0.98	0.98

We've produced separate ROC (Receiver Operating Characteristic) curves for the Multinomial Naive Bayes (NB) and Support Vector Machine (SVM) models. Figure 7 shows the ROC Curves of Four Parameter.

Figure 4 presents the confusion matrix for the five emotion categories. The efficacy of a classification model is summarized by comparing the predicted and actual class labels in a tabular format. In Figure 8 we show the confusion matrix of our system.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2024

DOI: 10.17148/IJARCCE.2024.131005

C. Implemented Model Result

We have implemented various novel techniques in our quest to improve E-Commerce Text recognition. SVM, Random Forest, KNN, LSTM, GRU, Multinomial NB, and SVM- Multinomial NB models were developed and applied as part of our methodology. The most promising of them, based on our evaluation's most convincing outcomes, was the SVM-Multinomial NB model.

TABLE II
DIFFERENT MODEL'S ACCURACY AND VALID ACCURACY WHAT WE IMPLEMENT FOR OUR SYSTEM

Model/Classifier	**Accuracy**	**Valid Accuracy**
SVM	95.72%	95.96%
Random Forest	91.94%	88.32%
KNN	72.84%	69.03%
LSTM	98.62%	89.11%
GRU	99.18%	86.79%
Multinomial NB	97.56%	94.54%
SVM-Multinomial NB(Proposed)	**99.83%**	**98.35%**

D. Comparison of Existing and Proposed Model

In our comparative review of recent developments in text detection and classification, we have compared several new methods with our model. Based on this analysis, our model has shown good performance, outperforming the alternatives. The detailed presentation of the models and their results demonstrates the effectiveness and superiority of our method in this field. In Table 3 we show the Comparison.

TABLE III
COMPARING PERFORMANCE OF PROPOSED AND EXISTING APPROACHES

Ref	**Model**	**Accuracy**
[15]	LSTMRNN	93.17%
[16]	Generative	77%
[17]	RF	98.1%
[18]	XGBoost	96%
[19]	BERT	96.49%
(Proposed)	**SVM-MNB**	**99.83%**

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2024

DOI: 10.17148/IJARCCE.2024.131005

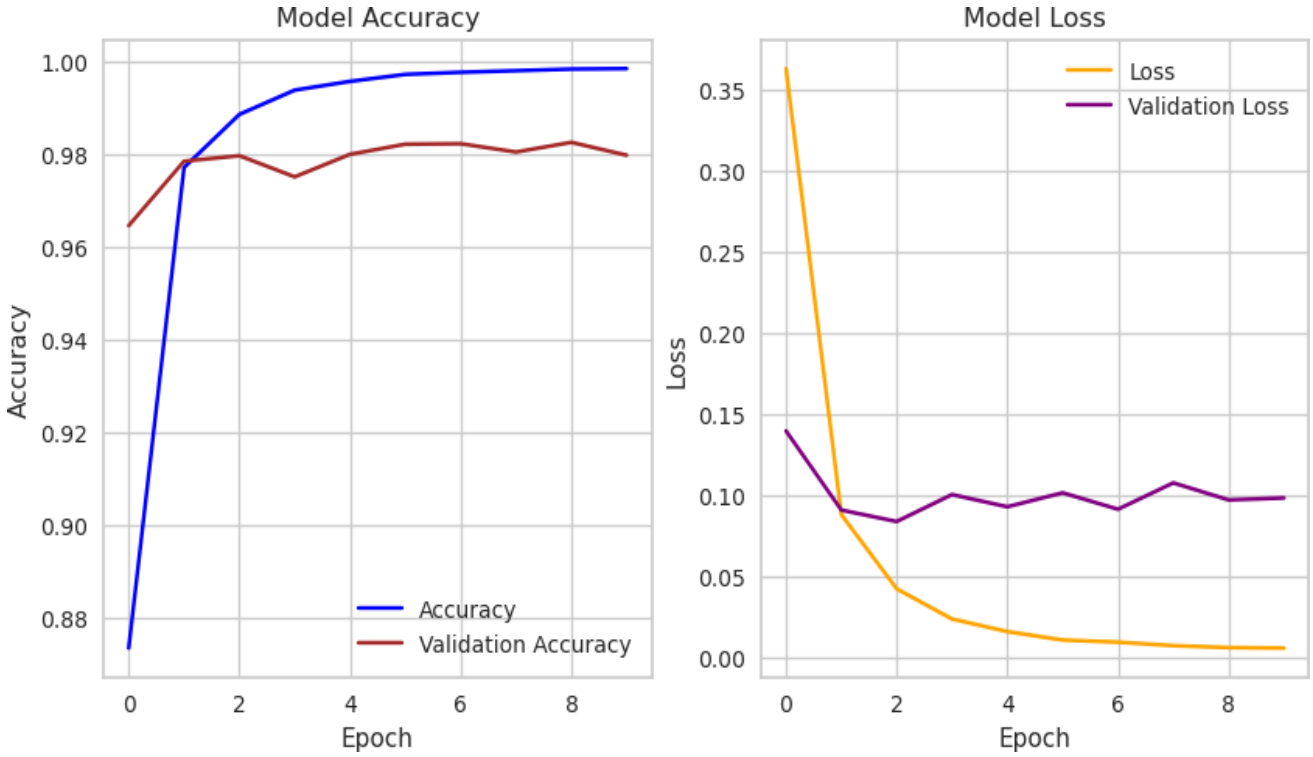

Fig. 6. Accuracy Of Proposed Model

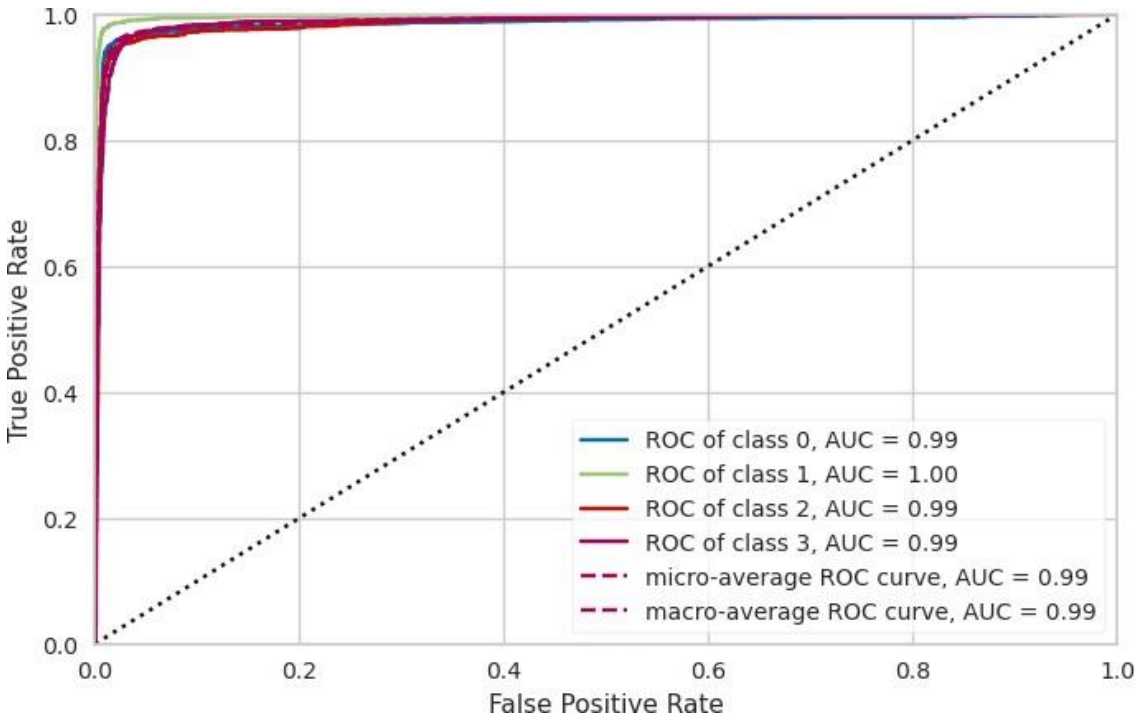

Fig. 7. ROC Curves of four Class

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2024

DOI: 10.17148/IJARCCE.2024.131005

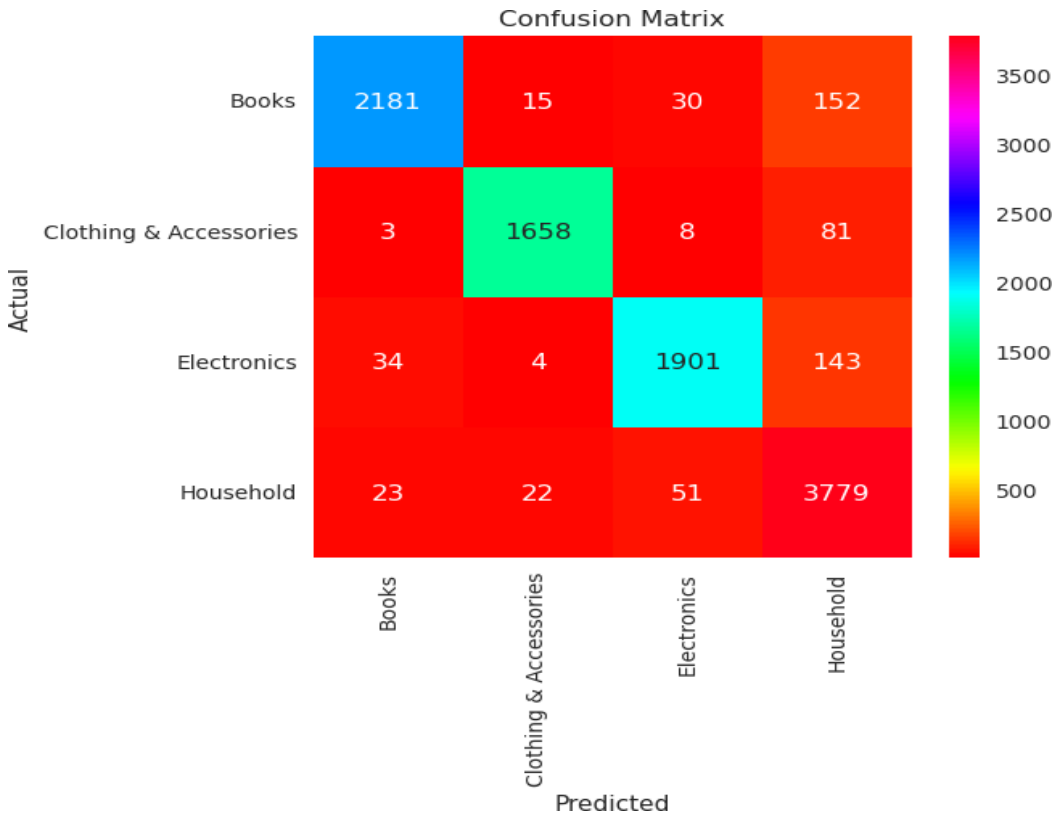

Fig. 8. Confusion Matrix of Four Class

V. CONCLUTION AND FUTURE WORK

By using a combination of behavioural and linguistic characteristics and an iterative multi-level algorithm, we have developed a new hybrid model for identifying E-commerce text. Our method emphasizes the importance of a complete feature set for accurate identification and takes into account the relationships among entities. We have incorporated structured frameworks into our algorithm to effectively handle multivariate features. Our research emphasizes the need for a comprehensive strategy that considers all behavioural and linguistic characteristics for optimal results. This work provides a strong foundation for further research to improve E-commerce text identification algorithms. By highlighting the importance of thorough feature assessment and sophisticated algorithms, we can develop more robust models that can navigate the ever- changing terrain of online text entities.

REFERENCES

[1]. Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., ... Li, X. (2023).Differentiate chatgpt-generated and human-written medical texts. arXiv preprint arXiv:2304.11567.

[2]. Xu, S. (2018). Bayesian Na¨ıve Bayes classifiers to text classification. Journal of Information Science, 44(1), 48-59.

[3]. Quinlan, J. R. (1996). Learning decision tree classifiers. ACM Computing Surveys (CSUR), 28(1), 71-72.

[4]. Suthaharan, S., Suthaharan, S. (2016). Support vector machine. Machine learning models and algorithms for big data classification: thinking with examples for effective learning, 207-235.

[5]. Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

[6]. Parmar, A., Katariya, R., Patel, V. (2019). A review on random forest: An ensemble classifier. In International conference on intelligent data communication technologies and internet of things (ICICI) 2018 (pp. 758-763). Springer International Publishing.

[7]. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2024

DOI: 10.17148/IJARCCE.2024.131005

[8]. Hossain, Muhammad Minoar, Reshma Ahmed Swarna, Rafid Mostafiz, Pabon Shaha, Lubna Yasmin Pinky, Mohammad Motiur Rahman, Wahidur Rahman, Md Selim Hossain, Md Elias Hossain, and Md Sadiq Iqbal. ”Analysis of the performance of feature optimization techniques for the diagnosis of machine learning-based chronic kidney disease.” Machine Learning with Applications 9 (2022): 100330.

[9]. Beresneva, D. (2016). Computer-generated text detection using machine learning: A systematic review. In Natural Language Processing and Information Systems: 21st International Conference on Applications of Natural Language to Information Systems, NLDB 2016, Salford, UK, June 22-24, 2016, Proceedings 21 (pp. 421-426). Springer International Publishing.

[10]. Jawahar, G., Abdul-Mageed, M., Lakshmanan, L. V. (2020). Automatic detection of machine generated text: A critical survey. arXiv preprint arXiv:2011.01314.

[11]. Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., Huang, F. (2023). On the possibilities of ai-generated text detection. arXiv preprint arXiv:2304.04736.

[12]. Floridi, L., Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694.

[13]. Chang, S. Y. (2022, July). Towards detection of AI-generated texts and misinformation. In Socio-Technical Aspects in Security: 11th International Workshop, STAST 2021, Virtual Event (p. 194). Springer Nature.

[14]. Crothers, E., Japkowicz, N., Viktor, H. L. (2023). Machine-generated text: A comprehensive survey of threat models and detection methods. IEEE Access

[15]. Katib, I., Assiri, F. Y., Abdushkour, H. A., Hamed, D., Ragab, M. (2023). Differentiating Chat Generative Pretrained Transformer from Humans: Detecting ChatGPT-Generated Text and Human Text Using Machine Learning. Mathematics, 11(15), 3400.

[16]. Islam, N., Sutradhar, D., Noor, H., Raya, J. T., Maisha, M. T., Farid, D.M. (2023). Distinguishing Human Generated Text From ChatGPT Generated Text Using Machine Learning. arXiv preprint arXiv:2306.01761.

[17]. Zaitsu, W., Jin, M. (2023). Distinguishing ChatGPT (-3.5,-4)-generated and human-written papers through Japanese stylometric analysis. arXiv preprint arXiv:2304.05534

[18]. Shijaku, R., Canhasi, E. (2023). ChatGPT Generated Text Detection. Publisher: Unpublished.

Mujahid, M., Rustam, F., Shafique, R., Chunduri, V., Villar, M. G., Ballester, J. B., ... Ashraf, I. (2023). Analyzing Sentiments Regarding ChatGPT Using Novel BERT: A Machine Learning Approach. Information, 14(9), 474.

