A Lightweight CCA-Based Framework for Complementary Product Recommendation Using Textual Embeddings

G. Kalyan Chakravarthi
Department of Computer Science and Engineering, GIET University, India
gkalyan.chakravarthi@giet.edu

Raghvendra Kumar
Department of Computer Science and Engineering, GIET University, India
raghvendra@giet.edu (corresponding author)

Ssvr Kumar Addagarla
Department of Computer Science and Engineering, Nadimpalli Satyanarayana Raju Institute of Technology, Visakhapatnam, Andhra Pradesh, India
ssvrkumar@gmail.com

Received: 7 February 2026 | Revised: 18 March 2026 and 23 March 2026 | Accepted: 26 March 2026

 DOI: https://doi.org/10.48084/etasr.18014

ABSTRACT

The rapid growth of online markets has increased the need for recommendation systems that identify not only similar products but also complementary ones. Conventional recommender systems rely primarily on similarity measures, co-purchase frequency, and past user history, but they fail to identify semantic and functional associations that retrieve true complementary products. As a result, customers do not receive meaningful complements for their products and instead are presented with substitutive item suggestions. The current study proposes an automated complementary recommendation framework based on a lightweight Canonical Correlation Analysis (CCA)-Lite approach that identifies semantic relationships between product categories using only textual metadata. This method uses Term Frequency–Inverse Document Frequency (TF-IDF) for feature extraction from product descriptions, followed by category-aware candidate generation. The CCA-Lite correlation mapping identifies relationships between primary and complementary product embeddings. The proposed framework models semantic relationships across product categories using the extracted textual information. Its lightweight and scalable design helps maintain conceptual simplicity while adapting to diverse e-commerce contexts. The presented text-based approach improves contextual relevance and supports cross-category recommendation through purely text-driven learning, offering an efficient, domain-independent, and computationally lightweight solution for modern e-commerce platforms.

Keywords-CCA-Lite; complementary product; textual metadata; category-aware; semantic relationships

I. INTRODUCTION

With the rapid increase of global e-commerce, the way customers identify, evaluate, and purchase products has been radically transformed. In 2004, the worldwide e-commerce market exceeded US $6.3 trillion according to Statista, accounting for over 22% of total retail sales, and is anticipated to reach US $8 trillion by 2027 [1]. In the evolving world of digital ecosystems, recommendation technologies lay the foundation for driving customers towards relevant products and thereby improving the conversion rate. McKinsey & Company discusses how personalization and recommendation-driven systems contribute to growth [2].

Conventional recommendation approaches like Content-Based Filtering (CBF) and Collaborative Filtering (CF) assume user preferences from past interactions or item attributes and are effective in retrieving similar products. In spite of that, they generally struggle to model complementary relationships, i.e., items that functionally or contextually enhance the primary purchase instead of acting as substitutes (e.g., a laptop with a backpack, cooling pad, or wireless mouse). Overlooking contextual complementarity limits user satisfaction and cross-selling potential [3, 4].

Advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) have enabled semantic representations of product text (titles, reviews, specifications).

Recent transformer-based text models and classical Term Frequency–Inverse Document Frequency (TF-IDF) have improved sentiment-aware personalization and similarity computation by capturing contextual and linguistic patterns in product metadata [5, 6]. However, complementary product recommendation remains underexplored compared to similarity-based retrieval systems. Earlier research highlights many multi-criteria and deep recommendation methods that emphasize similar item-item recommendation rather than cross-product associations [7, 8].

Current recommendation systems primarily optimize for substitutive similarity within the same category, relying heavily on dense user–item interactions that often face limitations such as data sparsity and cold-start problems. Although deep neural recommenders have improved ranking accuracy, they introduce significant computational overhead and limited interpretability, which hinders scalable, real-time, and explainable deployment [8, 9].

In parallel, textual metadata are utilized to a limited extent to discover cross-category and functionally complementary associations. Recent literature reviews identify several gaps, including semantic limitations, where existing methods focus on intra-category similarity over inter-category complementarity [3, 4]; an interpretability gap, where accurate but non-transparent deep architectures hinder trust and diagnosis [8, 9]; and a scalability gap, where high training and inference costs make such models impractical for large catalogs [8].

These limitations highlight the need for lightweight, text-centric, and interpretable recommendation approaches that rely on semantic cues rather than extensive user behavior. Enhancing complementary product identification supports cross-selling, basket building, and user satisfaction while advancing the development of context-aware and explainable recommender systems [3, 6].

The present study addresses these challenges by proposing a text-based complementary product recommendation framework that leverages product semantics and inter-category relationships. The proposed approach integrates TF-IDF and cosine similarity with a Canonical Correlation Analysis (CCA)-Lite mapping technique to capture latent correlations between primary and complementary product categories that traditional methods fail to identify. The framework is lightweight, scalable, and domain-independent, achieving a balance between predictive effectiveness and computational efficiency.

II. LITERATURE REVIEW

Several studies have investigated text-based and semantic approaches for recommendation systems in e-commerce environments.

Authors in [4] proposed a novel recommendation system called DeepIDRS, which operates in two stages. First, a language model is used to understand textual information, followed by an attention-based model that generates recommendations based on user history. The system is evaluated on real-world data and achieves an improvement of approximately 10% over existing models.

Authors in [5] presented a sentiment analysis recommendation system, which uses machine learning and NLP techniques to process customer reviews before performing sentiment analysis for rating validation. User-specific product ratings are derived from sentiment scores. To improve recommendation quality, cosine similarity is used to identify related products, including newly introduced items. Fuzzy logic is further applied to classify recommendation levels. Experimental results demonstrate the effectiveness of the system in modern e-commerce environments.

Authors in [6] presented a deep learning-based architecture for analyzing customer reviews and supporting business decision-making through sentiment understanding. The method preprocesses text and extracts features using a hybrid approach combining general and specific review attributes. Long Short-Term Memory (LSTM) networks are then used for sentiment classification. Evaluated on three datasets, the model achieves an F1-score of 92.81%, recall of 91.63%, and precision of 94.46%, demonstrating strong performance in sentiment summarization and analysis.

Authors in [7] presented an aspect-based deep learning model, AEMC, which analyzes multiple item attributes rather than relying on a single overall rating to improve product recommendations. The model is evaluated on two real-world datasets: the Yahoo Movies dataset, containing over 34,000 ratings across four aspects (story, acting, direction, and visuals), and a TripAdvisor dataset containing 29,000 hotel reviews across seven aspects, including location, service, and quality. The model is compared with several existing recommendation approaches and demonstrates improved performance for both datasets.

Authors in [8] proposed a recommendation system incorporating a mutual attention network that focuses on customer reviews to better understand user needs and improve recommendation quality. A prediction layer integrates all relevant information to generate accurate recommendations. Experiments conducted on Amazon and Yelp datasets show that the proposed method outperforms traditional approaches in both item ranking and rating prediction.

Authors in [10] proposed a smart e-commerce product recommendation system, T1OWA, which analyzes customer reviews using fuzzy logic and sentiment analysis. The system considers user reviews along with user interpretation of those reviews. It is based on an aggregation method that merges review content with user preferences and behavior to generate personalized product rankings. The system is evaluated on two datasets: one involving five SUVs using user-defined feature importance, and the other, a real-world dataset consisting of 10,000 hotel reviews collected from TripAdvisor, covering eight hotels and six aspects. The results show that the proposed method not only matches the accuracy of existing models but also provides greater flexibility by adapting to different user attitudes and review styles.

Authors in [11] introduced a smart product recommendation system that incorporates not only past user behavior but also the type of user interacting with the system. The approach uses NLP techniques to preprocess Amazon review data, identify positive and negative reviews, and compute product scores based on review sentiment and user type. Similar products are also identified using similarity measures. The system improves decision-making and provides better recommendations compared to traditional methods.

Authors in [12] discussed two major categories of recommendation systems, context-aware and context-based systems. Context-aware systems recommend products based on situational factors such as location, time, and surrounding users, whereas context-based systems recommend items based on past user activity. These approaches are applied in domains such as e-learning, mobile media recommendations, and personalized news delivery using hybrid systems, collaborative filtering, and content-based filtering. NLP techniques are also used to generate intuitive and relevant recommendations. Experimental analysis using semantic reasoning shows that such systems perform comparably to content-based recommenders.

Authors in [13] introduced dpFPN-Netv2 for product image detection and recognition. The model enhances Feature Pyramid Networks (FPN) by integrating two modules, DPFM and RFM, to improve detection accuracy. The proposed system reduces detection time from 90 ms to 52 ms compared with RetinaNet-50. Additionally, GTNet is incorporated for product classification, outperforming MWI-DenseNet in terms of accuracy while reducing computational complexity.

Authors in [14] proposed EPR-ML, a machine learning-based recommendation system that combines automatic tagging with machine learning and NLP techniques. Sentiment analysis is used to extract key product features, and logistic regression is applied for classification. The system is evaluated using Gaussian Naïve Bayes (GNB) and Linear Support Vector Machine (LSVM), where LSVM achieves the highest accuracy of 96%, demonstrating improved product tagging and recommendation performance.

Authors in [15] proposed a hybrid recommendation approach combining NLP and Convolutional Neural Networks (CNNs) for apparel recommendation. The model processes product titles using NLP and extracts image features using a VGG-16 CNN model on an Amazon dataset containing 180,000 clothing items. Feature vectors are compared using Euclidean distance to identify similar products, improving recommendation accuracy.

Authors in [16] introduced a personalized recommendation algorithm based on Frequent Pattern Growth (FP-Growth) and association rule mining. The method improves purchasing prediction by identifying frequent itemsets, achieving better probability estimates for next-item purchase recommendations.

Authors in [17] proposed PALR, a framework that leverages Large Language Models (LLMs) for personalized recommendation. The system first retrieves candidate items based on user–item interactions and then uses an LLM to rank items using natural language instructions. This improves ranking quality and personalization effectiveness.

Authors in [18] developed a sentiment-based recommendation framework using an e-commerce dataset of women's clothing reviews. The study evaluates multiple machine learning models to identify the most effective approach for sentiment analysis and product recommendation, highlighting the importance of customer behavior modeling.

Authors in [19] introduced Recformer, a transformer-based framework for sequential recommendation that integrates user preferences and item features. The model uses a bidirectional transformer architecture with pretraining and fine-tuning strategies that combine language understanding with recommendation objectives. Experimental results on six datasets show strong performance, particularly in low-resource and cold-start scenarios.

Authors in [20] presented a large-scale personalized recommendation system developed by the H&M Group. The system integrates popularity-based ranking, collaborative filtering, and personalized Bayesian ranking within a hybrid retrieval strategy to address scalability and cold-start challenges. Evaluation using LightGBM and deep neural networks shows that LightGBM outperforms Deep Neural Network (DNN) models in MAP@K and MAR@K metrics.

Authors in [21] compared traditional e-commerce classification systems with modern personalized recommendation frameworks. The proposed system integrates BERT-based embeddings with a nearest-neighbor algorithm to generate structured recommendations, demonstrating improved interpretability and performance.

Authors in [22] introduced LLM-based E-commerce Machine Translation (LEMT), a framework designed for domain-specific translation tasks in e-commerce. It incorporates bilingual resources, tokenizer optimization, and a two-stage self-contrastive fine-tuning strategy. Experimental results show that LEMT outperforms Neural Machine Translation models and GPT-4 in both translation quality and robustness.

Authors in [23] investigated the application of NLP and deep learning techniques for analyzing patient comments in sentiment-based healthcare recommendation systems. The study integrates linguistic, semantic, and statistical modeling approaches to enable emotion-aware drug recommendation systems that improve prescription accuracy.

Authors in [24] explored token-based language modeling for product recommendation using NLP techniques. Instead of full product names, tokenized representations (unigrams, bigrams, and trigrams) are used to better capture product relationships. Evaluations on Instacart and UK e-commerce datasets show improved performance in Mean Reciprocal Rank (MRR) and Hit Rate compared to non-tokenized approaches.

III. METHODOLOGY

This research is focused on developing a model that focuses on recommending complementary products for different electronic products based on textual queries using deep learning models and similarity search techniques. In this method, we developed a model that initially processes text descriptions and then performs feature extraction using TF-IDF. Subsequently, candidate generation is performed using cosine similarity, followed by complementary filtering through a CCA-Lite mapping procedure to obtain product recommendations. The proposed method adopts an efficient e-commerce recommendation algorithm that produces embeddings, preprocesses data in an orderly fashion, and provides text-based complementary suggestions for a given input query, as shown in Figure 1.

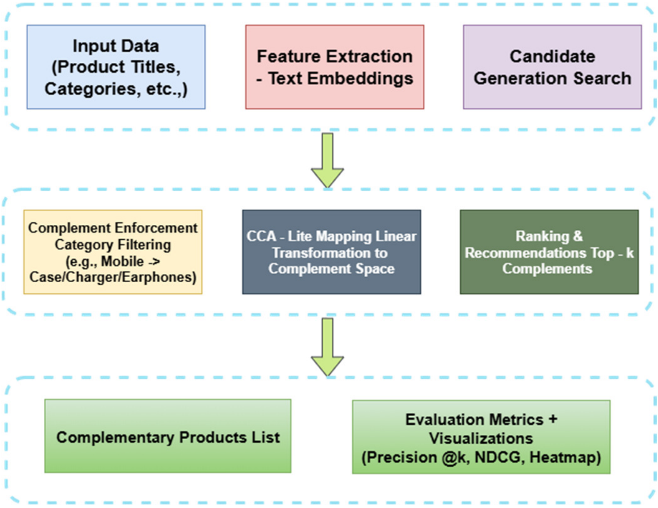

Fig. 1. Framework of the proposed complementary product recommender system.

A. *Dataset Preparation*

This study utilizes a dataset derived from the publicly available Amazon product dataset, specifically the Electronics category [25]. The dataset consists of product titles, descriptions, and category metadata. From this dataset, a subset of 9,500 electronic products across eight categories was curated and preprocessed to construct complementary product pairs. Each primary product is associated with 3–4 complementary products based on product relationships and category-level associations, as shown in Figure 2.

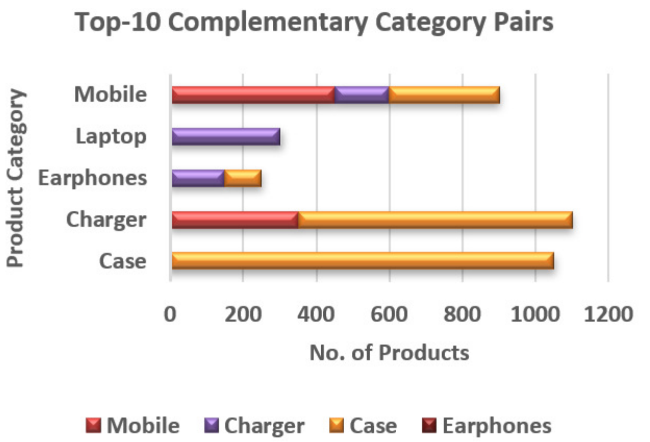

Fig. 2. Cross-category product pairs.

Each product pair is mathematically represented in (1). The dataset is further processed to remove noise and ensure consistency in textual metadata for effective feature extraction.

$$P = \{(p_i, c_j) \mid p_i \in \text{primary}, c_j \in \text{complement set of } p_i\} \quad (1)$$

where p_i represents the text feature representation of the primary product and c_j represents the feature representation of its complementary product. Mapping of (p_i, c_j) forms the training instances for the CCA-Lite mapping stage of the model.

B. *Feature Extraction through Term Frequency–Inverse Document Frequency*

The textual information of each product (i.e., product title, description, etc.) provides meaningful information about the type, function, and specifications of the product. These textual data are unstructured and cannot be directly interpreted by machine learning models. Therefore, a text embedding process is required to convert the text into numerical representations that capture the importance of words in each product description. In this study, TF-IDF is applied to product titles and descriptions to extract meaningful textual features representing the semantic content of each product.

The $\text{TF} - \text{IDF}$ value for term t in a product description d is computed as presented in (2):

$$\text{TF} - \text{IDF}\ (t, d) = \text{TF}(t, d)\ \times \text{IDF}\ (t) \quad (2)$$

where $\text{TF}(t, d)$ represents the term frequency and is computed as shown in (3):

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}} \quad (3)$$

Here, $f_{t,d}$ denotes the frequency of term t in text query d, normalized by the total number of terms in that text query.

The inverse document frequency $\text{IDF}(t)$ is computed as described in (4):

$$\text{IDF}(t) = \log\left(\frac{N}{n_t}\right) \quad (4)$$

where N denotes the total number of text queries, and n_t denotes the number of text queries containing the term t. The product of (3) and (4) assigns higher weight to terms that appear frequently in a specific text query, emphasizing words that uniquely identify the product.

C. *Complementary Category Rules*

In this stage, rules are defined based on traditional domain knowledge to establish logical associations between product categories. The rules are applied to filter and refine recommendations. For example, the rules are defined as:

$$\text{Laptop} \rightarrow \{\text{Charger}, \text{Mouse}, \text{Bag}\},$$

$$\text{Mobile Phone} \rightarrow \{\text{Earphones}, \text{Cover}, \text{Screen Protector}\}.$$

Formally, if C_p represents the category of a primary product and C_c represents the category of a complementary candidate, the rule-based filter is expressed as shown in (5):

$$R(C_p, C_c) = \begin{cases} 1, & \text{if } C_c \in Complement(C_p) \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Only the candidates satisfying $R(C_p, C_c) = 1$ are retained for the next stage.

D. Candidate Generation Using K-Nearest Neighbors

In this stage, the search space is narrowed down by retrieving products that exhibit high feature similarity with the query product. This allows the model to focus on a smaller and more relevant set of items before applying complement category rules and CCA-Lite mapping. This step reduces computational overhead and improves the precision of the subsequent mapping process.

Each product is represented as a feature vector extracted from the text query. Let f_i and f_j represent two feature vectors corresponding to products i and j, respectively. The similarity between these vectors is computed using cosine similarity to identify semantically related products in the embedding space.

The cosine similarity is computed as shown in (6):

$$Sim(f_i, f_j) = \frac{f_i \cdot f_j}{\|f_i\| \|f_j\|} \quad (6)$$

where $f_i \cdot f_j$ represents the dot product between the two feature vectors, and $\|f_i\|$ and $\|f_j\|$ denote their Euclidean norms. The cosine similarity values range between 0 and 1, where a value of 1 indicates identical direction (maximum similarity), and a value of 0 indicates no similarity.

After computing similarity scores, the values are sorted in descending order, and the top-k items with the highest similarity scores are selected as candidate complementary products. The corresponding product IDs are passed to the next stage for further processing.

In this study, K-Nearest Neighbors (KNN) is used to retrieve the top-k semantically similar products based on textual feature embeddings. The value of k is set to 10. The KNN function is defined as:

$$\text{KNN}(f_q, D, k) = \{f_1, f_2, \ldots, f_k\} \quad (7)$$

where D represents the dataset of all feature vectors, f_q is the query vector, f_i represents the i-th most similar item to the query, and k defines the number of neighbors controlling the candidate selection size.

This stage enables the identification of candidate products that are semantically similar to the query and are subsequently processed by the complement category rules and CCA-Lite mapping stages.

E. Canonical Correlation Analysis–Lite Mapping

The CCA-Lite mapping method is used for complementary association. It is a simplified yet effective approach designed to establish the final complementary mappings between primary and potential complementary products. The TF-IDF vectors obtained in the previous step undergo normalization and are reduced using Principal Component Analysis (PCA) to optimize computational complexity and reduce redundancy.

CCA-Lite mapping aims to identify how effectively two products complement each other and refines the results obtained from the candidate generation phase using a linear correlation function. The correlation function is calculated as shown in (8):

$$\rho = \max_{w_p, w_c} \frac{w_p^T \Sigma_{pc} w_c}{\sqrt{(w_p^T \Sigma_{pp} w_p)(w_c^T \Sigma_{cc} w_c)}} \quad (8)$$

where w_p and w_c are the projection vectors for the primary and complementary spaces, respectively. The Lite version of this method avoids full matrix inversion by employing truncated eigenvalue decomposition, thereby improving computational efficiency for large-scale datasets.

After learning the optimal correlation weights, the primary and complementary embeddings are projected into a common correlated subspace, where related products and their true complements are positioned close together, whereas unrelated products are positioned farther apart, as shown in (9) and (10):

$$P_i' = w_p^T P_i \quad (9)$$

$$C_j' = w_c^T C_j \quad (10)$$

In the extraction phase, a query product is provided, and its embedding is mapped using the learned w_p projection. It is then compared with the complementary embeddings C_j'. Finally, the top-k correlated products are selected based on cosine similarity within the correlated subspace to generate the final complementary product recommendations.

IV. RESULTS AND DISCUSSION

In this section, we present the experimental findings attained using the proposed model. All baseline models were implemented and evaluated on the same dataset to ensure a fair comparison. The experiments were performed using the electronics subset of the Amazon product dataset. The dataset was first preprocessed to retain product titles, categories, and relationships pertaining to complementary recommendation tasks. Text embeddings were extracted by applying the TF-IDF vectorizer to construct the complementary dataset for recommendation.

Table I presents the comparative performance of various models in terms of Precision@3, Recall@5, and Normalized Discounted Cumulative Gain (NDCG@3), whereas Table II presents the efficiency evaluation using cross-category rate, MRR, coverage, and inference time. The proposed Complementary CCA-Lite mapping is compared with baseline methods, including TF-IDF with KNN, Logistic Re-ranker, and Graph Personalized PageRank (PPR). The corresponding results are also illustrated in Figures 3 and 4.

TABLE I. PERFORMANCE EVALUATION OF EXISTING AND PROPOSED MODELS

Method	Precision@3	Recall@5	NDCG@3
TF-IDF + KNN [26]	0.65	0.58	0.61
Logistic Re-ranker [27]	0.72	0.66	0.69
Graph PPR [28]	0.75	0.69	0.73
Complementary CCA-Lite	0.82	0.77	0.79

TABLE II. EFFICIENCY EVALUATION OF EXISTING AND PROPOSED MODELS

Method	**Cross-category rate**	**MRR**	**Coverage (%)**	**Inference time (ms/query)**
TF-IDF + KNN	0.72	0.55	78.5	1.2
Logistic Re-ranker	0.77	0.61	81.3	2.1
Graph PPR	0.80	0.64	84.7	5.6
Complementary CCA-Lite	0.86	0.71	89.2	2.9

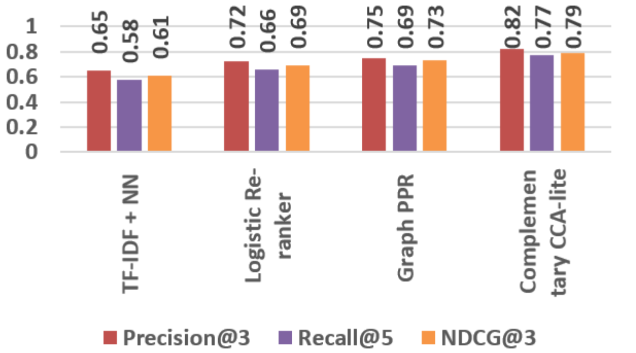

Fig. 3. Performance evaluation of CCA-Lite compared with baseline models.

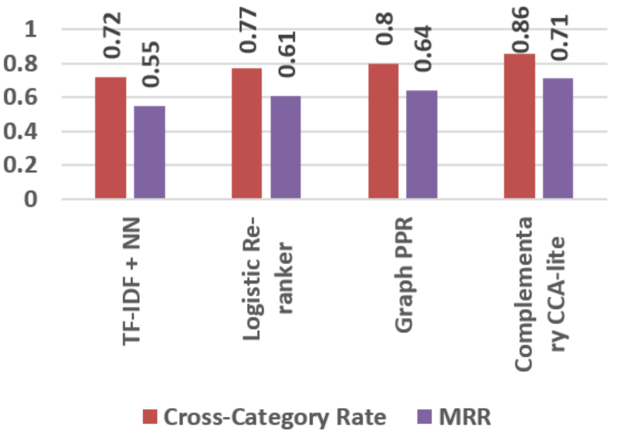

Fig. 4. Efficiency evaluation of CCA-Lite compared with baseline models.

The proposed Complementary CCA-Lite model outperforms all baseline models across the evaluation metrics presented in Tables I and II. Traditional TF-IDF and KNN-based methods primarily rely on vector similarity within a single latent space, thereby favoring substitutable products with high lexical similarity. However, complementary products are semantically linked through contextual and functional dependencies rather than lexical similarity.

In terms of precision, recall, and NDCG, the Logistic Re-ranker and Graph PPR approaches achieve precision values of 0.72 and 0.75 and NDCG values of 0.69 and 0.73, respectively. These values are lower than those of the proposed Complementary CCA-Lite approach, which achieves a precision of 0.82, recall of 0.77, and NDCG of 0.79, demonstrating improved performance.

Similarly, in the efficiency analysis, the Graph PPR approach achieves a cross-category rate of 0.80, MRR of 0.64, and coverage of 84.7%, whereas the Logistic Re-ranker achieves a coverage of 81.3%. In comparison, the Complementary CCA-Lite model achieves a cross-category rate of 0.86, MRR of 0.71, and coverage of 89.2%, demonstrating superior performance.

An ablation study is conducted to evaluate the contribution of each component of the proposed CCA-Lite framework, as shown in Table III. The complete model is compared with two variants: one without the mapping functionality and another without the complement filter. The results indicate that the full CCA-Lite framework achieves superior cross-domain alignment. The comparative performance in terms of Precision@3 and NDCG@3 is further illustrated in Figure 5.

TABLE III. ABLATION ANALYSIS

Variant	**Description**	**Precision@3**	**NDCG@3**
CCA-Lite (full)	Text + complement mapping (linear)	0.82	0.79
w/o mapping	Uses raw TF-IDF (no projection)	0.68	0.63
w/o complement filter	Includes all neighbors (substitutes + complements)	0.60	0.57

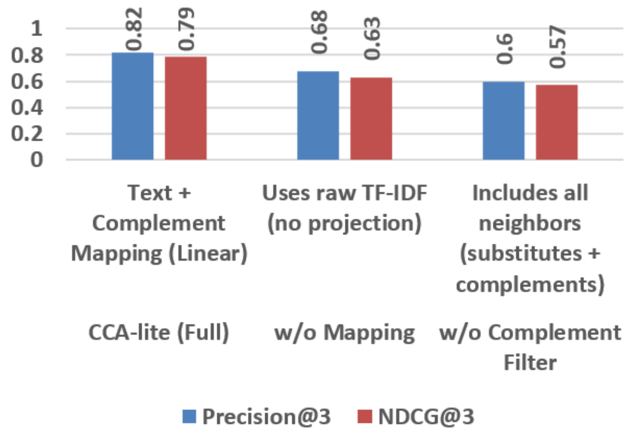

Fig. 5. Ablation analysis of CCA-Lite mapping using Precision@3 and NDCG@3.

To interpret the recommendation quality of the CCA-Lite model, Table IV presents a few top-3 recommendations generated by the proposed method.

TABLE IV. SAMPLE RECOMMENDATIONS FROM THE CCA-LITE MODEL

Query product	**Top-3 complementary products (predicted)**
Samsung Galaxy M33 5G	Spigen Rugged Armor Case, Samsung 25W Fast Charger, boAt Airdopes 131 Earbuds
HP Pavilion Laptop 15	HP 65W USB-C Charger, Logitech Wireless Mouse M235, HP Laptop Backpack
Canon EOS 1500D DSLR	SanDisk 128GB SD Card, AmazonBasics 60-inch Tripod, Canon LP-E10 Battery Pack

As presented in Table IV, the proposed Complementary CCA-Lite model provides accurate complementary recommendations. Each query product and its corresponding top-3 predicted complementary products represent functionally related accessories.

Overall, the proposed CCA-Lite model provides more effective complementary product recommendations compared with traditional approaches.

V. CONCLUSION

The proposed study presents a text-based complementary product recommendation framework intended to identify semantically and functionally related items within an e-commerce product dataset. In this approach, the use of Term Frequency–Inverse Document Frequency (TF-IDF) feature extraction, category-aware candidate generation, and complementary Canonical Correlation Analysis (CCA)-Lite mapping enables the projection of product embeddings into a complement-aware latent space. This approach employs lightweight linear mapping, which bridges the semantic gap between different product categories and enables the retrieval of relevant complementary products instead of similar or substitutable ones.

Several experiments were carried out on the Amazon Electronics dataset, which demonstrate that the proposed Complementary CCA-Lite model consistently outperforms baseline methods such as TF-IDF with K-Nearest Neighbors (KNN), Logistic Re-ranker, and Graph Personalized PageRank (PPR). The proposed CCA-Lite model achieves a Precision@3 of 0.82 and NDCG@3 of 0.79, whereas a cross-category rate of 0.86 indicates a 26% improvement in precision and a 15% improvement in MRR over baseline models. These results confirm that the proposed CCA-Lite framework is both effective and computationally efficient for complementary product recommendation tasks.

Furthermore, future work will explore multimodal approaches, where both textual and visual features are integrated to form fine-grained product clusters at category and subcategory levels. This is expected to enhance semantic richness, industrial adaptability, and real-time recommendation capability for complementary product retrieval.

DECLARATION OF COMPETING INTERESTS

The authors declare no competing interests.

ACKNOWLEDGMENT

Not applicable to this work.

DATA AVAILABILITY

The data that support the findings of this study are publicly available from the Amazon product dataset (Electronics category) [25].

REFERENCES

[1] Statista Research Department. "Retail e-commerce sales worldwide from 2022 to 2030." Statista. https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/.

[2] "No customer left behind: How to drive growth by putting personalization at the center of your marketing." McKinsey. https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/no-customer-left-behind.

[3] A. Tholib, T. Widiyaningtyas, and D. D. Prasetya, "An Intelligent Recommendation System Utilizing a Hybrid Deep Learning Method," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 25971–25977, Aug. 2025, https://doi.org/10.48084/etasr.12230.

[4] I. Islek and S. G. Oguducu, "A hierarchical recommendation system for E-commerce using online user reviews," *Electronic Commerce Research and Applications*, vol. 52, Mar. 2022, Art. no. 101131, https://doi.org/10.1016/j.elerap.2022.101131.

[5] H. Zhao, Z. Liu, X. Yao, and Q. Yang, "A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach," *Information Processing & Management*, vol. 58, no. 5, Sept. 2021, Art. no. 102656, https://doi.org/10.1016/j.ipm.2021.102656.

[6] G. Kaur and A. Sharma, "A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis," *Journal of Big Data*, vol. 10, no. 1, Jan. 2023, Art. no. 5, https://doi.org/10.1186/s40537-022-00680-6.

[7] Q. Shambour, "A deep learning based algorithm for multi-criteria recommender systems," *Knowledge-Based Systems*, vol. 211, Jan. 2021, Art. no. 106545, https://doi.org/10.1016/j.knosys.2020.106545.

[8] A. Da'u, N. Salim, and R. Idris, "An adaptive deep learning method for item recommendation system," *Knowledge-Based Systems*, vol. 213, Feb. 2021, Art. no. 106681, https://doi.org/10.1016/j.knosys.2020.106681.

[9] V. Christou *et al.*, "Performance and early drop prediction for higher education students using machine learning," *Expert Systems with Applications*, vol. 225, Sept. 2023, Art. no. 120079, https://doi.org/10.1016/j.eswa.2023.120079.

[10] J. Serrano-Guerrero, J. A. Olivas, and F. P. Romero, "A T1OWA and aspect-based model for customizing recommendations on eCommerce," *Applied Soft Computing*, vol. 97, Dec. 2020, Art. no. 106768, https://doi.org/10.1016/j.asoc.2020.106768.

[11] R. V. Karthik and S. Ganapathy, "A fuzzy recommendation system for predicting the customers interests using sentiment analysis and ontology in e-commerce," *Applied Soft Computing*, vol. 108, Sept. 2021, Art. no. 107396, https://doi.org/10.1016/j.asoc.2021.107396.

[12] U. Javed, K. Shaukat, I. A. Hameed, F. Iqbal, T. M. Alam, and S. Luo, "A Review of Content-Based and Context-Based Recommendation Systems," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 3, pp. 274–306, Feb. 2021, https://doi.org/10.3991/ijet.v16i03.18851.

[13] B. Zhao, W. Li, Q. Guo, and R. Song, "E-Commerce Picture Text Recognition Information System Based on Deep Learning," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, Jan. 2022, Art. no. 9474245, https://doi.org/10.1155/2022/9474245.

[14] V. Malik, R. Mittal, and S. V. SIngh, "EPR-ML: E-Commerce Product Recommendation Using NLP and Machine Learning Algorithm," in *2022 5th International Conference on Contemporary Computing and Informatics*, Uttar Pradesh, India, 2022, pp. 1778–1783, https://doi.org/10.1109/IC3I56241.2022.10073224.

[15] A. Kumar Sharma, B. Bajpai, R. Adhvaryu, S. Dhruvi Pankajkumar, P. Parthkumar Gordhanbhai, and A. Kumar, "An Efficient Approach of Product Recommendation System using NLP Technique," *Materials Today: Proceedings*, vol. 80, pp. 3730–3743, Jan. 2023, https://doi.org/10.1016/j.matpr.2021.07.371.

[16] M. Loukili, F. Messaoudi, and M. E. Ghazi, "Machine learning based recommender system for e-commerce," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 4, pp. 1803–1811, Dec. 2023, https://doi.org/10.11591/ijai.v12.i4.pp1803-1811.

[17] F. Yang, Z. Chen, Z. Jiang, E. Cho, X. Huang, and Y. Lu, "PALR: Personalization Aware LLMs for Recommendation." arXiv, June 07, 2023, https://doi.org/10.48550/arXiv.2305.07622.

[18] M. Loukili, F. Messaoudi, and M. El Ghazi, "Sentiment Analysis of Product Reviews for E- Commerce Recommendation based on Machine Learning," *International Journal of Advances in Soft Computing & Its Applications*, vol. 15, no. 1, pp. 1–13, Mar. 2023, https://doi.org/10.15849/IJASCA.230320.01.

[19] J. Li *et al.*, "Text Is All You Need: Learning Language Representations for Sequential Recommendation," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Long Beach, CA, USA, 2023, pp. 1258–1267, https://doi.org/10.1145/3580305.3599519.

[20] D.-N. Nguyen, V.-H. Nguyen, T. Trinh, T. Ho, and H.-S. Le, "A personalized product recommendation model in e-commerce based on retrieval strategy," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 10, no. 2, June 2024, Art. no. 100303, https://doi.org/10.1016/j.joitmc.2024.100303.

[21] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of E-commerce products based on machine learning," *Applied and Computational Engineering*, vol. 64, pp. 147–153, May 2024, https://doi.org/10.54254/2755-2721/64/20241365.

[22] D. Gao *et al.*, "LLMs-based machine translation for E-commerce," *Expert Systems with Applications*, vol. 258, Dec. 2024, Art. no. 125087, https://doi.org/10.1016/j.eswa.2024.125087.

[23] H. Zheng, K. Xu, H. Zhou, Y. Wang, and G. Su, "Medication Recommendation System Based on Natural Language Processing for Patient Emotion Analysis," *Academic Journal of Science and Technology*, vol. 10, no. 1, pp. 62–68, Mar. 2024, https://doi.org/10.54097/v160aa61.

[24] B. Jeong and K. J. Lee, "NLP-Based Recommendation Approach for Diverse Service Generation," *IEEE Access*, vol. 12, pp. 14260–14274, 2024, https://doi.org/10.1109/ACCESS.2024.3355546.

[25] "Amazon Electronics Products 10k items - 2023." Kaggle. [Online]. Available: https://www.kaggle.com/datasets/akeshkumarhp/electronics-products-amazon-10k-items.

[26] J. Suchal and P. Návrat, "Full Text Search Engine as Scalable k-Nearest Neighbor Recommendation System," in *Artificial Intelligence in Theory and Practice III: Third IFIP TC 12 International Conference on Artificial Intelligence*, Brisbane, Australia, 2010, pp. 165–173, https://doi.org/10.1007/978-3-642-15286-3_16.

[27] C. Pei *et al.*, "Personalized re-ranking for recommendation," in *Proceedings of the 13th ACM Conference on Recommender Systems*, Copenhagen, Denmark, 2019, pp. 3–11, https://doi.org/10.1145/3298689.3347000.

[28] A. J. N. Nzeko'O, M. Tchuenté, and M. Latapy, "A general graph-based framework for top-N recommendation using content, temporal and trust information," *Journal of Interdisciplinary Methodologies and Issues in Sciences*, vol. 5, June 2019, https://doi.org/10.18713/JIMIS-300519-5-2.