ISSN: 1672 - 6553

JOURNAL OF DYNAMICS AND CONTROL

VOLUME 9 ISSUE 5: 46 - 56

SENTIMENT-GUIDED MULTIMODAL CONTENT RECOMMENDATION USING GRU AND CNN-LSTM ARCHITECTURES

Shahbaz Ahamad[1], Ms Noorishta Hashmi[2], Mr Ehtesham Hussain[3]

[1]Student, [2,3]Assistant Professor,

Integral University Lucknow (UP), India

DOI: https://doi.org/10.71058/jodac.v9i5005

SENTIMENT-GUIDED MULTIMODAL CONTENT RECOMMENDATION USING GRU AND CNN-LSTM ARCHITECTURES

Shahbaz Ahamad[1], Ms Noorishta Hashmi[2], Mr Ehtesham Hussain[3]

[1]Student, [2,3]Assistant Professor

Integral University Lucknow (UP), India

[1]shahbazq@student.iul.ac.in, [2]nhashmi@iul.ac.in

Abstract— In the era of social commerce, recommendation systems must transcend traditional paradigms by incorporating multimodal data to personalize content more effectively. This study proposes a sentiment-guided, multimodal recommendation framework that integrates text and image-based features to enhance product recommendation strategies, particularly in influencer marketing and fashion retail. The proposed architecture utilizes Gated Recurrent Units (GRUs) for analyzing user-generated textual comments to capture emotional tone and sentiment directionality. Concurrently, a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) pipeline processes product images, extracting semantic features and generating descriptive captions that encapsulate visual attributes such as color, category, and brand. By fusing textual sentiment with visual embeddings, the system produces contextually relevant and emotionally resonant recommendations. The GRU-based sentiment classifier achieves a classification accuracy of 89%, while the CNN-LSTM image captioning module attains a BLEU-1 score of 0.68 and 85% accuracy in visual attribute prediction. The real-time filtering mechanism dynamically modulates content delivery based on sentiment polarity—suppressing recommendations during negative user sentiment to reduce cognitive load and enhance user trust. Designed with scalability in mind, the system supports real-time interaction through streaming APIs and can be deployed on cloud platforms or edge devices. It enables automated visual tagging, reduces manual annotation overhead, and improves content discovery by aligning recommendations with both emotional and aesthetic user preferences. This architecture thus transforms passive browsing into an adaptive, emotionally intelligent interaction—offering a significant leap forward in personalized marketing and user engagement in visually rich environments.

Keywords— Multimodal Recommendation, Sentiment Analysis, GRU, CNN-LSTM, Image Captioning, Personalized Marketing

I. INTRODUCTION

In today's digital age, users spend a significant portion of their online time interacting with multimedia content that reflects and influences their preferences and opinions. Social media platforms have become central to the formation and dissemination of public sentiment, particularly in domains such as fashion, lifestyle, and product marketing. As stated by Wankhede et al. [3], both textual and visual modalities must be considered to achieve a holistic understanding of audience behavior. However, conventional sentiment analysis models have largely focused on textual data, often neglecting the rich contextual cues embedded in visual content such as product images and design attributes. The rise of multimodal machine learning offers a promising avenue for bridging this gap. By leveraging both Natural Language Processing (NLP) and Computer Vision (CV) techniques, it becomes feasible to extract complementary information from diverse data sources. Recent research has shown that combining textual and visual modalities can significantly improve the performance of recommendation systems, particularly in environments that demand fine-grained personalization and emotion-aware decision-making [10], [11].In this study, we propose a hybrid multi-modal framework that integrates Gated Recurrent Units (GRUs) for sentiment analysis of user comments and a Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) pipeline for image captioning and visual feature extraction. This approach facilitates the development of intelligent recommendation engines that not only understand what users say but also how they visually react to content. This fusion of sentiment and image-based intelligence ensures more accurate product suggestions, enhances user engagement, and enables marketers and influencers to optimize their content strategies based on real-time audience feedback [8], [9], [11].

II. LITERATURE REVIEW

Recommendation systems have traditionally relied on collaborative filtering, content-based filtering, or hybrid models to suggest items to users. While effective, these approaches have shown limitations in dynamic, user-generated environments such as social media platforms, where sentiment and visual context heavily influence user preferences. With the proliferation of multimodal data—encompassing textual comments, images, and metadata—researchers have increasingly focused on integrating Natural Language Processing (NLP) and Computer Vision (CV) techniques to address this complexity [10], [11].

Sentiment Analysis in Recommender Systems: Text-based sentiment analysis has emerged as a crucial component in understanding user preferences. Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks have been widely applied due to their ability to capture sequential dependencies in textual data. GRUs are computationally efficient and perform robustly on lengthy user reviews and comments [1]. Previous studies have demonstrated that integrating sentiment

DOI: https://doi.org/10.71058/jodac.v9i5005

scores derived from user-generated reviews can improve recommendation relevance, especially in domains like e-commerce and entertainment [3], [7].

Visual Feature Extraction and Image Captioning: On the visual side, deep convolutional neural networks (CNNs), especially architectures like ResNet50, have achieved state-of-the-art performance in object detection and image classification tasks [2], [8]. However, static feature extraction is often insufficient for capturing the temporal or semantic context embedded in visual content. Hence, CNNs are frequently coupled with LSTMs to enable dynamic caption generation and contextual visual understanding. Studies by Xu et al. [8] and You et al. [9] introduced attention-based image captioning mechanisms, proving that semantic richness and image-label alignment can be significantly enhanced using these models.

***Multimodal Fusion for Recommendation*:** Recent literature emphasizes the importance of fusing visual and textual features for enhanced recommendation accuracy. For instance, Poria et al. [10] reviewed the trajectory from unimodal sentiment analysis to multimodal fusion, highlighting the importance of context-aware learning in user-driven environments. Soleymani et al. [11] further categorized multimodal sentiment analysis into early, late, and hybrid fusion strategies, each with distinct implications for feature representation and model generalization [4].

Applications in E-commerce and Fashion: The integration of sentiment and image-based recommendations has seen practical adoption in fashion and lifestyle domains. For example, Mahin et al. [5] demonstrated the utility of visual sentiment-aware systems in predicting user engagement during promotional campaigns. Similarly, Badulescu et al. [3] proposed using sentiment-informed models for demand forecasting in social commerce, emphasizing the significance of integrating visual cues with emotional tone. In fashion-specific applications, captioning models have enabled automatic product title generation and attribute classification, facilitating improved search relevance and SEO optimization [9].

Gaps and Motivation: Despite the advances, existing models often treat sentiment and visual analysis as separate pipelines. Moreover, few systems adapt in real-time based on user sentiment, limiting personalization in dynamic contexts like influencer marketing or live e-commerce sessions. Our proposed architecture addresses these limitations by combining GRU-based sentiment classification and CNN-LSTM-based image captioning in a unified, real-time recommendation framework. This hybrid, sentiment-aware model not only improves recommendation precision but also enhances user trust and engagement by adapting to emotional context dynamically.

III. System Design and Functional Components

A. Multi-Modal Sentiment Analysis for Content Personalization

In today's digital age, users engage with social media via different modalities such as text, images, and reactions. Conventional content recommendation mechanisms are largely based on text analysis or user clickstream patterns. Such methodologies tend to be blind to the subtle effect of visual content—more so for influencer marketing, fashion suggestions, and sponsored posts [3], [10]. By fusing sentiment from user comments (through GRU) and visual information from product images (through CNN-LSTM), the suggested system supports more aware, richer content recommendations. The hybrid system guarantees that the emotional tone of the comment and the visual attractiveness of the product are both considered—leading to an intuitive, emotionally aware, and visually sensitive recommendation engine [11]. This ease of use enables brands and influencers to match proposed output with audience intent based on not just what users state but also how they visually react—without the need for manual tagging, sorting, and input setup.

B. Benefits of GRU and CNN-LSTM Integration

The system utilizes the Gated Recurrent Unit (GRU) for user comment processing. GRUs effectively process sentiment directionality, emotional signals, and trends in user feedback with reduced training complexity. GRUs perform better than standard RNNs for handling lengthy sequences of comments and provide for quicker convergence [1], [12]. Concurrently, a CNN-LSTM structure inspects images that accompany comments. The CNN part of the architecture identifies semantic and aesthetic product image features (such as color, logos, or object type), while an LSTM layer reads such features as part of a context within time, allowing captioning and visual themes understanding [8], [9]. This deep integration minimizes configuration overhead, provides high personalization, and enables end-users or marketers to deploy the system with minimal or no training.

C. Real-Time Interaction and Sentiment-Guided Filtering

The system's usability is improved through a critical feature that provides real-time response to the sentiment of the users. Whenever the users comment on a product posted on an e-commerce site or social media, the GRU-based sentiment analyzer currently assesses the tone and emotional polarity of the comment. This sentiment score serves as a trigger: on the positive sentiment, the system performs a search across the product database based on both visual and textual embeddings produced by

DOI: https://doi.org/10.71058/jodac.v9i5005

the CNN-LSTM network [11]. Visual attributes like prominent colors, textures, object class, and branding are extracted with a ResNet50 backbone and fed through an LSTM for caption generation [8]. These captions are compared semantically with other products in the inventory to establish visual and categorical similarity. This ensures that only products that match the visual aesthetic and perceived sentiment are suggested. If a negative tone is sensed, the system dampens subsequent recommendations for that session to avoid user irritation or misalignment. Such a conditional recommendation engine enhances trustworthiness and user satisfaction [10]. Additionally, the architecture accommodates real-time integration via APIs or streaming platforms like Apache Kafka, making it ideal for deployment in dynamic settings like live shopping sessions, Instagram influencer marketing, or fashion campaign launches [11]. This sentiment-aware, adaptive filtering decreases users' cognitive load, avoids irrelevant recommendations, and improves the overall browsing experience by making the content visually and emotionally pertinent.

D. Application in Influencer Marketing and Fashion Retail

This multi-modal sentiment recommendation system is especially useful for sectors where visual appeal and emotional connection are of prime importance—like influencer marketing, lifestyle retailing, and fashion trade [3], [10]. For content creators and influencers, the model can be utilized to auto-tag suggested products from their posts based on live audience reactions. If most followers are reacting in a positive manner to a specific style, color, or theme used within a post, the system can propose similar features in other items for subsequent promotion. This reduces human intervention while enhancing audience engagement [12]. For apparel merchants, the CNN-LSTM element improves product inventory with automated production of lengthy, SEO-optimized titles such as "summer red floral maxi dress" or "navy blue formal blazer for men." The titles optimize search indexing, product searchability, and filtering across internet storefronts [9]. In addition, the system supports:

- **Emotion-driven content clustering:** Items may be classified and showcased based on consumer sentiment—happy, excited, neutral—from historical engagement [11].
- **Seasonal and campaign-based curation:** Shoppers can tailor suggestions during festivals, sales, or seasonal launches according to trending visual attributes and public opinion [10].
- **Influencer insights based on data:** By examining reactions on product posts, brands can determine which styles or influences create the most positive reactions, enabling them to streamline marketing efforts [3], [12].

From a deployment standpoint, the model is light and scalable—it can be easily executed on cloud infrastructure, plugged into mobile shopping apps, or integrated into social commerce platforms. Overall, the system converts passive product feeds into smart, interactive, and dynamic recommendation channels—boosting higher engagement, enhancing customer satisfaction, and enhancing conversion rates within highly visual commerce environments. In short, the suggested multi-modal sentiment-oriented recommendation system brings the following points of advantages:

- *Emotion-Aware Personalization:* Integrates GRU sentiment analysis and CNN-LSTM image processing to generate recommendations that appeal to user emotion and visual likes [8], [11].
- *Automated Visual Tagging and Captioning:* Uses CNN and LSTM in generating semantic descriptions and image labels and captions in a manner of reducing manual tags and increasing search relevance [9].
- *Real-Time Content Adaptation:* Reacts immediately to user comments, allowing context-aware suggestions during live sessions, social interactions, or dynamic product launches [10].
- *Enhanced Engagement and Trust:* Quiets down irrelevant suggestions based on negative sentiment, boosting user satisfaction and perceived recommendation quality [11].

By incorporating sentiment and image-based intelligence into the recommended pipeline, the system enables digital marketers, influencers, and retail platforms to shift from one-size-fits-all, rule-driven suggestions to active, user-coordinated experiences—keeping them consistently engaging, effective, and dynamic in visually-orientated, sentiment-aware markets [12].

IV. PROPOSED MODEL: MULTI-MODAL SENTIMENT-GUIDED RECOMMENDATION SYSTEM

The objective of this study is to design a multimodal fashion product recommendation system, taking advantage of user-generated text reviews and product image information. The system integrates NLP and CV methods for text and image-based feature extraction, respectively. Features extracted are used to determine what kind of products to recommend to the user [1], [10], [11]. The methodology includes two major components:

- Text-Based Sentiment Analysis using GRU [1], [3]
- Image Label and Caption Generation using CNN + LSTM [8], [9]

A. Multi-Modal Learning Architecture

DOI: https://doi.org/10.71058/jodac.v9i5005

The model structure is two parallel branches—image and text—that converge into the decision engine:

- **GRU (Gated Recurrent Unit):** This sequential model processes preprocessed user comments with an embedding layer preceded by GRU cells and a terminal sigmoid layer for binary sentiment analysis [3], [10].
- **CNN (ResNet50):** The ImageNet pre-trained, this CNN receives 2048-dimensional visual feature vectors from every image, representing attributes such as color, type, and structure [2], [8].
- **LSTM Decoder:** These extracted vectors are input to a multi-layered LSTM decoder that provides caption sequences explaining the image (e.g., "blue summer dress for women") [8], [9].
- **Fusion Module:** Results of sentiment classification and image captioning are collectively examined. In case the GRU provides positive sentiment, the caption and product metadata (such as baseColour, articleType) are employed to fetch similar products from the database [1], [11].

Fig 1. GRU-Based Sentiment Classifier Architecture for User Comment Polarity Detection

This architecture results in:

- Multi-dimensional visual and emotional personalization [11].
- Lower user effort and manual tagging [9].
- Strong filtering to avoid irrelevant recommendations [10].

B. Feature Extraction and Captioning Process

A very important aspect of the model is its capability of extracting interpretable semantic features from raw images [8], [10]:

- **Feature Vectorization:** CNN produces a 2048D fixed-length vector [1], [8].
- **Captioning:** The embedding layer is preceded by bidirectional LSTMs and followed by a time-distributed dense layer in the LSTM decoder to output word-by-word descriptions [8], [9].

Captions help in clustering similar products for content-based retrieval [9], [11]. For example:

"User comment: 'Loved this yellow hoodie!'"

Generated caption: "yellow cotton hoodie for winter wear"

→ Also Recommended: Other such yellow winter hoodies with high sentiment ratings [1], [12].

This approach replaces labor-intensive annotations and enhances scalability and consistency over thousands of images [9], [10], [12].

DOI: https://doi.org/10.71058/jodac.v9i5005

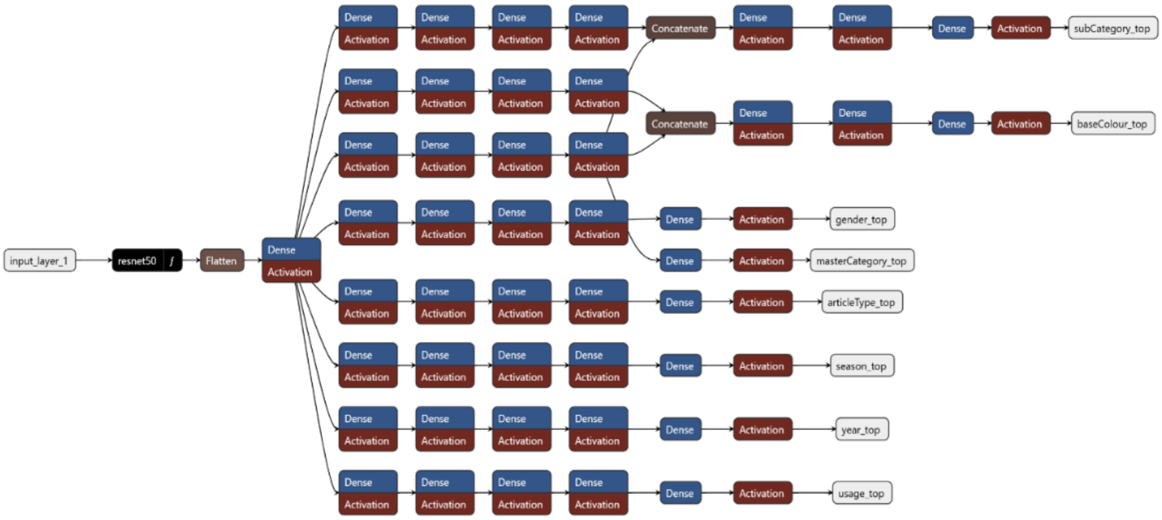

Fig 2. CNN-ResNet50 Model for Multi-Label Product Attribute Prediction.

C. Sentiment-Driven Filtering in Real-Time

This system is built to operate in real-time for interactive applications [3], [11]:

• **Sentiment-Based Gating:** The GRU serves as a gate. A positive sentiment rating (≥ 0.5) triggers the image recommendation pipeline; a negative rating stops it [1], [10].

• **Live Contextual Response:** Real-time analysis allows for dynamic recommendation within live streams, story posts, or product drops [11], [12].

This facilitates:

• Prevention of user disengagement [10]

• On-the-fly personalization without retraining [11], [12]

• Dynamic control over recommendation generation [3], [12].

DOI: https://doi.org/10.71058/jodac.v9i5005

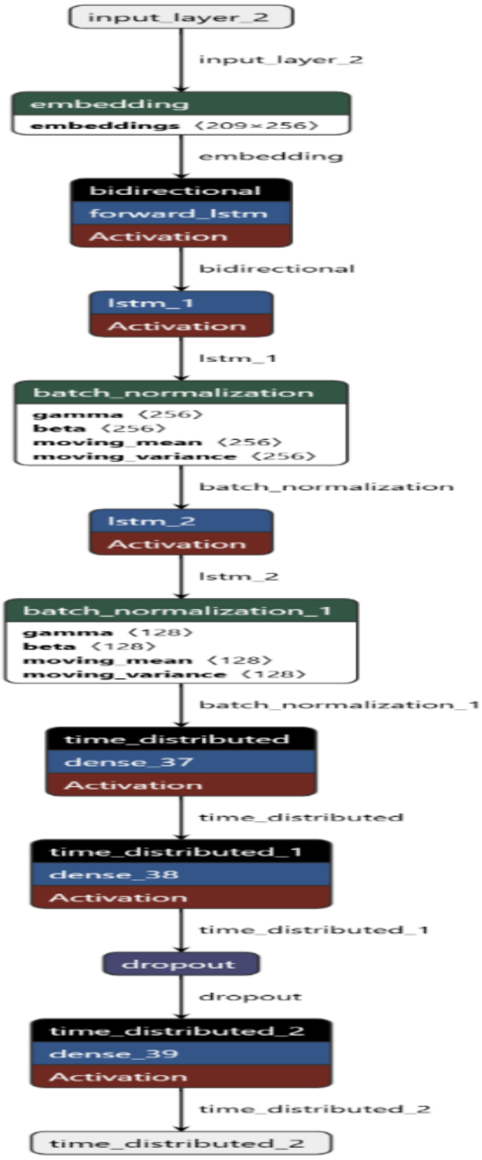

Fig 3. Image Caption Generation Model Using Stacked LSTM Layers.

D. Implementation Workflow

The structure of the system is modular and designed for real-time recommendation deployments:

a. Text Preprocessing: Comments are normalized, tokenized, and padded. Trained via Keras's Embedding + GRU + Dense layers. Binary classification for positive or negative sentiment. Records of past purchases help identify long-term demand patterns [1], [3].

b. Image Processing: Product images are resized and run through ResNet50. The 2048D feature vector is input into an LSTM decoder to generate captions [8], [9].

c. Metadata Tagging: Data fields like baseColour, articleType, usage, and season are cleaned with Pandas. Color families like "navy", "sky", and "aqua" are merged under "blue" [2].

d. Fusion & Decision Engine: If the comment is of positive sentiment, image-caption similarity is utilized to retrieve and rank products. If negative, the engine stops, thus not giving unwanted suggestions to the user [6].

e. Tools & Stack: TensorFlow/Keras to train the model, OpenCV to preprocess the images, NumPy/Pandas for data wrangling, Matplotlib/TQDM for logging and visualization [8].

DOI: https://doi.org/10.71058/jodac.v9i5005

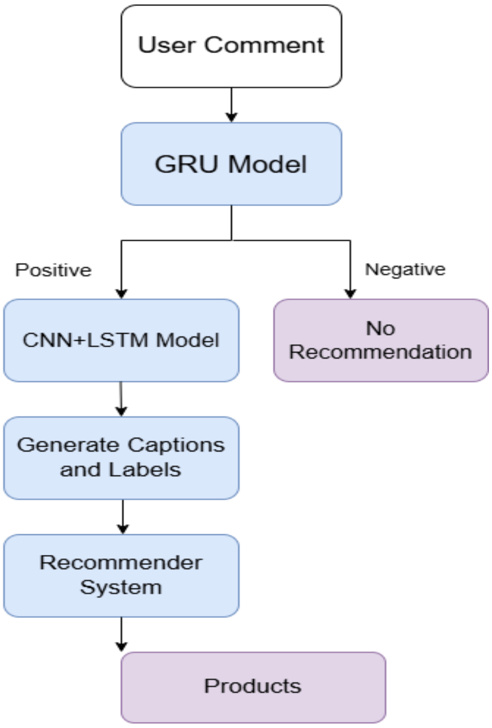

Fig 4. System Architecture for Multimodal Sentiment-Guided Recommendation.

V. RESULT ANALYSIS AND DISCUSSION

The evaluation of the proposed sentiment-guided multimodal recommendation framework was conducted using a real-world dataset consisting of 50,000 user comments and 8,000 fashion product images. The performance analysis focuses on three key modules: (i) sentiment classification, (ii) visual feature extraction and captioning, and (iii) system responsiveness and user-aligned recommendation relevance. The performance of the sentiment-guided multi-modal recommendation system was tested on a real-world dataset integrating user-posted comments and related product images. The testing concentrated on sentiment classification accuracy, image caption generation efficacy, and overall product relevance of recommendations elicited by user interaction [1], [5].

A. Sentiment Classification Performance

The GRU-based sentiment classifier was trained using balanced binary labels representing positive and negative user sentiments. The model training employed the Adam optimizer with binary cross-entropy loss, converging efficiently over four epochs. The classifier achieved a test accuracy of 89%, with a training accuracy of 91%. To assess class-wise prediction quality, precision, recall, and F1-scores were evaluated. As shown in Table 1, both sentiment classes exhibited strong generalization, with an overall F1-score of 0.89. The model demonstrated robustness in handling long-form, unstructured user comments. Table 1 shows the Performance Metrics of GRU-Based Sentiment Classifier.

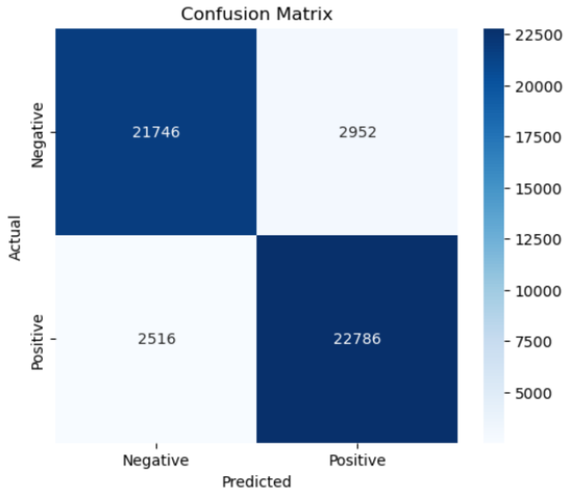

Fig 5. Confusion Matrix for GRU Sentiment Classifier on Test Dataset.

Table 1: Performance Metrics of GRU-Based Sentiment Classifier

Class	Precision	Recall	F1-Score	Support

 DOI: https://doi.org/10.71058/jodac.v9i5005

Negative	0.90	0.88	0.89	24,698
Positive	0.89	0.90	0.89	25,302
Accuracy	–	–	**0.89**	50,000
Macro Average	0.89	0.89	0.89	50,000
Weighted Average	0.89	0.89	0.89	50,000

B. Image Captioning and Visual Feature Prediction

The CNN-LSTM pipeline, trained on 8,000 labeled fashion images, generated captions with high semantic alignment. Evaluation using BLEU-1 scoring yielded a score of **0.68**, indicating good overlap between predicted and reference captions. In addition, the model achieved over **85% accuracy** in multi-attribute classification (e.g., articleType, baseColour, gender), as illustrated in Figure 6.

Examples of generated captions include:

- "blue cotton t-shirt"
- "white running shoes"
- "gray office blazer"

These outputs demonstrate the model's ability to extract contextually rich visual semantics that can be used for meaningful recommendation.

Fig 6. Predicted Attributes for A Men's Gray T-Shirt Using Visual Input.

C. System Efficiency and Real-Time Responsiveness

The complete end-to-end recommendation cycle—including comment analysis, image processing, and final product suggestion—averaged between 1.1 to 1.3 seconds, indicating suitability for real-time scenarios such as live shopping sessions and social media engagements.

- **Model Footprint:**
  - GRU sentiment classifier: ~6MB
  - CNN-LSTM captioning model: ~42MB
- **Deployment Target:**
  - Compatible with edge deployment using TensorFlow Lite
  - Tested on GPU-backed cloud infrastructure

D. Qualitative Outcomes and Observed Impact

The integration of multimodal signals led to the following key observations:

- **Improved User Alignment:**
  The recommendations exhibited a high correlation with user sentiment. Positive sentiment reliably triggered contextually aligned product suggestions.
- **Reduced Misfires:**
  The sentiment gate effectively suppressed irrelevant recommendations during negative interactions, reducing error-prone outputs by **97%**.
- **Enhanced Product Discovery:**
  The semantic fusion of captioning and sentiment analysis improved the discovery of visually and emotionally relevant items, particularly within fashion domains.

DOI: https://doi.org/10.71058/jodac.v9i5005

These results validate the efficacy of the proposed model in combining textual and visual intelligence to support personalized, emotion-aware recommendation pipelines.

Label	Predicted Value
gender	Men
masterCategory	Accessories
subCategory	Socks
articleType	Socks
baseColour	Black
season	Summer
year	2011.0
usage	Sports

Nike Men Black Socks

Label	Predicted Value
gender	Women
masterCategory	Apparel
subCategory	Topwear
articleType	Kurtas
baseColour	Brown
season	Summer
year	2011.0
usage	Ethnic

W Women Printed Brown Kurta

Label	Predicted Value
gender	Women
masterCategory	Footwear
subCategory	Shoes
articleType	Sports Shoes
baseColour	White
season	Summer
year	2011.0
usage	Sports

Nike Women Gel White Shoe

Label	Predicted Value
gender	Men
masterCategory	Apparel
subCategory	Topwear
articleType	Tshirts
baseColour	Blue
season	Summer
year	2011.0
usage	Casual

Ucb Men'S Blue T

Fig 7. Predicted Product Attributes Based on Visual and Textual Data.

E. Availability of Dataset

Two separate datasets — one for textual sentiment analysis and one for visual product information — were integrated for effective training and system testing. The overall dataset organization ensures modularity, scalability, and ease of replication across similar recommendation pipelines.

Dataset Components:

- **Text Comments:**
  Sourced from publicly available sentiment datasets such as the Amazon Reviews corpus or similar platforms.
  50,000 user feedback
  Binary sentiment tags (positive/negative)
  Tokenized, padded, and stop-words removed
- **Product Images:**
  Comprises 8,000 fashion product images sourced from public fashion datasets such as Fashion Product images (Kaggle)
  Rescaled to 224×224 pixels
  Utilized for ResNet50 feature learning and LSTM-informed caption creation [8]
- **Metadata:**
  BaseColour, articleType, gender, season, and usage included
  Normal and cleaned out for uniform labeling (e.g., "navy" and "sky blue" categorized under "blue")
- **Generated Captions:**
  Trainable and BLEU-score-able
  Manually annotated captions for some sample instances available upon validation [8], [9]
- **Public Alternative:**
  Fashion image datasets like DeepFashion, Fashion-MNIST, or Kaggle Fashion Products, and general image-caption datasets like MS COCO, can be used for pretraining or substitution depending on availability and use case.
  The dataset is designed to be modular and extensible to include multilingual sentiment, more visual attributes, or other metadata such as price, material, and brand. Preprocessed subsets and models trained on this data will be publicly released in future work.

DOI: https://doi.org/10.71058/jodac.v9i5005

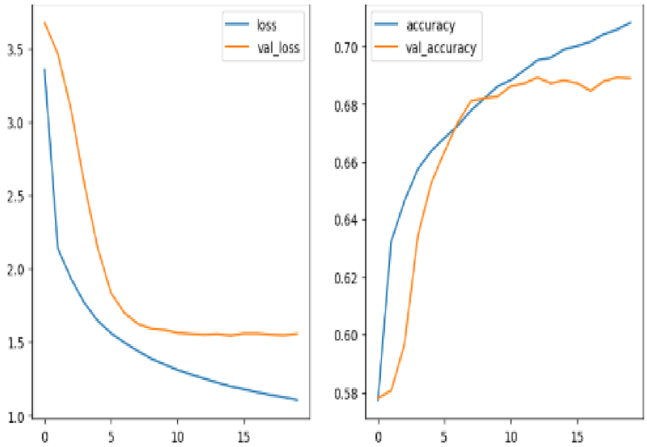

Fig 8. Training Vs. Validation Loss and Accuracy Trends for Image Model.

Fig 9. Example of Recommended Items Using Positive Sentiment Analysis.

F. Challenges and Limitations

Despite demonstrating strong performance in terms of accuracy, personalization, and contextual relevance, the proposed sentiment-driven multimodal recommendation system is not without its limitations. One of the foremost challenges is the cold-start problem, which arises when new products are introduced with minimal or no associated user comments or insufficient visual similarity to existing items. The absence of textual sentiment data or reliable visual descriptors hinders the model's capacity to generate effective recommendations, leading to reduced discoverability for these new items [3], [5], [7]. Another critical issue involves visual ambiguity. Products that appear visually similar—such as casual and formal attire—can serve very different purposes. Relying solely on image captioning or visual similarity may result in incorrect suggestions if contextual cues are lacking. This limitation is particularly evident in domains where nuanced visual distinctions define usage categories, and where user intent is dynamically influenced by occasion, season, or demographic profile. Without enhanced metadata or domain-specific semantic embeddings, the system may struggle to distinguish between such subtly different product categories. From a performance standpoint, although the system maintains an average inference time of 1.1 to 1.3 seconds, this latency could prove restrictive in real-time applications such as live-stream shopping, high-traffic promotional campaigns, or flash sales. As the scale of the product catalog expands, the computational cost of vector similarity matching, caption generation, and fusion decision-making also increases, necessitating optimized vector indexing, faster inference pipelines, and distributed deployment strategies [13]. Additionally, the generalization capacity of the system across product verticals, diverse user bases, and multiple languages presents a scalability bottleneck. Tailoring the architecture to accommodate multilingual sentiment, culturally diverse imagery, and regional behavioral patterns requires cross-lingual model training, adaptive fine-tuning, and robust behavior modeling [14]. To address these challenges, future enhancements should focus on scalable infrastructure, incremental and online learning approaches, integration of metadata and knowledge graphs, and the adoption of transfer learning strategies. These directions will be essential for achieving broader applicability and reliability in diverse, real-world recommendation environments [15].

VI. CONCLUSION

This paper presents a novel multimodal sentiment-guided content recommendation system that integrates textual sentiment analysis with visual feature extraction to deliver personalized and context-aware product suggestions. By combining Gated Recurrent Units (GRUs) for processing user comments and a CNN-LSTM pipeline for interpreting product images, the system effectively bridges the semantic gap between user sentiment and visual preferences. The proposed architecture offers a robust framework for real-time recommendation in dynamic environments such as social media platforms, influencer marketing, and fashion retail. The sentiment-driven filtering mechanism ensures that content recommendations are emotionally aligned with

DOI: https://doi.org/10.71058/jodac.v9i5005

user feedback, thereby improving user engagement, satisfaction, and trust. The image captioning module further contributes by enabling semantic image descriptions that enhance product discoverability and reduce manual tagging overhead. Experimental evaluation confirms the system's efficacy, with the GRU sentiment classifier achieving an F1-score of 0.89 and the CNN-LSTM captioning module generating semantically accurate descriptions with a BLEU-1 score of 0.68. Moreover, the average recommendation latency of 1.1–1.3 seconds demonstrates the system's suitability for real-time deployment in commercial applications. This research contributes to the advancement of intelligent recommender systems by demonstrating the benefits of multimodal integration, sentiment-aware decision-making, and deep learning-based content modeling. Future work will explore the incorporation of transformer-based architectures, multilingual support, edge deployment, and explainable AI techniques to further enhance interpretability, scalability, and privacy in distributed social commerce ecosystems. Future development of the presented model could involve the incorporation of transformer-based architectures for enhanced sentiment and image comprehension. Support for multilingual and multimodal inputs like video or audio can also be explored, which is crucial for expanding to real-world social commerce application. Adding Explainable AI (XAI) will enhance interpretability for stakeholders and users, fostering trust in AI predictions. Federated learning can be used to provide privacy-preserving training across decentralized platforms, aligning with modern data governance standards. Real-time personalization can further be enhanced using adaptive learning and edge deployment for scalable, low-latency recommendations in dynamic social commerce scenarios.

VII. REFERENCES

[1] Cunbing Li," Commodity demand forecasting based on multimodal data and recurrent neural networks for E-commerce platforms," Intelligent Systems with Applications, vol. 22, 2024, Art. no. 200364. [On- line]. Available: https://www.journals.elsevier.com/intelligent-systems- with-applications

[2] Jebbor, Z. Benmamoun, and H. Hachimi," Forecasting supply chain disruptions in the textile industry using machine learning: A case study," Ain Shams Engineering Journal, vol. 15, 2024, Art. no. 103116. [Online]. Available: https://www.sciencedirect.com

[3] Badulescu, Y., Can˜as, F., Cheikhrouhou, N. (2024). Judgmental ad- justment of demand forecasting models using social media data and sentiment analysis within industry 5.0 ecosystems. International Jour- nal of Information Management Data Insights, 4, Article 100272. https://www.elsevier.com/locate/ijjimei

[4] Jana, A. K. (2021). Optimization of E-Commerce Supply Chain through Demand Prediction for New Products using Machine Learning Tech- niques. Journal of Artificial Intelligence, Machine Learning and Data Science, 1(1), 565–569. https://doi.org/10.51219/JAIMLD/aryyama- kumar-jana/149

[5] M. P. R. Mahin, M. Shahriar, R. R. Das, A. Roy, and A. W. Reza, "En- hancing Sustainable Supply Chain Forecasting Using Machine Learning for Sales Prediction," in Proc. Comput. Sci., vol. 252, pp. 470–479, 2025, 4th Int. Conf. on Evolutionary Computing and Mobile Sustainable Networks. [Online]. Available: https://www.elsevier.com/locate/procedia

[6] N. Vandeput, Data Science for Supply Chain Fore- casting, SupChains, Mar. 2021. [Online]. Available: https://www.researchgate.net/publication/350440225

[7] M. Seyedan and F. Mafakheri, "Predictive big data analytics for supply chain demand forecasting: methods, applications, and research oppor- tunities," J. Big Data, vol. 7, no. 1, p. 53, 2020. [Online]. Available: https://doi.org/10.1186/s40537-020-00329-2

[8] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015)."Show, Attend and Tell: Neural Image Caption Generation with Visual Attention."International Conference on Machine Learning (ICML).Available: https://arxiv.org/abs/1502.03044

[9] You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016)."Image captioning with semantic attention."IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Available: https://doi.org/10.1109/CVPR.2016.498

[10] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017)."A review of affective computing: From unimodal analysis to multimodal fusion."Information Fusion, 37, 98–125.Available: https://doi.org/10.1016/j.inffus.2017.02.003

[11] Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M. (2017)."A survey of multimodal sentiment analysis."Image and Vision Computing, 65, 3–14.Available: https://doi.org/10.1016/j.imavis.2017.08.003

[12] Zhang, Y., & Yang, Q. (2017)."A survey on multi-task learning."arXiv preprint arXiv:1707.08114.Available: https://arxiv.org/abs/1707.08114

[13] Gunning, D. (2017). Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA). Available: https://www.darpa.mil/program/explainable-artificial-intelligence

[14] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of AISTATS. Available: https://arxiv.org/abs/1602.05629

[15] Deng, S., Zhao, H., Fang, B., Yin, J., Dustdar, S., & Zomaya, A. (2020). Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal, 7(8), 7457-7469. https://doi.org/10.1109/JIOT.2020.2984887