International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

FusionNet: A Supervised Multi-Modal Deep Learning Framework for Enhanced Product Recommendation Using SBERT and MobileNetV2

Tharaa Melhem
Master in Web Science, Syrian Virtual University, Damascus, Syria.

Dr. Sira Astour
Faculty of Information and Communication Engineering, Arab International University (AIU), Daraa, Syria.

Abstract - The rapid expansion of e-commerce platforms has rendered recommender systems indispensable for alleviating information overload and facilitating product discovery. However, conventional collaborative filtering and content-based approaches are fundamentally constrained by data sparsity, the cold-start problem, and their reliance on single-modality signals, which collectively undermine recommendation quality in dynamic environments. To address these limitations, this paper presents FusionNet, a supervised multi-modal deep learning framework that jointly learns unified product representations from unstructured textual descriptions and visual imagery without requiring historical user–item interactions. Specifically, textual attributes are encoded using a Sentence-BERT model fine-tuned on the Stanford Natural Language Inference (SNLI) and Multi-Genre Natural Language Inference (MNLI) corpora to capture nuanced semantic relationships. Concurrently, visual features are extracted via a lightweight MobileNetV2 backbone and projected into a compatible embedding space. The MobileNet model is specifically designed for high-efficiency operation on smart devices and embedded systems. It achieves this by drastically reducing its computing resources compared to traditional neural networks, requiring only very low processing power (low computation) because it relies on Depthwise Separable Convolutions instead of traditional convolution. Furthermore, MobileNet occupies a very small amount of storage space (between 16 and 32 MB depending on the version used) and reduces the number of calculations, thus minimizing the load on the processor. These heterogeneous features are subsequently fused through a multilayer perceptron-based fusion network, which is explicitly supervised by product category labels to learn a discriminative shared embedding space of 256 dimensions. Recommendation retrieval is then performed via cosine similarity over this unified space. Experiments conducted on Amazon product corpus comprising 459,229 items across 112 categories demonstrate the efficacy of the proposed framework. FusionNet achieves a Precision@10 of 0.860, an NDCG@10 of 0.861, and a HitRate@10 of 0.933, indicating that over 93% of the generated recommendation lists contain at least one relevant item within the top ten positions. Furthermore, at k=7, the model attains an NDCG of 0.902 and a product coverage of 0.920, confirming its ability to surface diverse and relevant items effectively. Comprehensive ablation and robustness analyses further corroborate that the synergistic integration of semantic and visual modalities, guided by categorical supervision, significantly mitigates the cold-start challenge and enhances overall system reliability. The results establish FusionNet as a highly accurate, scalable, and deployable content-driven solution for modern e-commerce and digital content platforms.

***Keywords*: Recommender Systems, Multimodal Learning, Sentence-BERT (SBERT), MobileNetV2, Deep Learning, Product Recommendation, Multimodal Fusion, E-Commerce, Content-Based Recommendation.**

1. INTRODUCTION:

With the current technological revolution going on, people are getting a lot of information, which makes it difficult for people to get some interesting or relevant information. One of the efficient methods that can solve this particular issue is recommendation system. Recommendation systems filter out the item and recommend those items according to the preferences and behaviors of users, such as movies, products, news articles, and songs [1]. Many recommendation techniques have been devised that have proved to be very effective, including collaborative filtering and content-based filtering techniques. The basic premise behind collaborative filtering is that people who have agreed in the past are expected to agree in the future too, whereas content-based filtering relies on attributes of the item and the user profile [2]. However, there are some shortcomings in these techniques as well. Collaborative

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

filtering often suffers from data scarcity and cold start issues [3], while content-based approaches can be too narrow and fail to generalize to the diverse needs of users [4].

To address these challenges, the field has increasingly turned to more intelligent systems powered by artificial intelligence, particularly deep learning which offers a unique ability to learn rich representations from large-scale, high-dimensional data and captures subtle patterns and nonlinear relationships that traditional models might miss [5].

In terms of recommendation systems design, the shift to the multimodal system design, in other words, the integration of several types of data such as descriptions of the objects of interest, images, and classes, becomes a major breakthrough in comparison to the unimodal designs. The core advantage of multimodal processing of data can be considered its capability to exploit the complementary information presented in different perceptual channels. The objects in the context of e-commerce are inherently multi-sensory; the description provides explicit features and functionalities of the product while the image implies the implicit features that cannot be expressed in language. Hence, taking into account two pieces of information at once, the recommendation algorithm creates a detailed and complete representation of the object in order to close the semantic gap between raw data and user preferences. Therefore, there are certain advantages that multimodal processing has for the recommendation algorithms: (1) increased representational capacity as text and vision combination enables distinction between semantically similar and visually dissimilar objects ("black dress" with various shapes) and vice versa; (2) intrinsic robustness and fault tolerance as the lack of one modality (absent images or description) does not influence negatively on the performance of the algorithm as the missing information can be compensated by another modality; (3) increased generalization and prevention of overfitting as multimodal learning prevents overfitting as the algorithm is not capable of finding the patterns in the noisy data in one modality but has to discover cross-modal patterns; (4) considerable improvement of the cold-start problem as the objects without any history of the user interaction become recognizable due to rich multimodal representations and can be recommended; and (5) increased user satisfaction as the recommendations provided with the help of multimodal reasoning reflect human cognitive abilities and incorporate both textual and visual features of the objects.

The core problem addressed by this research is the lack of intelligent, language-aware, and visually conscious frameworks that can effectively combine unstructured textual content and diverse product images to deliver rich, personalized recommendations. This research justifies the growing demand for more adaptive and comprehensive recommendation systems that go beyond traditional, rule-based approaches. This study aims to expand the boundaries of what recommendation systems can do, leading to more beneficial user experiences, improved system performance, and broader applications in areas such as e-commerce and beyond.

The aim of the research is to examine the use of the S-BERT model [6] to improve the intelligence, adaptability and intention understanding of recommendation systems. In particular, there are some disadvantages of traditional recommendation methods connected with the usage of structured interaction between the user and items and, therefore, their inability to obtain relevant information from unstructured input data such as natural language input data. Thus, the appearance of BERT and SBERT models makes it possible to solve this issue and make the system analyze natural language text [7]. The main task of the current research is to develop a hybrid recommendation system which utilizes a finely tuned Sentence-BERT (SBERT) model, trained on Stanford Natural Language Inference (SNLI) and Multi-Type Natural Language Inference (MNLI) datasets, to extract semantic implications from product titles and descriptions, while the MobileNetV2 model is used to extract visual features from product images. These heterogeneous features are then combined through a supervised FusionNet to identify a shared embedding space guided by product category classifications.

2. LITERATURE REVIEWS:

There are different recommendation algorithms designed to solve different problems, such as the cold start problem, the multi-behavioral learning problem, and the time distribution problem. While some approaches rely on teacher-student frameworks for effective knowledge transfer, others push boundaries through complex graphical models or federated learning-based models. In this context, the articles by [8] and [9] utilize teacher-student learning methods to address the cold-start problem. With respect to the article by [8], the issue that this article focuses on relates to distillation, which will help in enhancing recommendation systems for new users and new items. A good starting point can be provided due to the transfer of knowledge from the pre-trained model. On the contrary, the article by [9] reinforces thSe concept through the introduction of uncertainty-aware learning along with the teacher-student learning process. Thus, through this method, the goal of accommodating the new and old relationships between users and

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

items can be achieved. Although both address the same problem, [9] introduces a level of nuance by explicitly addressing uncertainty, while [8] offers a more direct implementation.

The past studies conducted by [10], [11], [12], and [13] were focused exclusively on graph-based methods. In addition, each of the cited works brings its unique view on how the use of graphs may help to improve recommendation systems. For example, [10] present an approach based on the use of GNNs and self-attention that makes it possible to deal with multi-behavioral signals of users. The use of attention layers allows dynamic weight assignment to user behavior. Meanwhile, [11] propose hypergraph alignment to provide neighborhood consistency between domains, making their model applicable in the cross-domain recommendation task. In addition, [12] suggest an approach based on policy learning and hypergraph differential learning and applicable in the task of conversational recommendation. Finally, the paper by [13] offers an innovative solution based on GNNO that applies graphs not for passing messages but for sampling negative instances. Although these designs achieve high accuracy and responsiveness, they rely heavily on a graph neural network, which is known for its high computational requirements.

The research paper [14] departs from the traditional centralized model and explores decentralized collaborative learning by leveraging federated learning. The federated learning enables this research paper to perform local evolution of the user model without requiring any data exchange. It is the strength of this research paper that enables the provision of enhanced privacy to the users. However, there are certain issues related to synchronization and communication in unreliable network connectivity in this research paper.

There are also other methods of doing so, discussed in the paper [15]. In this method, pre-trained semantic compression encoder-decoder framework is employed. The use of such a multitask framework makes it easier for the system to comprehend the contextual cues embedded in the news articles, hence improving the recommendation of trendy or niche content. Although structurally different from graph-based models, this method plays a similar role in capturing complex user-content relationships, but through language and semantics rather than structural topology.

Both [16] and [17] papers are related to the topic of changing user preferences through time and structure. The issue of feature leakage in GNNs for [16] is resolved with the use of de-edgement approach which will help to correct the inclusion of nodes before their usage and, therefore, improve representation. Meanwhile, the paper of [17] relates to the user behavior at various timescales of short-term and long-term with the help of semantic content. The Long-Term and Short-Term Attention Network (TLSAN) provides a multi-layered perspective on how users change over time, particularly in e-commerce environments.

Analysis of the research papers reveals that significant progress has been made in the field of recommendation systems, especially toward the development of hybrid or intelligent systems. Graph Neural Network (GNNO), Teacher-Student architecture, attention-based models, and differential learning are no longer novel technologies but constitute the base of modern sophisticated models. For example, exploitation of the structure of the data in GNNO and MMNC helps reveal that proper utilization of such concepts can result in better recommendations. Another concept gaining more importance in recent times is privacy, which can be achieved through the development of federate learning models. Furthermore, time-sensitive and semantically rich models are gaining traction, enabling systems to better understand user behavior dynamics and content context.

However, there are several questions that have not been answered yet. The first question concerns computational complexity. Systems that are based on neural networks with GNNs, attention mechanisms, and complex passive sampling have very low scalability. Thus, such systems cannot be applied in the situation where there is a lack of computational resources or when there are time limitations. Another question is the question of model generalization. It is evident that most of the described systems, which utilized particular structures of cipher-decipher or scale-compliant hypergraphs, demonstrated excellent results in the process of working with training datasets, but could not work effectively with various hidden sources of data. Besides, there is a problem of interpretability.

To bridge these gaps, future research should prioritize the development of lightweight and efficient models that maintain high performance without overburdening system resources. Exploring knowledge extraction techniques and improved parameter training may help achieve this balance. There is also a clear need to invest in multimodal recommendation techniques, such as combining natural language processing models like S-BERT with image processing models like the lightweight MobileNet model, to deliver more interactive and semantically rich recommendations.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Progress made in 2026 has added to the frontier of multimodal and deep learning recommendation systems. For instance, CineFinder [18] is a multimodal recommendation system for movies which uses VGG-16, ResNet-50, and MobileNet visual features and BERT, RoBERTa, and SBERT textual features. This study proves that the use of multiple pre-trained models can effectively address the cold-start problem. In literature related to e-commerce, paper [19] proposed a multimodal recommendation approach based on user-product ratings and reviews and images of products via an LLM-driven review parsing pipeline combined with GraphSAGE-based textual representation learning and CLIP-based visual encoding enriched with psychology-informed prompts, achieving robust cold-start performance across four Amazon datasets. The research paper by [20] presents a RDNMF method that addresses the cold start problem through a neural matrix factorization approach, which incorporates the use of pre-trained BERT with LSTM to enable temporal preference modeling and also integrates reviews and item description in its recommendation model, resulting in an increase in HR@10 by 3.43% and NDCG@10 by 5.85%. Overall, these research papers from 2026 indicate a clear consensus about the significance of multimodal fusion irrespective of whether it is with SBERT, CLIP or LLM-enhancement.

3. METHODOLOGY:

Based on the research gaps identified and analyzed in Section 2, primarily the high computational cost of graph network-based (GNN) models, poor performance in cold start scenarios due to over-reliance on behavioral interaction data, and the inadequacy of fully integrated and efficient use of rich multimedia content (text and images), this paper proposes a multimedia recommendation system model based on the multimedia integration of textual and visual features. The suggested architecture consists of two main stages: first stage includes extracting semantic information of texts using Sentence-BERT (SBERT) model fine-tuned on SNLI and MNLI datasets; and the second stage includes extracting visual features using lightweight pre-trained MobileNetV2 model projected to one dimensional feature space. In turn, the second approach suggests fusing of heterogeneous features using FusionNet, which is a supervised multi-layer perceptron (MLP) neural network. MLP projects features into a shared embedding space (256 dimensions) guided by product category classifications (112 categories). This space is then used to build a recommendation system based on cosine similarity to return the best K recommendations. This framework aims to provide a highly accurate solution capable of handling data scatter and cold-start functionality without requiring user interaction, while maintaining computational efficiency suitable for deployment in e-commerce applications. The proposed solution consists of four main stages:

1) Identifying the datasets to be used.

2) Training the SBERT model to learn semantically similar text inclusions.

3) Using the MobileNet v2 model to extract image features [21].

4) Building the proposed FusionNet, which projects the extracted, heterogeneous textual and visual features into a shared design space.

5) Building the recommendation system.

6) Evaluating the quality of the recommendations.

Figure 1 illustrates the general structure of the proposed methodology.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

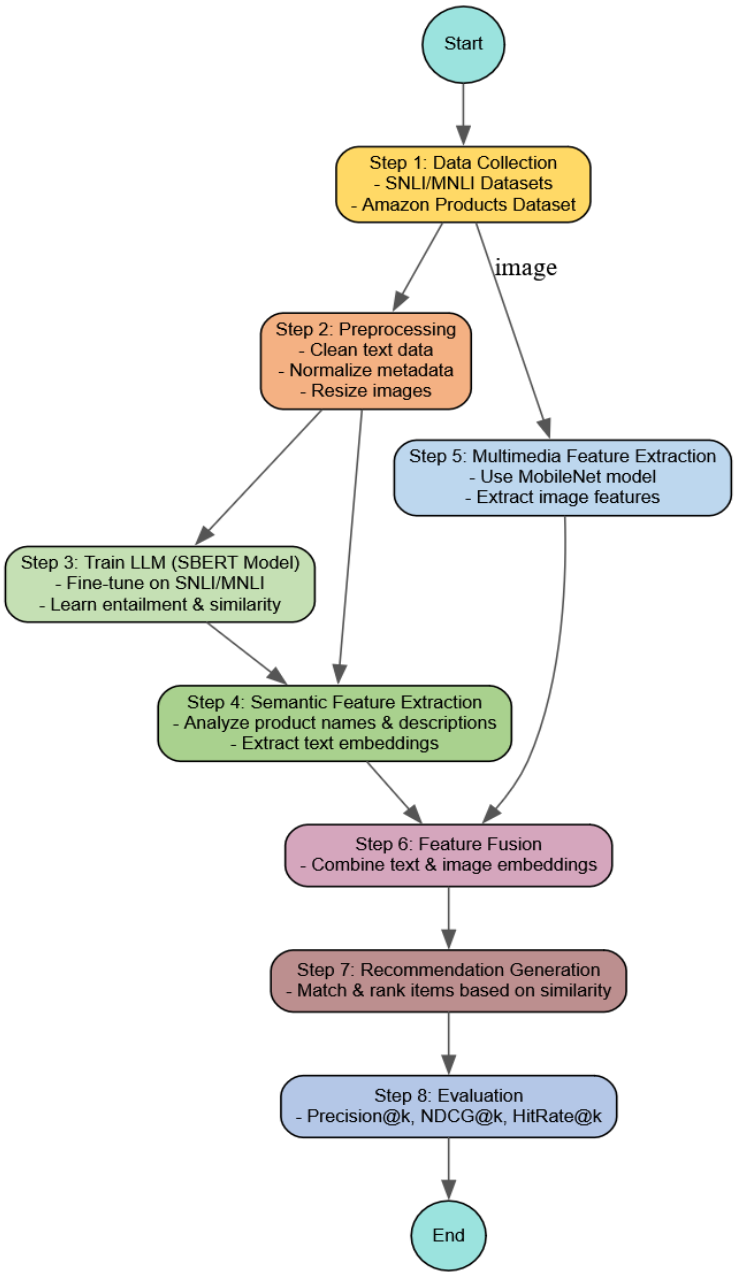


Figure 1. The general structure of the proposed methodology

3.1 Learn text embedding using the SBERT model:

In this research, two large natural language inference datasets, SNLI [22] and MNLI [23], were used to train an SBERT model for sentence embedding to capture semantic similarity between text pairs.

The choice of SBERT as a textual encoder in the proposed framework is driven by several advantages in comparison with other approaches to sentence embedding, specifically the classic BERT cross-encoder architecture. While BERT is recognized as a model producing state-of-the-art results on various sentence-pair regression tasks, including semantic textual similarity (STS), the fundamental characteristic of the model makes it impossible to utilize BERT in a cost-effective way for similarity search and clustering [24]. Namely, BERT needs two sentences to be fed simultaneously through the network in order to compare them, resulting in $O(n^2)$ complexity of the algorithm in terms of the size of the dataset [25]. For instance, to find the most similar pair of sentences among the 10,000 sentences, one would have to make about 50 million forward passes, requiring around 65 hours of computations [24]. In order to overcome this problem, SBERT proposes an architectural modification of BERT through implementing the Siamese network architecture involving two BERT models with identical weights that transform sentences into

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

fixed-size vectors separately. Comparison of such vectors via cosine similarity allows reducing the amount of time needed to conduct the above similarity search from 65 hours to about 5 seconds with the preservation of accuracy [24]. The trick behind the efficiency of SBERT is the utilization of the pooling mechanism, such as mean pooling of all token embeddings, that converts the variable-size output of BERT into the fixed-size sentence embeddings. Besides the improvement in terms of efficiency, SBERT proves to generate semantically meaningful sentence embeddings, outperforming the simple BERT output averaging approach which often generates poor representations being inferior even to the simple averaging of GloVe embeddings [24]. The architecture of SBERT used in the study is based on the BERT-base variant with the following specifications: 12 Transformer layers, 768 hidden dimensions and 12 attention heads resulting in approximately 110 million trainable parameters [26]. The model seems to provide a decent balance of computational performance and representational power for practical application in the e-commerce setting. SBERT generates sentence embeddings of 384 dimensions for each product title and description that are then combined with the visual features generated by MobileNetV2 model.

Assume that D_{snli} and D_{mnli} be the SNLI and MNLI datasets, respectively. We combine these two datasets to train the model on a larger number of sentences, resulting in the unified dataset D as shown in formula (1):

$$D = D_{snli} \cup D_{mnli} \qquad (1)$$

The entailment samples are then filtered out, since each sample in the dataset D has a label ($y \in \{0,1,2\}$) where each of these symbols represents entailment, neutrality, and contradiction, respectively. At this stage, only the entailment pairs are retained. The goal here is to train the model on semantically similar sentences, enabling it to later make recommendations based on the semantic similarity of products with similar descriptions. Formula (2) shows the remaining data set after retaining only the entailment pairs.

$$D^{+} = \{(p,h) \in D | y = 0\} \qquad (2)$$

Where:
p: The premise.
h: The hypothesis.

Table 1 shows samples of the datasets used.

Table 1. Samples of the datasets used

Class	hypothesis	premise
neutral	A person is training his horse for a competition.	**A person on a horse jumps over a broken down airplane.**
contradiction	A person is at a diner, ordering an omelette.	**A person on a horse jumps over a broken down airplane**
entailment	A person is outdoors, on a horse.	**A person on a horse jumps over a broken down airplane**
neutral	They are smiling at their parents.	**Children smiling and waving at camera**
entailment	There are children present.	**Children smiling and waving at camera**
contradiction	The kids are frowning	**Children smiling and waving at camera**

Fine-tuning of the SBERT model on SNLI and MNLI entails the full end-to-end fine-tuning procedure when all the layers of transformers are updated in the course of training, but there is no layer freezing or parameter isolation whatsoever. This is

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

required because of the semantic nature of the task, which is impossible to complete without understanding the linguistic relation between premise-hypothesis sentence pairs and, thus, is impossible without the adaptability of all the layers of representation of the network from low-level syntactic to high-level semantic representations. Hence, there is no layer freezing applied. The only element of the architecture that remains pre-trained is the tokenizer. This end-to-end fine-tuning strategy, as proposed by [24, 27], has proven to yield better results for sentence-pair tasks compared to the strategies that involve layer freezing or task-specific heads updating only.

The S-BERT model is fine-tuned on SNLI and MNLI datasets, focusing solely on entailment pairs to learn to link semantically similar sentences together. The `MultipleNegativesRankingLoss` loss function [28], a commonly used loss function for training sentence embedding models, particularly in information retrieval and semantic search tasks, was employed. This is a form of contrastive learning that uses negatives within a batch to increase the similarity between a baseline text and its corresponding text pair, while decreasing the similarity with all other samples (text pairs) in the same batch [29].

Through this stage, we obtain the SBERT model, which has been trained to distinguish sentences that are semantically similar to each other, and which will be used within the Amazon Products dataset to extract products that are similar in terms of name and description. Figure 2 illustrates the mechanism for adjusting and training the SBERT model using the SNLI and MNLI datasets.

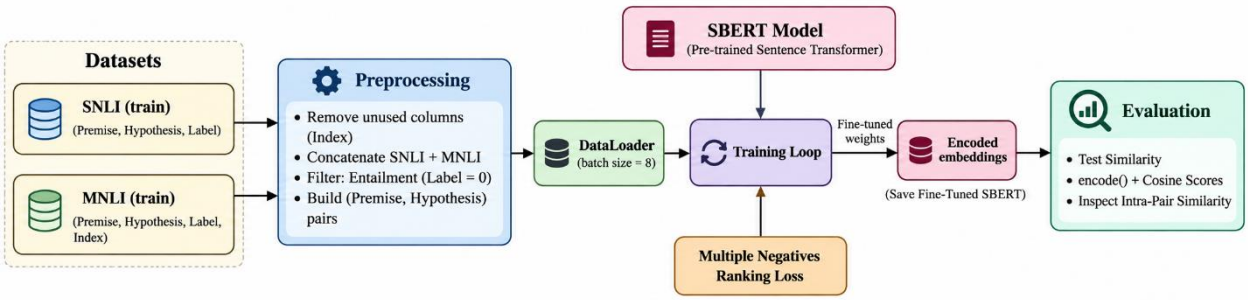


Figure 2. The mechanism for adjusting and training the SBERT model using the SNLI and MNLI datasets

3.2 Multimodal Feature Extraction:

This research uses an Amazon Products dataset described as $\mathcal{P}(t_i, v_i, y_i)$ where:

t_i: Product name and description.

v_i: Image URL.

y_i: Product category.

First, the features for all product names and descriptions, which are textual, are extracted using the SBERT model which fine-tuned on SNLI and MNLI datasets as shown in formula (3):

$$t_i = SBERT\ (t_i) \in R^{384} \qquad (3)$$

Image features are extracted using a MobileNetV2 model that extracts convolutional features from images, where the model is truncated in the final convolutional layer. These features are compiled and converted into a one-dimensional v vector of size 1280. A Linear Projection layer is then added to convert the dimension of the vector v from 1280 to 384 so that it is equal to the dimensions of the text embedding. Figure 3 illustrates the mechanism for extracting text and image features.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

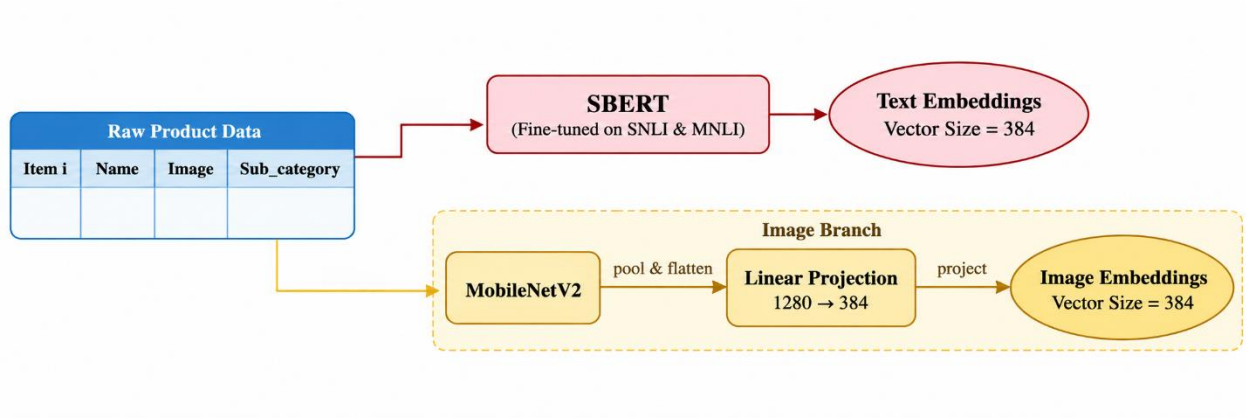


Figure 3. The mechanism for extracting text and image features

3.3 Supervised Fusion Network:

To enable the model to learn the shared embedding space of multimodal features and thus make accurate product recommendations, we propose a multilayer perceptron artificial neural network that combines textual and image features, and converts them into a smaller-dimensional shared embedding space, and then assigns the shared embeddings to product categories using supervised learning [30].

The artificial neural network consists of four layers:

1. Input layer: 768 neurons representing 384 text features extracted using SBERT and 384 image features extracted using the MobileNetv2 model.
2. Hidden layer 1: 512 neurons .
3. Hidden layer 2: 256 neurons .
4. Output layer: 112 neurons representing the number of product categories within the Amazon Products dataset.

Figure 4 illustrates the structure of the proposed FusionNet.

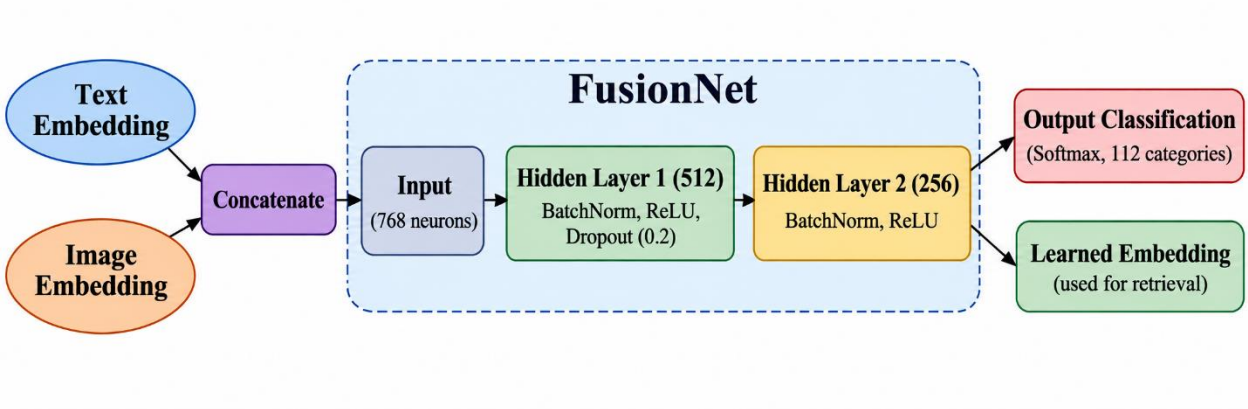


Figure 4. The structure of the proposed FusionNet

3.4 Building the recommendation system:

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

For all products in the dataset, the final embeddings are extracted using the trained proposed FusionNet as shown in formula (4):

$$e_i = f_{fusion}(t_i, v_i) \in \mathbb{R}^{256} \quad (4)$$

These embeddings are stored in an $E \in \mathbb{R}^{M \times 256}$ matrix where M represents the number of products within the dataset.
In order to extract recommendations for product q, the cosine similarity s_i is calculated between the product q embedding and all products embeddings in the dataset as shown in formula (5):

$$s_i = \frac{e_q e_i}{\|e_q\| \, \|e_i\|} \quad \forall i \in \{1, \dots, M\} \quad (5)$$

The recommendation system recommends the highest K products, which represent the products with the highest similarity as shown in formula (6):

$$\mathscr{R}_q = \{p_{i1}, p_{i2}, \dots, p_{ik} \mid s_{i1} \geq s_{i2}, \dots, s_{ik}\} \quad (6)$$

Figure 5 illustrates the mechanism for obtaining recommendations and evaluation.

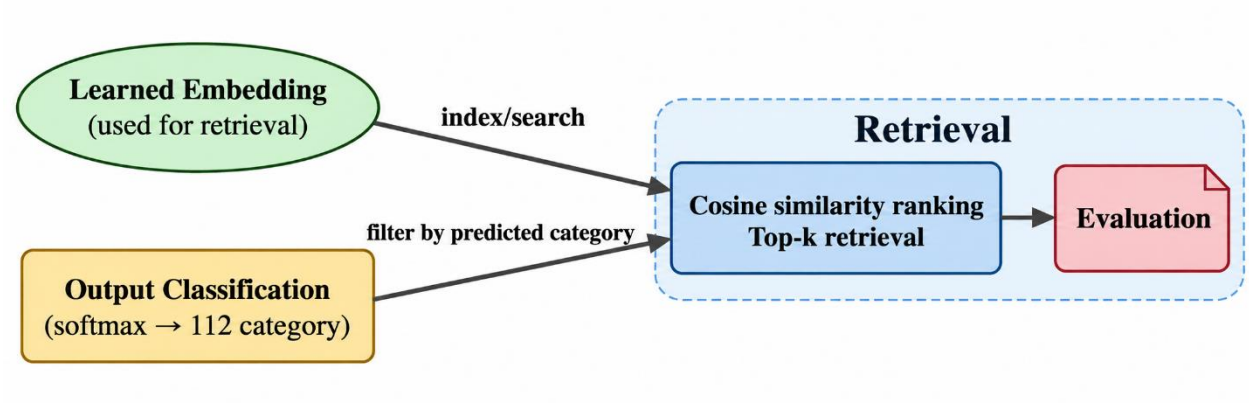


Figure 5. The mechanism for obtaining recommendations and evaluation

3.5 Evaluation Metrics:

To evaluate the effectiveness of the proposed recommendation system, information retrieval metrics tailored to the product recommendation settings are used. Relevance is defined as two relevant items belonging to the same product category.

The Precision@k value is calculated as described in formula (7). A high Precision@k value indicates that the recommendation system retrieves few unrelated items among its primary recommendations [31].

$$Precision@k = \frac{Number\ of\ relevant\ items\ in\ top}{k} \quad (7)$$

The Normalized Discounted Cumulative Gain (NDCG) metric is also used. This metric not only considers the importance of the recommended items but also focuses on their ranking [32], where higher-ranking, more important items contribute more to the score. The NDCG value is calculated as shown in formulas (10):

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

$$DCG_i = \sum_{j=1}^{k} \frac{2^{rel_{i,j}} - 1}{\log_2(j+1)} \qquad (8)$$

$$IDCG_i = \sum_{j=1}^{\min(\mathcal{R}_i^{all},k)} \frac{1}{\log_2(j+1)} \qquad (9)$$

$$NDCG = \frac{1}{M} \sum_{i=1}^{M} \frac{DCG_i}{IDCG_i} \qquad (10)$$

Where:
rel_i: The binary relevance of the product with the product at position i, where its value is 1 if the two products are related; otherwise, it takes the value 0.

HitRate is a metric that measures the proportion of recommendations where at least one relevant item was found within the top K scores [33]. The HitRate is formulated as shown in formula (11):

$$HitRate = \frac{1}{M} \sum_{i=1}^{M} 1 \ : \ \exists\, j \le k \ and \ rel_{i,j} = 1 \qquad (11)$$

The Mean Average Precision (MAP) measures the average precision across all relevant elements, calculated for all queries. Assume that Q represent the set of all queries (products). For a single query q, assume R_q is the set of relevant products (of the same class as q). The average precision at k for query q is defined as shown in formula (12):

$$AP@k = \frac{1}{\min(R_q, k)} \sum_{r=1}^{k} P@r \qquad (12)$$

Where $P@r$ represents the precision value at r. Therefore, $MAP@k$ is the average $AP@k$ of all queries as shown in formula (13):

$$MAP@k = \frac{1}{Q} \sum_{q \in Q} AP_q@k \qquad (13)$$

The Mean Reciprocal Rank (MRR) focuses on the ranking of the first relevant item in a list of recommendations. For each query q, assume $rank_q$ is the position of the first relevant item. The Reciprocal Rank (RR)is shown in formula (14):

$$RR_q = \frac{1}{rank_q} \qquad (14)$$

Therefore, $MRR@k$ is the average RR for all queries as shown in formula (15):

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

$$MRR@k = \frac{1}{Q}\sum_{q \in Q}\frac{1}{rank_q} \qquad (15)$$

MRR is not affected by additional relevant elements after the first one, it only cares about the first correct recommendation.

The Coverage Index measures the proportion of featured items that appear in at least one recommendation list across all queries. Assume N is the set of all items. R_q is defined as the set of items recommended for the query q (excluding q itself). The set of items recommended across all queries is shown in formula (16):

$$R_{total} = \bigcup_{q \in Q} R_q \qquad (16)$$

Therefore, the value of $Coverage@k$ is shown in formula (17):

$$Coverage@k = \frac{R_{total}}{N} \qquad (17)$$

A value close to 1 indicates that almost all products are recommended to some users, thus avoiding the trap of highly popular products where only a few products are suggested frequently.

4. RESULTS AND DISCUSSIONS:

The experiments were conducted on the Amazon Products dataset [34], which includes 459,229 products across 112 different categories. Each product record contains the following:

1) Textual data: Product title and description.

2) Visual data: Links to product images.

3) Category classifications used in supervised training.

4.1 Ablation Study:

The proposed multimedia recommendation system demonstrated outstanding performance across all evaluation metrics. Table (2) shows the ablation study that evaluating the results of: (1) textual features extracted using only the S-BERT model, (2) visual features extracted using only the MobileNet v2 model, and (3) the proposed model representing the FusionNet, which integrates textual and visual features within a shared embedding space. These results demonstrate the benefits of integrating these components and the reduced performance of each when used in isolation. The results are extracted for ten recommendations (K=10) per product within the dataset.

Table 2. Ablation Study

Metric	Text-only SBERT	Image-only MobileNet	Proposed System
Precision@10	0.2751	0.1051	**0.8602**
NDCG@10	0.3027	0.1134	**0.8609**
HitRate@10	0.7179	0.3816	**0.9329**

The proposed system achieved a Precision@10 score of 0.86 for each item in the dataset, indicating that 86% of the top 10 recommendations were relevant to the query item. The NDCG@10 score of 0.86 also indicates excellent ranking quality, with

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

the most relevant items appearing at the top of the recommendations list. Figure 6 shows examples of the recommendations provided by the model.

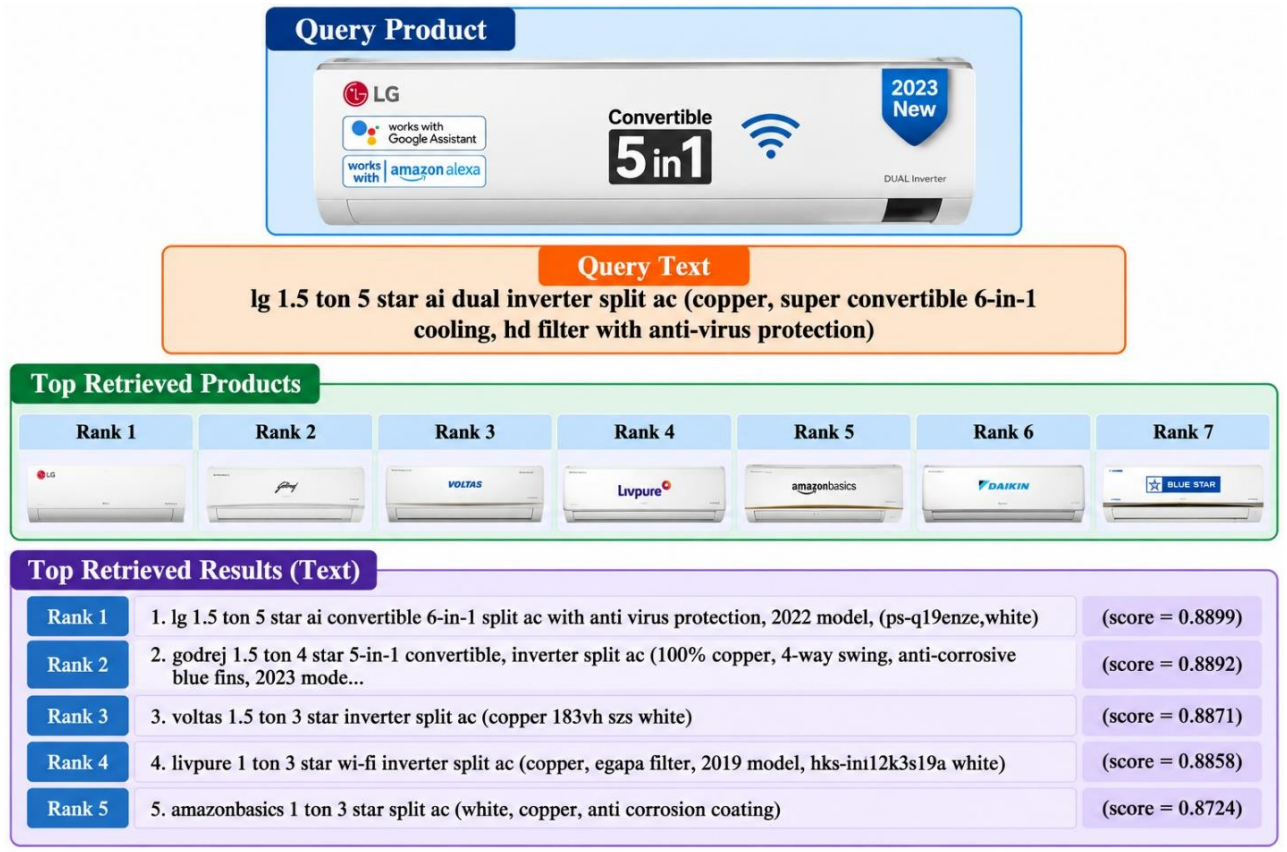


Figure 6. Examples of the recommendations provided by the model

Table 3 shows a comparison of the proposed model with a range of previous studies.

Table 3. A comparison of the proposed model with a range of previous studies

Ref	Year	Method	Precision	NDCG	HitRate
[35]	2024	Graph Neural Networks (GNNs)	0.50	0.53	-
[36]	2025	Large Language Models (LLM)	-	0.49	-
[37]	2025	multilayer perceptron (MLP)	0.618	0.567	-
[38]	2025	Graph Neural Networks (GNNs)	0.859	0.452	-
[39]	2025	Graph Neural Networks (GNNs)	-	0.046	0.279
[40]	2026	GRAFT-Rec (Graph-Augmented Fusion with Gradient-Boosted Meta-Ranking)	0.0108	0.0570	0.1080
[41]	2026	Hybrid NeuMF + LSTM (Early-Fusion Deep Learning)	0.143	0.139	-

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

[42]	2026	LSTM + Bayesian Personalized Ranking (Deep Learning with BPR)	0.0068	0.1892	0.6763
Proposed model	2025	S-BERT+ MobileNet v2 + FusionNet	0.8602	0.8609	0.9329

4.2 The overall performance of the proposed recommendation system:

Table 4 summarizes the overall performance of the proposed recommendation system.

Table 4. The overall performance of the proposed recommendation system

Metric	Proposed System
Precision@7	**0.8626**
NDCG@7	**0.9020**
MAP@7	**0.8915**
MRR@7	**0.8959**
Coverage@7	**0.9196**

Based on the results in Table 4, the proposed model achieves a high NDCG@7 index value of 0.902, indicating that relevant items are ranked high on the list. The Precision@7 value of 0.863 shows that over 86% of the recommended items are relevant. The Coverage@7 value of 0.92 demonstrates that the system recommends a wide range of items (only about 8% of products never appear on any recommendation list).

Figure 7 illustrates the distribution of NDCG@7 across individual queries. The majority of queries achieve near-perfect scores, with a long tail on the left representing difficult cases involving diverse shapes, colors, and specifications, such as clothing.

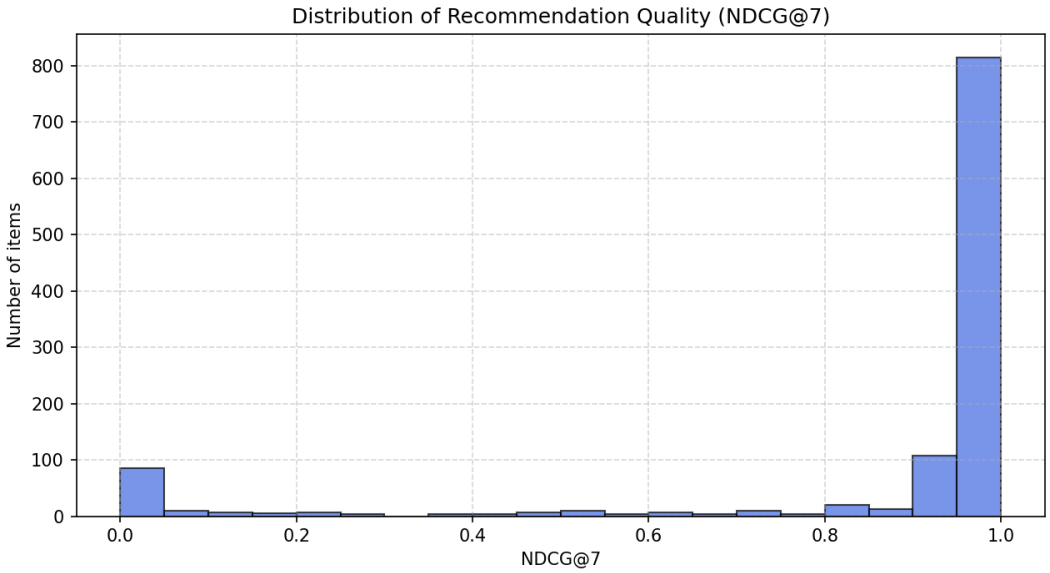


Figure 7. The distribution of NDCG@7 across individual queries

4.3 Diversity and fairness:

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Intralist diversity (ILD) measures the average difference between pairs of items within a list of recommendations, using (1-cosine similarity) as shown in fromula (18).

$$ILD@k = \frac{1}{k(k-1)} \sum_{i \neq j} \left(1 - sim(z_i, z_j)\right) \qquad (18)$$

When k is 7, the ILD is 0.1064 (with a standard deviation of 0.0800). This value indicates that the recommended items are often very similar. While this achieves high accuracy, it limits the likelihood of unexpected results, as users may prefer more varied recommendations. Nevertheless, this value is suitable for our primary goal of tailoring recommendations to each product category.

We assess fairness by comparing the NDCG@10 between head categories (top 20% frequency – ranked by original frequency in raw data) and tail categories (bottom 20% frequency). The difference is shown in formula (19):

$$Disparity = NDCG_{head} - NDCG_{tail} = -0.0147$$

With a head NDCG value of 0.894 and a tail NDCG value of 0.908, the negative variance indicates that the tail classes perform slightly better, meaning that the model is not biased towards the most common classes within the dataset.

4.4 Analysis of the embedding space:

To analyze the structure of the extracted multimodal embedding space using the proposed model, we calculated cluster purity by applying the K-means algorithm with 112 clusters (the number of classes in the dataset).

The cluster purity (88.3%) shows that when embeddings are forced to belong to 112 classes (clusters), most clusters contain elements from a single correct class. Only 11.7% of the elements were incorrectly classified.

4.5 Evaluate the model robustness:

To simulate real-world scenarios where a media component might be missing (e.g., image loading failure or text extraction error), we evaluate three variables:

1) Full multimedia (text and image).

2) Missing image (image embedding replaced with a zero vector).

3) Missing text (text embedding replaced with a zero vector).

Table 5 shows the NDCG@7 value for each case.

Table 5. The NDCG@7 value for each case

Condition	NDCG@7	Absolute drop	Relative drop
Full multimodal	0.9020	-	-
Missing image	0.8523	0.0497-	%5.5-
Missing text	0.7206	0.1814-	%20.1-

Removing the image results in a slight performance decrease of 5.5%, indicating that the text alone carries most of the distinctive signal. Conversely, removing the text leads to a sharp 20% decrease, highlighting the importance of textual product names and specifications. The system is robust when the image is sometimes unavailable, but becomes weaker when the text is absent.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

5. CONCLUSIONS AND FUTURE WORKS:

In this paper, the employment of multimedia representation was indeed a highly effective means of improving the performance of recommendation systems. The proposed FusionNet whereby text embeddings from the SBERT model are combined with image embeddings derived from MobileNetV2 has shown to possess a high capability of detecting the connection between different products in terms of semantics and visuals. As a result, there has been a significant improvement in the quality of recommendations made compared to situations whereby only one kind of data is used for making recommendations. This is evident from the high scores in evaluation parameters such as Precision@10, NDCG@10, and HitRate@10. The experiments have also shown that the proposed Supervised Fusion Network is able to learn the efficient joint embedding space that would allow representing the products of various categories. One of the most significant results is that the proposed approach is able to solve the problem of cold start by making recommendations for unseen products. It means that the usage of deep semantics and multimedia fusion is the perspective field of recommendation systems development. The results showed that the text-based model based on SBERT performed acceptably on its own (Precision@10 = 0.275 and NDCG@10 = 0.303), confirming its ability to capture semantic similarity between product descriptions. However, it remained limited in handling the visual diversity of products, justifying the need for multimedia integration. On the other hand, the results also showed that relying on images only performed very poorly (Precision@10 = 0.105 and NDCG@10 = 0.113), indicating that visual features alone are insufficient. However, they make a significant difference when combined with text, increasing the Precision@10 =0.860, thus proving that multimedia integration is the true key to improvement. The proposed FusionNet demonstrated its ability to learn a unified embedding space, achieving a clustering purity of 88.3% when the embeddings were divided into 112 categories. This confirms the network's success in placing similar products in close proximity within the space. Compared to previous studies on the same topic, the proposed model outperformed the latest models, achieving an NDCG@10 value of 0.861, compared to 0.567, 0.53, and 0.452 in other studies based on MLP, GNN, and LLM. Finally, the proposed model achieved a HitRate@10 of 93.3% and covered 92% of the products in the recommendations (Coverage@7), confirming that the integration of textual embeddings and visual features under category supervision produces a robust system capable of overcoming the challenges of data scatter.

Based on the analysis of the results obtained, the following directions for the future development of the system are recommended. First, the system can be improved by integrating Large Language Models (LLMs) directly into the process of user intent recognition or conversation recommendation generation. Second, it would be interesting to study more complex approaches to LLMs' integration, such as multimodal attention models and dynamic integration, instead of static integration that would give more weight to the most important mediator based on the product type. Moreover, it would be interesting to evaluate the system's performance under more realistic conditions using interaction data of users, such as clicks or purchases.

REFERENCES:

[1] A. J. E. Andrews and B. Furht, "A comprehensive survey of recommendation methods and techniques," *Journal of Big Data,* 2026.
[2] F. Li, S. Liu, and H. Wang, "Collaborative filtering or content-based recommendation? The effects of digital platform recommendation type on consumer's intention," *Journal of Retailing and Consumer Services,* vol. 90, p. 104694, 2026.
[3] J. Guan, B. Chen, and S. Yu, "A hybrid similarity model for mitigating the cold-start problem of collaborative filtering in sparse data," *Expert Systems with Applications,* vol. 249, p. 123700, 2024.
[4] N. Pittaras, G. Giannakopoulos, P. Stamatopoulos, and V. Karkaletsis, "Content-based and knowledge-enriched representations for classification across modalities: a survey," *ACM Computing Surveys,* vol. 55, no. 14s, pp. 1-40, 2023.
[5] A. Wilson and M. R. Anwar, "The future of adaptive machine learning algorithms in high-dimensional data processing," *International Transactions on Artificial Intelligence,* vol. 3, no. 1, pp. 97-107, 2024.
[6] R. Patil and V. Gudivada, "A review of current trends, techniques, and challenges in large language models (llms)," *Applied Sciences,* vol. 14, no. 5, p. 2074, 2024.
[7] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, "BERT applications in natural language processing: a review," *Artificial Intelligence Review,* vol. 58, no. 6, pp. 1-49, 2025.
[8] F. Huang, Z. Wang, X. Huang, Y. Qian, Z. Li, and H. Chen, "Aligning distillation for cold-start item recommendation," in *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, 2023, pp. 1147-1157.
[9] T. Liu *et al.*, "Uncertainty-aware Consistency Learning for Cold-Start Item Recommendation," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2466-2470.
[10] J. Xu *et al.*, "Multi-behavior self-supervised learning for recommendation," in *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, 2023, pp. 496-505.
[11] W. Tang *et al.*, "Multi-order matched neighborhood consistent graph alignment in a union vector space," in *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, 2023, pp. 963-972.
[12] S. Zhao *et al.*, "Multi-view hypergraph contrastive policy learning for conversational recommendation," in *Proceedings of the 46th International ACM SIGIR conference on research and development in information retrieval*, 2023, pp. 654-664.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

[13] L. Fan, J. Pu, R. Zhang, and X.-M. Wu, "Neighborhood-based hard negative mining for sequential recommendation," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2042-2046.
[14] J. Long, T. Chen, Q. V. H. Nguyen, G. Xu, K. Zheng, and H. Yin, "Model-agnostic decentralized collaborative learning for on-device poi recommendation," in *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, 2023, pp. 423-432.
[15] X. Xiao, Q. Li, S. Liu, and K. Zhou, "Improving news recommendation via bottlenecked multi-task pre-training," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2082-2086.
[16] I. Kumar, Y. Hu, and Y. Zhang, "Eflec: Efficient feature-leakage correction in gnn based recommendation systems," in *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*, 2022, pp. 1885-1889.
[17] S. Liang, Y. Zhang, and Y. Wang, "C-TLSAN: Content-Enhanced Time-Aware Long-and Short-Term Attention Network for Personalized Recommendation," *arXiv preprint arXiv:2506.13021,* 2025.
[18] M. T. Sarıçiçek, R. Orman, M. Dener, and H. Kınacı, "CineFinder: A Movie Recommendation System Using Visual and Textual Deep Features."
[19] H.-W. Lee and R. Kuo, "Integrating Large Language Model and Graph Neural Network for Knowledge Graph-Based Multi-Source Recommender System," *Applied Soft Computing,* p. 115749, 2026.
[20] K. Li and N. A. M. Kamal, "Deep recommendation algorithm based on reviews and descriptions neural matrix factorization under cold-start," *Complex & Intelligent Systems,* vol. 12, no. 2, p. 88, 2026.
[21] Y. Saini, A. Tyagi, and P. Kumar, "AI-Driven Real-Time Waste Classification and Nearest Bin Recommendation System," in *2026 IEEE 18th International Conference on Computational Intelligence and Communication Networks (CICN)*, 2026: IEEE, pp. 1510-1515.
[22] I. M. S. Putra, D. Siahaan, and A. Saikhu, "SNLI Indo: A recognizing textual entailment dataset in Indonesian derived from the Stanford Natural Language Inference dataset," *Data in Brief,* vol. 52, p. 109998, 2024.
[23] S. Giddaluru, S. M. Maturi, O. Ooruchintala, and M. Munirathnam, "Stance Detection in Hinglish Data using the BART-large-MNLI Integration Model," *Engineering, Technology & Applied Science Research,* vol. 14, no. 4, pp. 15477-15481, 2024.
[24] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 2019, pp. 3982-3992.
[25] A. Manenti, "Deep Learning techniques for Natural Language Processing: A multilingual encoder model for NLI task," Politecnico di Torino, 2022.
[26] Y. Santander-Cruz, S. Salazar-Colores, W. J. Paredes-García, H. Guendulain-Arenas, and S. Tovar-Arriaga, "Semantic feature extraction using SBERT for dementia detection," *Brain sciences,* vol. 12, no. 2, p. 270, 2022.
[27] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 3651-3657.
[28] L. Fitriani, S. F. Muzahidar, A. Sutedi, and F. Nuraeni, "A Bilingual Academic Chatbot Based on Semantic Retrieval Using mBERT," *JUITA: Jurnal Informatika,* pp. 313-322, 2026.
[29] S. Dehghan and M. F. Amasyali, "SupMPN: supervised multiple positives and negatives contrastive learning model for semantic textual similarity," *Applied Sciences,* vol. 12, no. 19, p. 9659, 2022.
[30] C. K. J. Hou and K. Behdinan, "Dimensionality reduction in surrogate modeling: A review of combined methods," *Data science and engineering,* vol. 7, no. 4, pp. 402-427, 2022.
[31] S. Zhang *et al.*, "Talos: Optimizing top-k accuracy in recommender systems," in *Proceedings of the ACM Web Conference 2026*, 2026, pp. 6666-6677.
[32] M. Louhichi, R. Nesmaoui, and M. Lazaar, "DyHuCoG: A Dynamic Hypergraph Cooperative Game for Preference-aware Recommendation," *International Journal of Intelligent Engineering & Systems,* vol. 19, no. 2, 2026.
[33] A. Aly, A. Ibrahim, and R. Kashef, "RAG-Rec: Retrieval Augmented Generation for Robust Personalized Recommendation Systems," *Procedia Computer Science,* vol. 275, pp. 1-8, 2026.
[34] Y. M. Latha and B. S. Rao, "Amazon product recommendation system based on a modified convolutional neural network," *ETRI journal,* vol. 46, no. 4, pp. 633-647, 2024.
[35] A. Medisetty, "Think buy: a scalable, context-rich AI model for personalized e-commerce recommendations," *International Journal of Computational Mathematical Ideas (IJCMI),* vol. 16, no. 1, pp. 3078-3091, 2024.
[36] Q. Lu, H. Lyu, J. Zheng, Y. Wang, L. Zhang, and C. Zhou, "Research on E-Commerce Long-Tail Product Recommendation Mechanism Based on Large-Scale Language Models," in *Proceedings of the 9th International Conference on Electronic Information Technology and Computer Engineering*, 2025, pp. 997-1002.
[37] H. Xia and Y. Wang, "Enhancing Neural Collaborative Filtering for Product Recommendation by Integrating Sales Data and User Satisfaction," *Electronics,* vol. 14, no. 16, p. 3165, 2025.
[38] M. Cao, F. F. Yang, Y. Jin, and Y. Yan, "Graph Neural Network for Product Recommendation on the Amazon Co-purchase Graph," *arXiv preprint arXiv:2508.14059,* 2025.
[39] C. Fan and S. Fujita, "Reinforcement learning-based recommender systems enhanced with graph neural networks," *IEEE Access,* 2025.
[40] P. Muniraja and S. M. Satapathy, "GRAFT-Rec: a hybrid multi-signal framework with gradient-boosted meta-ranking for collaborative filtering," *Social Network Analysis and Mining,* 2026.
[41] L. Efrizoni, "Optimizing Content Recommendations Using a Hybrid Filtering Algorithm to Enhance User Relevance and Engagement," *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer,* vol. 25, no. 2, pp. 275-286, 2026.
[42] S. Siet, S. Peng, I. Sadriddinov, and K. Park, "Personalized Recommendation System Using Deep Learning with Bayesian Personalized Ranking," *Computers, Materials and Continua,* vol. 86, no. 3, 2026.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)