


Empowering Retail Dual Transformer-Based Profound Product Recommendation Using Multi-Model Review

Deema mohammed Alsekait
Princess Nourah bint Abdulrahman University, Saudi Arabia

Hanaa Fathi
Applied Science Private University, Jordan

Mohamed Taha
 <https://orcid.org/0000-0003-0885-0985>
Benha University, Egypt

Ahmed Taha
Benha University, Egypt


Ayman Nabil
Misr International University, Egypt

Asif Nawaz
PMAS-Arid Agriculture University, Pakistan

Zohair Ahmed
Islamic University, Pakistan

Mohammad Alshinwan
Applied Science Private University, Jordan

Mohamed F. Issa
University of Pannonia, Hungary

Diaa Salama AbdElminaam
 <https://orcid.org/0000-0003-0881-3164>
Jadara Research Center, Egypt

ABSTRACT

Advancements in technology have significantly changed how we interact on social media platforms, where reviews and comments heavily influence consumer decisions. Traditionally, opinion mining has focused on textual data, overlooking the valuable insights present in customer-uploaded images—a concept we term Multus-Medium. This paper introduces a multimodal strategy for product recommendations that utilizes both text and image data. The proposed approach involves data collection, preprocessing, and sentiment analysis using Vti for images and SpanBERT for text reviews. These outputs are then fused to generate a final recommendation. The proposed model demonstrates superior performance, achieving 91.55% accuracy on the Amazon dataset and 90.89% on the Kaggle dataset. These compelling findings underscore the potential of our approach, offering a comprehensive and precise method for opinion mining in the era of social media-driven product reviews, ultimately aiding consumers in making informed purchasing decisions.

KEYWORDS

Multus-Medium Reviews, Recommendation, Sentiment Score, SpanBERT, Fusion, Vti Transformer

INTRODUCTION

Electronic commerce, commonly known as e-commerce, refers to the buying and selling of goods and services via the internet. This digital evolution allows consumers to shop from the convenience of their homes or offices at any time, enhancing accessibility and flexibility. E-commerce benefits businesses by enabling them to offer round-the-clock service, which often leads to increased transactions and customer satisfaction. Platforms such as review sites, online shopping portals, and blogs empower nearly everyone to share their views on products and services (Kanwal et al., 2024). A 2024 survey by Adobe Digital Economy Index found that 85% of online shoppers now conduct

DOI: 10.4018/JOEUC.358002

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Figure 1. Textual review example with customer ratings



extensive research online before making a purchase decision, highlighting the influential role of the internet in consumer behavior (Kumar et al., 2024).

Moreover, the emotional range expressed in online reviews—from positive to negative sentiments—can profoundly impact consumer choices and corporate reputations. These reviews not only inform potential buyers about the quality of products and services but also play a crucial role in shaping a company’s image and potentially its revenues. Positive reviews can boost consumer trust and attract more customers, while negative feedback can harm a company’s credibility and deter potential buyers (Alslaity et al., 2024). In today’s competitive market, with a plethora of options available, customers highly value the experiences shared by others. Managing and responding to customer feedback is essential for survival in the competitive realm of e-commerce. For instance, if someone is considering buying headphones, they are likely to consult online reviews to gauge the opinions of past customers, which will significantly influence their purchasing decision.

Reviews can generally be classified into two main types: quantitative, such as star ratings, and qualitative, such as lists of pros and cons, as illustrated in Figure 1. These reviews encompass annotations in the form of words or images that convey the reviewer’s sentiments. Quantitative measures, notably star ratings, are commonly used on many online retail sites to gauge consumer opinions. Classification involves organizing different types of reviews or feedback into distinct categories (Singh et al., 2024). Reviews can be emotionally categorized in various ways, including positive, negative, neutral, and these sentiments can be broken down into other specific emotions such as anger, sadness, happiness, and annoyance.

Consequently, consumers increasingly rely on product reviews to gather information and make well-informed purchasing decisions (Mei et al., 2023). However, with the overwhelming amount of available information, approximately 32% of consumers report feeling confused and 30% express frustration. Despite this, many still prefer to read through several reviews before choosing a product. To mitigate these challenges, sentiment analysis (SA) has become an essential tool. It helps extract, evaluate, and present the sentiments of individuals on social media platforms.

SA, a branch of natural language processing, automates the identification of opinions from textual data (Miah et al., 2024). Its primary aim is to classify user-generated reviews as negative or positive based on the expressed sentiment regarding a specific topic. According to Liu ([date]), the terms “opinion mining” and “SA” are interchangeable and involve the study of behaviors, attitudes, feelings, emotions, evaluations, categorizations, opinions, and sentiments related to various aspects of services, products, or individuals.

Various methods of opinion mining are utilized to provide product recommendations, primarily based on textual content provided by consumers. However, with the advent of Multus-Medium, which includes both text and images, consumers can now convey their feedback more comprehensively. As depicted in Figure 2, users express their thoughts, feelings, and emotions by posting both text and photos on social media platforms, such as their experiences with delayed flights.

Figure 2. Multus-medium reviews



When recommending products to new customers, many analysts utilize traditional machine learning techniques and opinion classification methods such as deep multi-modal attentive fusion (DMAF), VGG-Net-16, and attention-based modality-gated networks (AMGN). Multus-Medium reviews, which include both textual and image data, provide a wealth of information which requires careful analysis before processing. Consequently, sophisticated methodologies are employed to analyze customer feedback from Multus-Medium image and text data. This research focuses on the Multus-Medium approach, which extracts features from both images and text, aiding in the effective recommendation of products to customers.

Research Contribution

The key contributions of the proposed research are as follows:

- the formation of Dual sentiment model based on Vti transformer with SpanBERT for better Sentiment score calculation which could be further used for SA tasks;
- a new proposed mechanism for profound product recommendation model which utilizes sentiment score, product usage, and user profiling; and
- the experimental results of the proposed model, which have been compared with baseline approaches, demonstrating superior accuracy, precision, and recall.

The rest of the paper is organized as follows. The next section presents a literature review of the recommendation techniques, and proposed methodology is discussed in the following section. After that, we analyze the data. Finally, we present conclusions, recommendations, and future research guidelines.

LITERATURE REVIEW

In recent years, product recommendations have become crucial for understanding customer feedback before making any purchasing decisions regarding specific products or services. This section explores cutting-edge developments in the field, which are further divided into three categories: text-based reviews, image-based reviews, and Multus-Medium reviews. We discuss each category in detail.

Text Reviews-Based Recommendation

Research in analysis of people's sentiments is being carried out extensively nowadays (Rana et al., 2023). Onan (2020), for example, analyzed panel data of commercial U.S. banks. Their study employed two measures: "Garcia's sentiment measure," which included the analysis of positive or

negative financial news from the *New York Times*, and “news implied volatility,” which is a measure of text uncertainty. Their research underscores the adverse impact of the Great Financial Crisis on the sentiments of the investors, which had an impact on declining lending behavior in the United States. This was a factor in the banking sector becoming instable.

Yang et al. (2020) proposed a deep learning model to categorize people’s sentiments in their reviews from the text. The authors observed that a comprehensive deep learning-based approach incorporating a unified feature set, including embedding of words, knowledge of sentiments, rules of sentiment shift, and linguistic and statistical knowledge was not used for SA. In their work they used recurrent neural networks (RNN) with long short-term memory (LSTM) for sequential processing and took a sentence level sentiment approach to classify the text reviews.

Meena et al. (2018) introduced a deep learning technique for performing SA on product reviews by using a Twitter dataset. Their proposed method used a combination of TF-IDF weighted GloVe word embeddings and a convolutional neural network (CNN)-LSTM architecture. They performed numerous experiments to analyze and compare the performance of various word embedding schemes with traditional deep neural network architectures.

Image Reviews-Based Recommendation

In the field of image SA, the work of Wang et al. (2021) addressed the challenge of recognizing high-level abstractions of image data. They concluded that image and local region information contain significant sentimental information. To achieve this, they proposed a framework which utilized affective regions of images. Their approach involved using an off-the-shelf abjectness tool to generate candidates, followed by a candidate selection method to remove redundant and noisy proposals. Next, they employed a CNN to compute the sentiment scores for each candidate, considering both abjectness and sentiment scores to discover effective regions automatically.

Chaudhry et al. (2021) proposed two-stage deep learning architecture for fashion picture recommendations. They utilized a visually aware feature extractor, which was driven by data, using a neural network classifier. They subsequently used this classifier as input for ranking algorithm-based suggestions which were similarity-based, and they evaluated the proposed work on the fashion dataset, which was publicly available. To enhance the robustness and performance of existing content-based recommendation systems, the authors used their approach in combination with other established systems. This was aimed at better aligning with specific client styles, among other benefits.

Huang et al. (2020) proposed a model that takes both images and its features as inputs to provide users with personalized recommendations. Their model used various image features, such as color and shape, to differentiate images into different categories. They used the mean and standard deviation of image matrices to classify images and calculated the distance matrix between images to distinguish them. Their model had 83% accuracy.

Multis-Medium Reviews-Based Recommendation

El-Affendi et al. (2021) performed SA on the Weibos posts about travelers’ experiences with commercial air travel. Different travel-related occurrences, such as a flight delay, may have an adverse effect on passengers’ moods. They used a multi-task structural arrangement to evaluate events along with sentiment instantaneously by modeling the cross-modal linkages to provide more discriminative representations. The authors extracted features from a prescribed dataset. Numerous tests demonstrated that the suggested procedure outperforms the most recent cutting-edge methods. The accuracy of their proposed model was 91%. Li et al. (2024) presented the multi-level textual-visual alignment and fusion network, which integrates three auxiliary tasks to enhance multi-modal alignment. It first converts multi-level image data into descriptions of images, faces, and optical characters. These descriptions are then combined with textual input, creating a unified textual-visual representation that facilitates comprehensive alignment between the two modalities. This combined input is subsequently processed by an integrated text model which incorporates relevant visual features, enabling more

accurate SA and recommendation outcomes. Similarly, Zulqarnain et al. (2024) proposed a two-stage GRU model based on a feature attention mechanism designed to improve sentiment polarity detection through sequential modeling and word-feature capture. This model includes a pre-feature attention layer which enhances the connection between words within a sentence by focusing on key features, thereby improving sentiment polarity categorization. Goularte et al. (2024) introduced SentPT, a polarity classifier specifically designed for SA of Portuguese-language texts across various genres. This classifier utilizes a transfer learning approach, fine-tuning a [define here] (BERT) model, and it is evaluated on a diverse set of texts, including product reviews, literary works, news articles, and game comments. SentPT demonstrates the effectiveness of transfer learning in adapting to different text genres for SA.

Gastaldo et al. (2013) proposed a mixed-fusion architecture for the image-text SA which leverages the inherent correlation and discriminative properties between visual and semantic contents. Their model utilized two distinct unimodal attention models to classify emotions for the text and image modalities separately. These attention models focus on identifying the essential words and discriminative regions closely associated with the sentiment.

Medhat et al. (2014) proposed an effective model to incorporate extensive social information to enhance the effectiveness of multi-model SA. They utilized a regional attention schema to highlight emotional areas based on the attended channels. To create high-quality representations of social images, they developed a heterogeneous relation network and expanded the graph convolutional network to order content data from social contexts. To evaluate their approach, they performed an experiment on the benchmark datasets of Flickr and Getty. Their model achieved a high accuracy rate of 87% in rating photos from these two datasets.

Wu et al. (2021) introduced a novel approach to SA classification. Their method employed discriminative feature extraction from both images and text. The multi-modal SA was performed by exploiting the correlations between image and text modalities. To achieve this, they utilized a visual-semantic attention paradigm to obtain attended visual aspects for each word, which were further used to develop a semantic self-attention model for automated identification of distinguishing features for sentiment categorization. They tested the approach on a manually annotated dataset without machine labeling. Table 1 offers a summary of other methods addressing SA.

The discussion highlights that much of the existing research tends to concentrate on either text or images, often overlooking Multus reviews. Multus medium reviews, which encompass both text and images, provide richer information that can enhance decision support systems and potentially improve customer satisfaction. This study proposes a Multus-Medium-based recommendation approach aimed at leveraging Multus-Medium for more effective product recommendations.

PROPOSED METHODOLOGY

This section describes in detail the proposed frameworks for product recommendation. Figure 3 shows the architecture of the proposed model which consists of numerous phases to ensure comprehensive and accurate recommendations. The proposed model demonstrates the use of an enhanced vision transformer (ViT) for extracting image features and SpanBERT for generating span-aware text representations from review datasets which include both images and text. The enhanced ViT processes the visual content of the product, while SpanBERT handles the textual data, including SA. The extracted features from both the image and text are then combined in a dense layer, creating item representations. These representations, which incorporate product attributes, descriptions, and sentiment scores, are used for user profiling and similarity calculations. Ultimately, this approach facilitates top-N product recommendations, tailored to the user's preferences and history, enhancing the accuracy and relevance of the recommendations.

Table 1. Summarized literature review

Ref	Proposed Methodology	Result Accuracy	Limitation
[18]	Empirical Analysis	79%	Loses sequential information.
[19]	RNN-LSTM	74.78%	Uses a non-reliable, weight-based method for feature extraction.
[21]	CNN-LSTM	93%	Features may be lost during extraction.
[22]	R-CNN	83.05%	Searching for opinionated words is costly and time-consuming.
[23]	CNN	87%	Struggles with managing implicit features due to a lack of information.
[24]	EHFAO	87.8%	Domain Specific
[25]	Regression	83%	Often dictionary-based and domain-specific.
[26]	Multi-modal event-aware network	91%	Needs a more reliable model.
[27]	Deep learning	76%	Domain specific
[28]	Attention-based heterogeneous relational model (AHRM)	87%	Unable to normalize the polarity intensity of words.
[29]	AMGN	88%	Features may be lost during extraction.
Current Research	Multus-Medium opinion mining (MMOM) approach	91.55% (Amazon), 90.89% (Kaggle)	Introduces an integrated approach for text and image data, surpassing previous methods in accuracy and robustness. Focuses on accuracy; complexity analysis to be explored in future work.

Data Collection

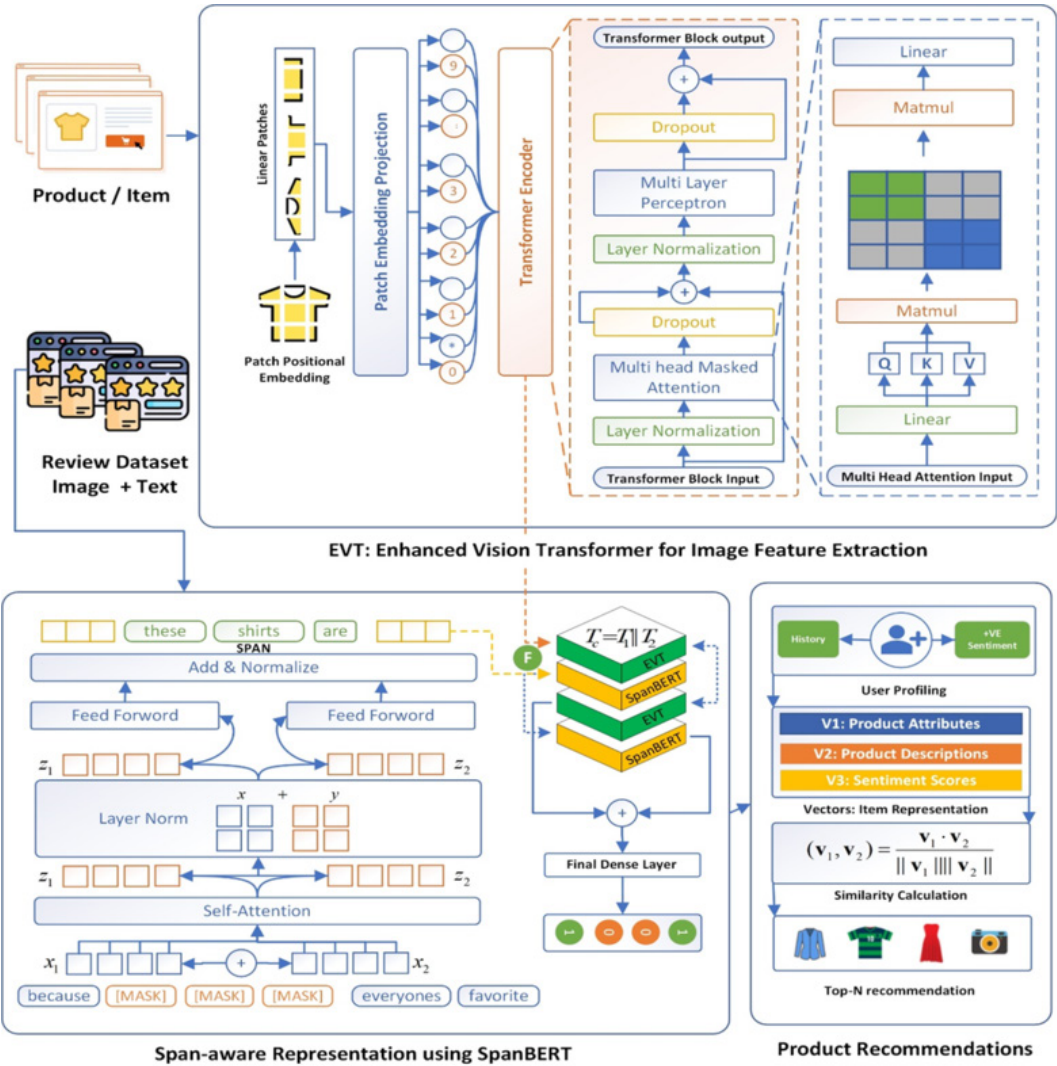
Two large-scale social Multus-Medium datasets, Amazon Multi-Model and Kaggle, have been used in this study (Ghorbanali et al., 2022; Al-Sammarraie et al., 2022). The very first is the updated version of the Amazon review dataset, originally released in 2014, which now contains an expanded collection of 233.1 million reviews, spanning from May 1996 to October 2018. The dataset not only includes reviews with ratings, text, and helpfulness votes but also comprehensive product metadata such as descriptions, categories, price, brand, and image features. Additionally, it includes transaction metadata for each review, detailed attributes such as color, size, and package type, along with product images post-delivery. The dataset also offers enhanced metadata on the product landing pages, including bullet-point descriptions, technical details tables, and similar products tables. Moreover, it introduces five new product categories, broadening its scope and utility. The details are shown in Table 2.

The very next dataset comprises of six distinct sets, each focusing on different product categories or media, with each dataset duplicated across seven unique combinations of image and text encoders, resulting in a total of 42 folders. The folders are named according to the dataset and the specific encoder used for the visual and textual data, for example, “bookcrossing-vit_bert.” The included datasets cover a range of items and media: clothing, shoes, and jewelry (Amazon), home and kitchen (Amazon), musical instruments (Amazon), movies and TV (Amazon), book-crossing, and movielens 25M. Each dataset is encoded with various combinations of visual and text encoders, enhancing the data’s utility for tasks that require integrated analysis of images and textual information.

Table 2. Data collection

Dataset Name	Type	Web link
Amazon Multi-Model	Product Reviews and Rating with categories, price, brand and image features such as color, size and package type	https://nijianmo.github.io/amazon/index.html
Kaggle	Clothing, Shoes and Jewelry (Amazon) Home and Kitchen (Amazon) Musical Instruments (Amazon) Movies and TV (Amazon) Book-Crossing Movielens 25M	https://www.kaggle.com/datasets/ignacioavas/alignmacrid-vae

Figure 3. Proposed dual transformer architecture for product recommendation



Data Preprocessing

Data processing for both textual and visual datasets is thoroughly detailed. For the text dataset, the process involves systematically transforming each review through several stages: tokenization, conversion to lowercase, normalization, stemming, and transliteration. These steps ensure that the raw text is stripped of variability and complexity, resulting in a clean, standardized list of words stored in a preprocessed set, ready for further analysis. For the image dataset, the process begins with resizing the images to a standard dimension, followed by conversion to grayscale or color channel normalization (Ahmed et al., 2021; Rana et al., 2022). The images then undergo noise reduction, typically using filters like Gaussian blur, and contrast enhancement if necessary. More advanced techniques such as edge detection (for example, Sobel operator) and key feature extraction (for example, SIFT or SURF) are applied to optimize the images. The final output consists of the fully processed images, now optimized for improved performance in subsequent analysis or machine learning tasks.

Sentiment Score Calculation

This section delves into the process of calculating the final sentiment score, meticulously outlined in Algorithm 1. This robust calculation unfolds in two distinct phases. Initially, we employ a cutting-edge ViT transformer-based model, specifically designed to derive sentiment scores from image-based reviews. Following this, the second phase leverages the capabilities of a SpanBERT-based model to assess sentiment from textual reviews (Luong et al., 2015; Szegedy et al., 2014). Detailed descriptions of each step in this sophisticated analytical journey are provided in the subsequent subsections.

Computing the Sentiment Score of Image-Based Reviews

ViTs offer significant advantages over traditional image feature extraction methods like CNNs, particularly due to their global context awareness and scalability. Unlike CNNs, which process images through progressively larger local receptive fields, ViTs analyze the entire image at once through the mechanism of self-attention (Simonyan & Zisserman, 2014). This approach allows ViTs to capture intricate dependencies between disparate parts of an image, potentially leading to a more holistic understanding of the scene. Moreover, ViTs are highly scalable; they can be efficiently trained on larger datasets and easily adjusted in size (depth and width) to suit different computational budgets and performance requirements. Additionally, ViTs do not rely on the inductive biases inherent in CNNs, such as translation invariance and locality, which makes them more flexible when learning from diverse and complex image datasets. This flexibility and global perspective make ViTs particularly potent for tasks where context and detailed spatial relationships are crucial, offering a robust alternative to traditional methods in many advanced image recognition and classification tasks.

The first step in a ViT model is to split the image I into fixed-size patches. If the image has dimensions $H \times W$ (height and width) and the color channels C , and if each patch is of size $P \times P$, then the number of patches N is given by:

Multi-headed self-attention allows the model to focus on different parts of the image. For each head kk , the attention scores are computed as follows in Equation 1:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where Q , K , and V are the query, key, and value matrices, respectively, derived from the input embeddings, and d_k is the dimensionality of each head. The outputs from all heads are concatenated and linearly transformed.

Each position passes through the same feed-forward network, as shown in Equation 2:

Algorithm 1. Multi-modal sentiment for product recommendations

1.	Input: $DatasetD = (I_i, T_i, y_i)_{i=1}^N$; An image I_i , text T_i , and sentiment label y_i
2.	Output: y_i Predicted sentiment labels; Product P Top_N
3.	Initialize for each sample (I_i, T_i, y_i) in D :
4.	for $X = 1 \dots do$
5.	Preprocess image I_i to obtain feature representation F_i^I
6.	Tokenize and encode text T_i to obtain token embeddings E_i^T
7.	for each sample (I_i, T_i, y_i) in D :
8.	Extract features from image I_i : $F_i^I = ViT(I_i)$
9.	Extract embeddings from text T_i : $E_i^T = SpanBERT(T_i)$
10.	for each sample (I_i, T_i, y_i) in D :
11.	$\sum_{i=0}^n F_i^I E_i^T \leftarrow multi-modal M_i = [F_i^I, E_i^T]$
12.	Initialize model parameters Θ
13.	for each epoch:
14.	for each sample (M_i, y_i) in D :
15.	Train multi-modal feature M_i and sentiment label y_i
16.	Given a new sample (I_{new}, T_{new}) :
17.	Preprocess image I_{new} and text T_{new} representation M_{new}
18.	Predict I_{new} and T_{new}
19.	for each review (P_i, S_i) in D :
20.	Map product P_i to a feature vector F_i^P representing its attributes
21.	Include sentiment $score S_i$ as an additional feature
22.	Recommendation model M feature vectors F_i^P sentiment scores S_i
23.	Given a new user U_{new} and their preferences $P_{U_{new}}$:
24.	$P (Top_N)$
25.	End

$$FFN(x) = \max(0, xW1 + b1) W2 + b2, \quad (2)$$

where $W1, W2$ are weight matrices and $b1, b2$ are biases.

After processing through the transformer layers, the output corresponding to a special classification token or a pooled representation of all output embeddings is used to predict the class of the image. Typically, a linear layer is followed by a softmax function to output probabilities over classes, as shown in Equation 3:

$$Probability = \text{softmax} (W_{class} h_{cls}). \quad (3)$$

However, for the loss function, cross-entropy loss is used for classification tasks such that, as shown in Equation 4:

$$L = - \sum_{i=1}^c y_i \log(y_i^{\wedge}), \quad (4)$$

where, y_i is the true label vector and y^{\wedge} is the predicted probability vector.

Computing the Sentiment Score of Text-Based Reviews

SpanBERT is especially effective for text review classification due to its focus on understanding and predicting spans of text, rather than individual tokens. This span-based approach during pre-training allows SpanBERT to develop richer contextual embeddings which capture more nuanced relationships within the text. It masks and predicts entire spans, training the model to integrate context over longer sequences, which is crucial in grasping the overall sentiment and subtle expressions in reviews. Additionally, the span-boundary objective enhances the model's comprehension of phrase and sentence boundaries, making it adept at discerning the sentiment influenced by complex linguistic structures. Thus, SpanBERT provides a more contextual and nuanced understanding of text, leading to superior performance in text review classification tasks.

Initially, on text review, padding has been performed to ensure that all sequences in a batch have the same length for matrix operations in deep learning models. In this research, SpanBERT improves upon BERT by focusing its pre-training on spans of text rather than individual tokens. This has been done using two key modifications. Firstly, span is performed pre-training; rather than masking individual tokens as BERT does, SpanBERT masks contiguous spans of text. Mathematically, let $T = t_1, t_2, \dots, t_n$ be a sequence of tokens. A random span t_i, \dots, t_k is masked and the model predicts the entire masked span. This encourages the model to understand contextual relationships over longer sequences of text. Secondly, a span-boundary objective predicts the tokens at the boundaries of a masked span given the context and the content of the span, such that for a given a span t_i, \dots, t_k , the model predicts t_i and t_k based on the context C (surrounding tokens) and the masked span representation. This helps in learning better representations for start and end semantics of phrases.

The core of SpanBERT is a multi-layer transformer, which processes the input sequence. The self-attention allows each position to attend to all positions and is computed as follows. Firstly, queries Q , Keys K , and Values V are derived from the input embeddings X , as shown in Equation 5:

$$Q = XW^Q, K = XW^K, V = XW^V. \quad (5)$$

The attention output for each head is shown in Equation 6:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (6)$$

where Q, K, V are the query, key, and value matrices derived from the input embeddings, and d_k is the dimensionality of each head. The outputs from all heads are concatenated and linearly transformed.

However, for multi-head attention, calculated in Equation 7:

$$\text{Multi-Head}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O. \quad (7)$$

The position-wise feed-forward networks performed each position independently transformed. After attention, each position x goes through a feed-forward network, as shown in Equation 8:

$$\text{FFN}(x) = \max(0, xW1 + b1)W^2 + b^2. \quad (8)$$

Finally, using the output from SpanBERT, a classification layer predicts the sentiment. In the context of review classification, a simple logistic regression layer can be sufficient, which uses the embeddings (often from the special CLS token placed at the beginning of each input sequence) to predict the sentiment of the review. The logistic regression model computes a probability of the input belonging to a positive class, based on a sigmoid function applied to a linear combination of the input features. The pooled output h from SpanBERT (often the CLS token's embedding) has been computed such that, as shown in Equation 9:

$$P(y = 1 | h) = \sigma(ht + b), \quad (9)$$

where σ is the sigmoid function for binary classification.

In the loss function, the training process involves minimizing the binary cross-entropy loss. For predicted probabilities y^i and true labels y , the loss function L is calculated as shown in Equation 10:

$$L = -\sum_{i=1}^N [y^i \log(y^i) + (1 - y^i) \log(1 - y^i)]. \quad (10)$$

Applying Fusion

Fusion involves combining outcomes derived from various features extracted from both text and images. In the referenced studies, SA utilized both modalities, integrating results from textual and visual features through fusion. Ultimately, the results from the Vti transformer model and the SpanBERT model were amalgamated to create a cohesive feature vector set for learning. The model then calculated the final sentiment score using a specific equation mentioned in Huang et al. (2019) and shown in Equation 11:

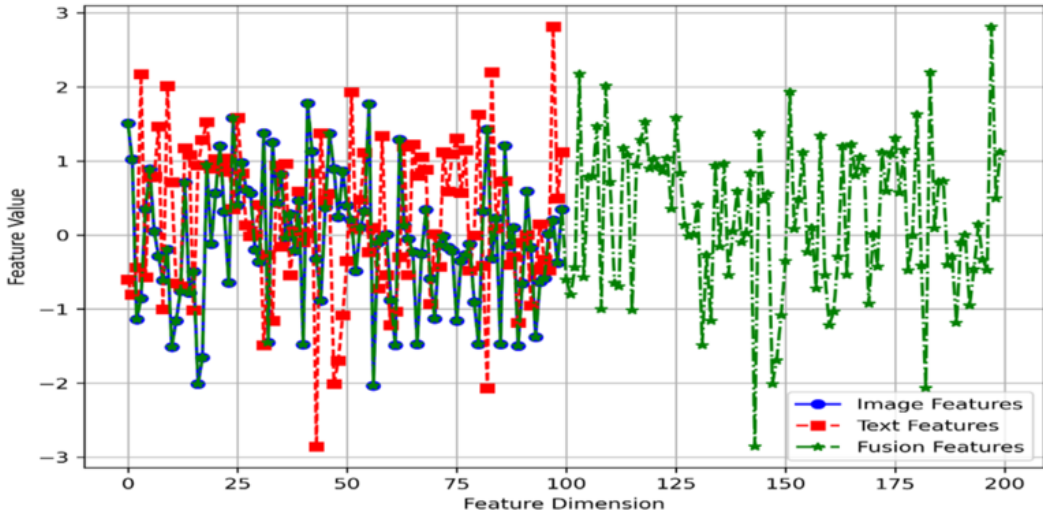
$$Z_i = \begin{cases} \frac{1}{1 + \alpha + \sigma} (Z_i^{(m)} + \alpha Z_i^{(v)} + \sigma Z_i^{(o)}) & \text{input: } (V_i, T_i) \\ Z_i^{(v)} & \text{input: } (V_i) \\ Z_i^{(o)} & \text{input: } (T_i) \end{cases}, \quad (11)$$

where α and δ are hyper-parameters which determine how much emphasis should be placed on classifiers. Figure 4 shows the visualization of text- and image-based sentiment fusion after applying the proposed strategy.

Final Product Recommendation

Collaborative filtering and content-based strategies form the core of traditional recommender systems. Collaborative filtering systems utilize a user-item interaction matrix to track past interactions and generate recommendations. These systems are further divided into memory-based and model-based approaches. Memory-based approaches recommend items by identifying users similar to a given user and suggesting items those similar users liked, though they struggle with variations and biases. On the other hand, model-based techniques build a generative model on the interaction matrix to predict new user preferences but can suffer from model bias and volatility, especially with sparse data. Content-based systems, however, incorporate user preferences and product characteristics like popularity and descriptions alongside user-item interactions. They use this information to feed a machine learning model that optimizes for errors, aiming to recommend products that are most likely to elicit positive reviews from users. These systems typically have higher biases but lower variance, and they focus on aligning recommendations closely with user preferences.

Figure 4. Image and text based sentiment visualization using fusion



To construct a more comprehensive product recommendation system, this research utilized user profiling, item representation, and similarity calculations to recommend profound products. The user profiling was based on two major components: the maintenance of history of products and sentiment score. The history maintenance involves tracking each user's interaction with products, including purchases, views, and ratings. Let U represent a user and P represent a set of products, $P = \{p_1, p_2, \dots, p_n\}$, with which the user has interacted. The sentiment score has been obtained from the previously discussed Vti transformer and SpanBERT, and we analyzed the sentiment of reviews written by the user. For each product p_i , we computed a sentiment score s_i using an SA model: $s_i = \text{SentimentAnalysis}(r_i)$, where r_i is the review text for product p_i .

Each user is represented as a profile vector based on their interactions and sentiments towards products. This vector can be composed of various features such as User Vector $= u = [v_1, v_2, v_3, \dots, v_m]$, average sentiment score, product categories, frequency of purchases, and more, where each v_i could be derived from user behaviors and preferences like average sentiment scores, favorite product categories, or frequency metrics. Similarly, each item in the catalog is also represented as a vector.

Finally, to match users with products that they are likely to enjoy, one should calculate the cosine similarity between the user vector and each item vector in the catalog. The cosine similarity for a user u and an item i is given in Equation 12:

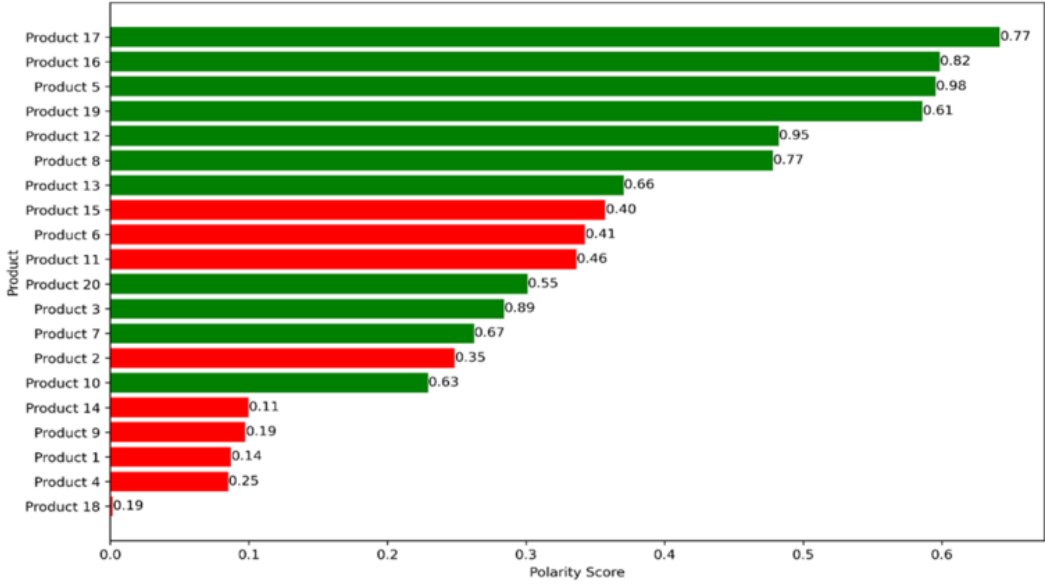
$$\text{Cosine Similarity}(u, i) = \frac{u \cdot i}{\|u\| \|i\|}, \quad (12)$$

where \cdot denotes the dot product of the vectors, and $\|u\|$ and $\|i\|$ are the norms of the vectors.

Figure 5 shows the similarity scores-based recommendation of the top- N items to the user. To sort the items by descending similarity score and select the top N , the calculation should be carried out as in Equation 13:

$$\text{Top} - N \text{ Items} = \text{sort}(\{\text{Cosine Similarity}(u, i_1), \dots, \text{Cosine Similarity}(u, i_n)\})[1:N] \quad (13)$$

Figure 5. Recommendation of top N item to end user



RESULTS AND DISCUSSION

This section details the experimental outcomes and evaluates the effectiveness of the proposed method. Numerous experiments were conducted to assess the accuracy and efficiency of the proposed system. The experimental evaluation demonstrated that the proposed technique significantly outperforms existing cutting-edge methods.

Performance Evaluation Measure

Precision, recall, F1 score, and ultimate accuracy are utilized as standard benchmarks to assess the effectiveness of the proposed Multus-Medium model. These qualities are calculated in Equations 14-17 using the corresponding measures true positive (TP), false negative (FN), and false positive (FP):

$$precision = \frac{TP}{TP + FP} ; \quad (14)$$

$$recall = \frac{TP}{TP + FN} ; \quad (15)$$

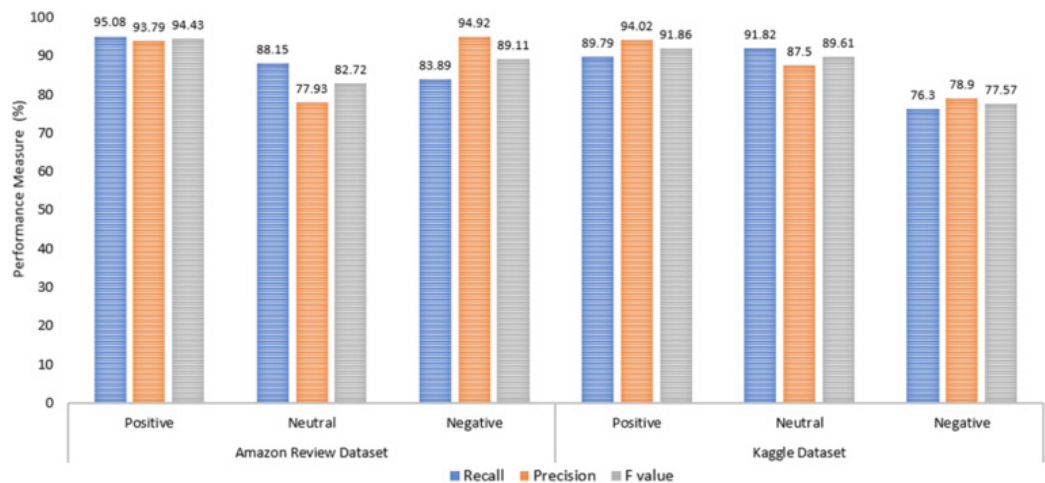
$$F1 \text{ score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} ; \quad (16)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} . \quad (17)$$

It has been common practice to use a receiver operating characteristic (ROC) curve to assess a classifier's efficiency. Equation 18 shows the formula for the performance indicator (Huang et al., 2020).

$$ROC = \frac{CP(\frac{i}{positive})}{CP(\frac{i}{negative})} . \quad (18)$$

Figure 6. Performance measure analysis on each dataset



Baselines Models

To assess the performance of the proposed model, the following baseline models were compared. Medhat et al. (2014) utilized a multi-modal approach (positive, negative) to analyze social images through an AHRM. Gastaldo et al. (2013) introduced a method known as DMAF for SA. This technique enhances sentiment categorization accuracy by focusing on specific locations and words linked to emotions, gathering both complementary and non-redundant information. Ghorbanali et al. (2022) presented a novel strategy named AMGN which capitalizes on the interplay between the modalities of images and texts to identify critical features for multi-modal SA.

Results

The initial experiment highlighted the effectiveness of the proposed method by evaluating its precision, accuracy, and recall. The results are graphically represented in Figure 6, which illustrates the performance of the proposed approach on various datasets, specifically Amazon and Kaggle datasets.

The graphical representation in Figure 6 clearly demonstrates that the proposed technique achieves high precision, recall, and accuracy scores across various datasets. Specifically, on the images dataset, the proposed method achieved a precision of 82.74%, recall of 75.15%, and accuracy of 87.45%, showcasing impressive results. On the voice dataset, the method scored even higher, with a precision of 85.47%, recall of 77.2%, and an accuracy of 84.24%, further highlighting its effectiveness. Additionally, the technique also performed commendably on the These results underscore the efficiency and superior performance of the proposed approach in identification, excelling in terms of precision, accuracy, F-score, and recall across diverse datasets.

In a further evaluation, the performance of the proposed work was also demonstrated using ROC curves. An ROC curve serves as a visual tool that illustrates the effectiveness of a binary classification model under various categorization thresholds. It plots the sensitivity (true positive rate) against the specificity (false positive rate) for different thresholds. The ROC curves for each dataset are displayed in Figures 7 and 8, providing a clear and comprehensive overview of the model's ability to classify data across varying conditions and thresholds accurately. This depiction helps in assessing the robustness and precision of the proposed approach in differentiating between classes. The ROC analysis evaluates the TP rate and FP rate across each dataset. The classification method introduced in this study achieved an accuracy of 88.54% and 90.38% on the Twitter and Flickr8k datasets, respectively. It is essential to delve into a more detailed technical discussion to understand

Figure 7. The ROC-based performance measure of proposed model for amazon review dataset

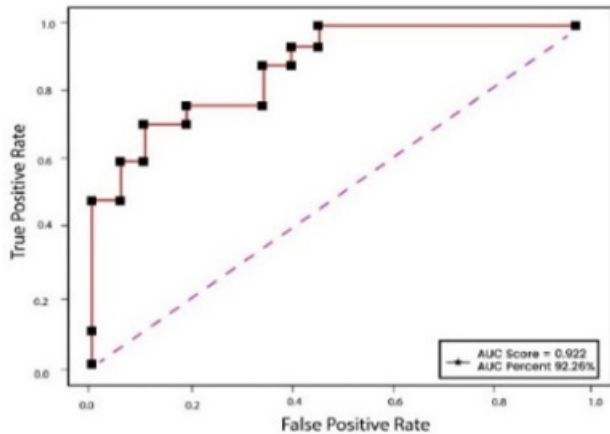
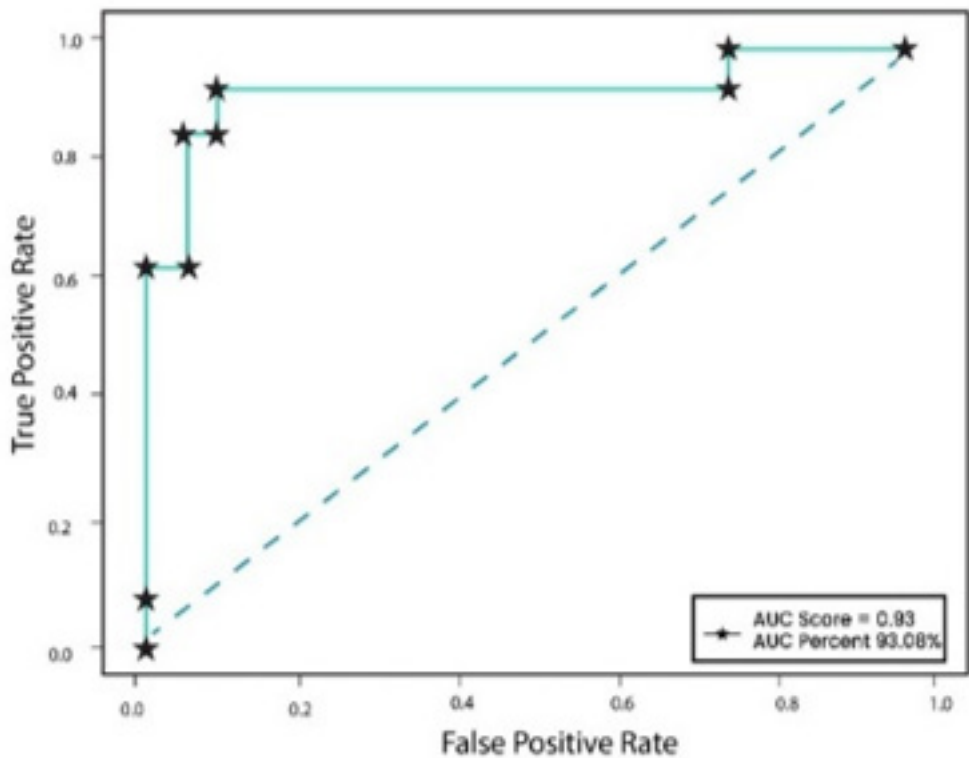


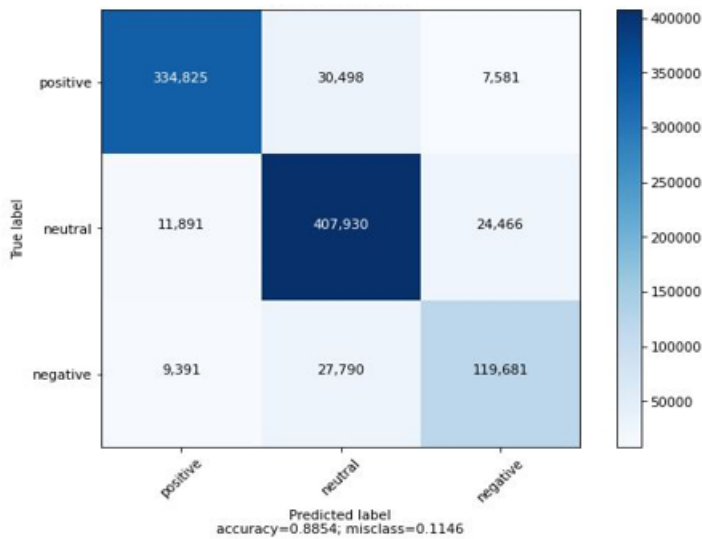
Figure 8. The ROC-based performance measure of proposed model for Kaggle dataset



these performance metrics and their implications better. The area under the ROC curve is 0.92 for dataset 12a and 0.93 for dataset 12b.

A confusion matrix, utilized as a performance metric, was constructed to showcase the efficacy of the proposed method, highlighting TP, TN, FP, and FN, as seen in Figure 7. This matrix illustrates

Figure 9. The confusion matrix-based performance measure of proposed model for Amazon review dataset



two distinct scenarios of predicted versus actual values. The accuracy of the proposed system is recorded at 88.54% in Figure 7(a) and 90.38% in Figure 7(b), demonstrating that the classification approach is effective and yields favorable results. Additionally, metrics such as precision, recall, F1 score, and accuracy were used to assess the performance of the proposed model.

Table 3 provides a detailed summary of the performance of the proposed model across different sentiment categories within the Flicker8k and Twitter datasets, highlighting its capacity for accurate sentiment classification. For the Amazon dataset, the model demonstrated excellent recall of 95.08% for the positive class, with a precision of 93.79%, resulting in an F value of 94.43%, indicating a strong balance between recall and precision. The neutral class showed a recall of 88.15% and a precision of 77.93%, leading to an F value of 82.72%. For the negative class, the recall was 83.89% and the precision was impressively high at 94.92%, with an F value of 89.11%. Overall, the average recall, precision, and F value across all classes were 89.04%, 88.88%, and 88.75%, respectively.

Turning to the Kaggle dataset, the model achieved a recall of 89.79% and a precision of 94.02% for the positive class, with an F value of 91.86%. The neutral class showed a recall of 91.82% and a precision of 87.5%, resulting in an F value of 89.61%. The negative class had lower scores, with a recall of 76.3% and a precision of 78.9%, and an F value of 77.57%. The average performance metrics for this dataset were a recall of 85.97%, a precision of 86.80%, and an F value of 86.34%.

The results indicate that the model is highly effective in evaluating product sentiment scores. It demonstrates notable performance, achieving high recall and precision specifically for positive sentiments. However, its performance varies for neutral and negative sentiments across the two datasets used. The consistently high performance observed in both the training and validation phases suggests that the proposed model reliably classifies sentiments with considerable accuracy and minimal error. This consistent performance underscores the model's robustness and its ability to generalize well from training data to unseen data. To gain a comprehensive understanding of these results, a deeper technical analysis is required. This would involve exploring the underlying factors contributing to the model's success and limitations in sentiment classification. Such an exploration would not only shed light on the reasons behind the varying performance for different sentiment categories but also allow us to discuss the broader implications of these findings. This can enhance the understanding of the model's operational dynamics and improve its effectiveness in real-world applications.

Table 3. MMOM performance in sentiment score calculation against each class

Dataset	Class	Recall	Precision	F value
Amazon Review Dataset	Positive	95.08	93.79	94.43
	Neutral	88.15	77.93	82.72
	Negative	83.89	94.92	89.11
	Average	89.04	88.88	88.75
Kaggle Dataset	Positive	89.79	94.02	91.86
	Neutral	91.82	87.5	89.61
	Negative	76.3	78.9	77.57
	Average	85.97	86.80	86.34

In the very last experiment, the performance of the proposed model was compared with the benchmark approaches. The Figure 11 illustrates a comparison of SA model accuracies on Amazon and Kaggle datasets. The proposed work model exhibits superior accuracy on Amazon and remains competitive on Kaggle, surpassing AHRM and closely matching AMGN on both datasets. DMAF shows a notable increase in accuracy on Kaggle compared to Amazon, indicating its varying performance across different datasets.

Figure 10. The confusion matrix-based performance measure of proposed model for Kaggle review dataset

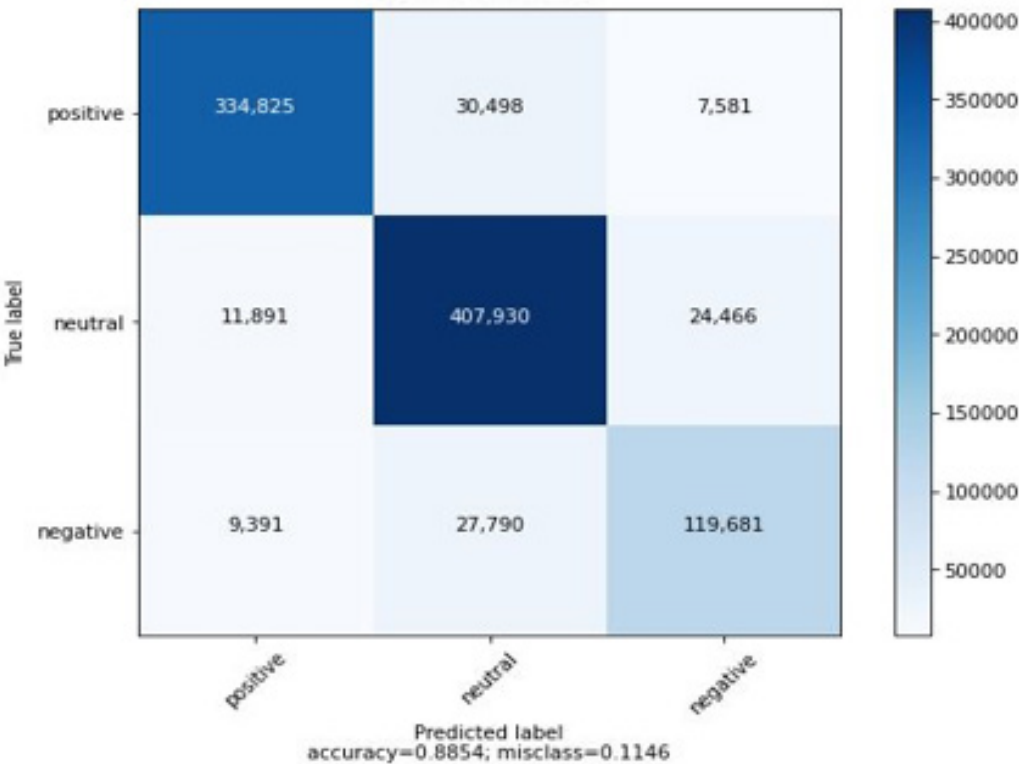
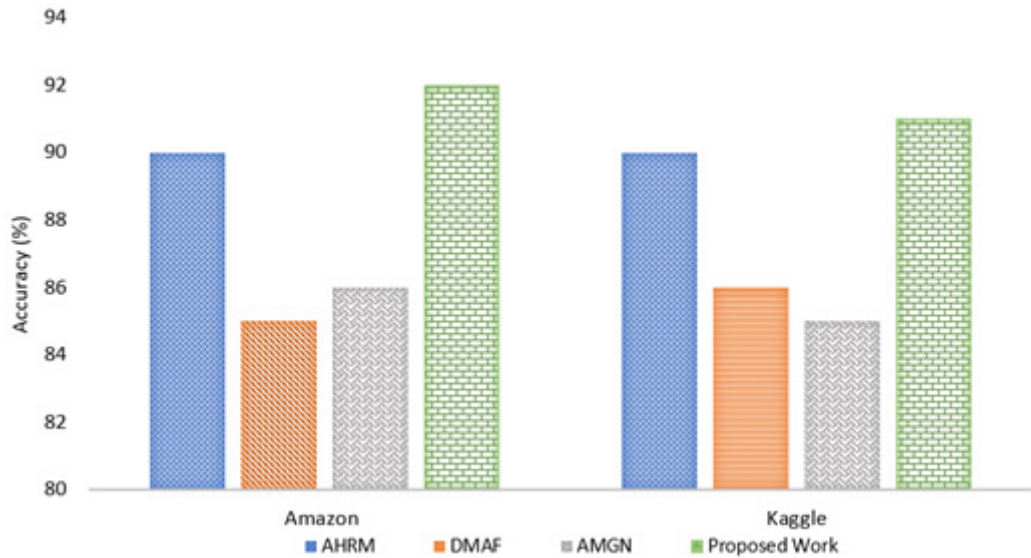


Figure 11. Comparison of proposed work with baseline approaches



Generalizability of Experimental Results

The MMOM approach demonstrated strong performance on two distinct datasets, achieving an accuracy of 91.55% on the Amazon Product Reviews dataset and 90.89% on the Kaggle Dataset. These datasets were selected due to their representativeness of real-world e-commerce environments, where user reviews are a mix of textual and visual content. The results suggest that the MMOM model is robust across different types of products and user-generated content, indicating good generalizability within the domain of e-commerce. To assess the broader applicability of the MMOM approach, we consider the following technical factors.

Diversity of Datasets

The Amazon and Kaggle datasets cover a wide range of products and review styles, including variations in language, sentiment, and image content. The model's consistent performance across these datasets indicates its ability to generalize well to other e-commerce platforms with similar data characteristics. This suggests that the MMOM model could be effective in different contexts within the retail sector, such as fashion, electronics, or home goods.

Multi-Modal Integration

The MMOM approach uniquely integrates text and image data through advanced fusion techniques. This multi-modal strategy is designed to handle various forms of input data, making the model adaptable to different scenarios where users express opinions not just through text, but also through images. The ability to analyze both modalities effectively enhances the model's potential to generalize across various domains that involve rich multi-media content, such as social media SA, product reviews in niche markets, and more.

Model Robustness

The use of Vti for image analysis and SpanBERT for text analysis contributes to the model's robustness, as both are well-established techniques known for their high performance in their respective

domains. The fusion technique employed ensures that the insights from both text and image data are coherently combined, reducing the risk of performance degradation when applied to different datasets.

Potential Limitations

While the MMOM model shows promise, its generalizability might be constrained in environments where the review content significantly deviates from the datasets used in this study. For instance, in highly specialized industries with unique jargon or image types, the model may require fine-tuning or adaptation to maintain its accuracy. Additionally, datasets with extreme imbalances in the distribution of text and image content may challenge the model's ability to integrate the two modalities effectively.

ANALYSIS AND PRACTICAL IMPLICATIONS

The field of SA has evolved significantly, with various methodologies being proposed to tackle the challenges of understanding sentiment in user-generated content. Traditional approaches, such as empirical analysis, have been foundational but often fall short in preserving sequential information, which is crucial for capturing context. Techniques like RNN-LSTM were introduced to address this by modeling sequential dependencies, though their effectiveness can be limited by the reliability of feature extraction methods. More recent approaches, like CNN-LSTM and R-CNN, combine CNNs and RNNs to leverage the strengths of both architectures. These methods have demonstrated high accuracy in specific contexts, such as image and text-based SA. However, challenges remain, particularly in the extraction and preservation of key features, which can impact the overall performance of these models.

Multi-modal approaches, such as the multi-modal event-aware network, have emerged to integrate text and image data, aiming to capture a more comprehensive understanding of sentiment. While these models show promise, their reliability and generalizability across different domains still need improvement. Deep learning methods, despite their powerful capabilities, often suffer from domain specificity, limiting their applicability in diverse scenarios.

In contrast, the MMOM approach proposed in this study builds on these advancements by offering an integrated method that effectively combines text and image analysis. By addressing the limitations of previous methods, such as feature loss and domain specificity, MMOM provides a more robust and versatile solution for SA. This approach not only improves accuracy but also enhances the ability to generalize across various datasets and application areas.

While the primary focus of this study has been on developing and evaluating the accuracy of the MMOM approach, it is crucial to consider how this model can be applied in real-world scenarios. Practical applications of MMOM span across multiple industries, particularly within e-commerce, social media analysis, and customer experience management.

One of the most significant applications of the MMOM approach is in e-commerce platforms, where customer reviews play a pivotal role in shaping purchasing decisions. By integrating text and image data from customer reviews, the MMOM model can provide more accurate and nuanced product recommendations, leading to enhanced customer satisfaction and potentially higher sales conversion rates. For instance, an online retailer could use MMOM to analyze both the written reviews and the images uploaded by customers to assess the overall sentiment towards a product, thereby refining their recommendation algorithm.

Social media platforms generate vast amounts of user-generated content, often comprising both text and images. The MMOM approach can be employed to monitor brand perception and consumer sentiment in real-time by analyzing posts, comments, and images together. This comprehensive SA could help brands to understand public opinion better, allowing them to respond more effectively to consumer concerns and trends.

In the hospitality and service industries, customer feedback often includes detailed reviews along with images of the facilities or services. The MMOM model can be used to analyze this feedback, providing businesses with actionable insights to improve their services. For example, a hotel chain

could use MMOM to analyze customer reviews and photos and identify specific areas of their services that need improvement, such as room cleanliness or food quality.

While these practical applications illustrate the potential impact of the MMOM approach, detailed case analyses were beyond the scope of the current study. The in-depth case studies that apply the MMOM model in various industry settings could also be considered. These case studies will provide a more comprehensive understanding of the model's practical utility and its effectiveness in real-world applications.

CONCLUSION AND FUTURE WORK DIRECTION

Online product reviews and comments wield considerable influence over purchasing decisions, sales figures, and product quality enhancements while also shaping recommendation systems for new users. This study employed a deep learning strategy for SA on product reviews, utilizing multi-modal data. It introduced an innovative approach that leverages both textual and visual features to provide superior product recommendations, harnessing unique and discriminative characteristics. By merging sentiment scores derived from the Vti transformer and SpanBERT models through novel fusion techniques, the proposed model achieved enhanced SA, exhibiting superior recommendation accuracy compared to existing models. Experimental findings highlighted the proposed model's impressive accuracy, F1 score, and ROC values, with notable achievements on both the Amazon and Kaggle datasets. The Multus-Medium approach, particularly the proposed model, presented a robust framework for SA and recommendation systems, effectively integrating textual and visual data. The success of this model suggests opportunities for future research to explore advanced decision-making techniques for accuracy improvement. Moreover, the proposed framework holds potential for extension to other sentiment-driven applications, such as hospital recommendations, agricultural advisories, and medical diagnostics, thereby broadening its applicability and impact.

Future research could focus on several avenues to enhance the capabilities of the proposed model further. One direction is the exploration of more sophisticated fusion techniques which can better integrate textual and visual features, potentially leading to even higher accuracy in SA and recommendation systems. Additionally, investigating the model's applicability in real-time processing scenarios and its adaptability to other domains with sentiment-driven data could provide valuable insights. Moreover, extending the model to incorporate additional data modalities, such as audio or video, could open new possibilities for more complex and diverse applications. Finally, expanding the framework to support multi-lingual SA could significantly increase its relevance in a global context.

AUTHOR NOTE

Mohamed Taha (<https://orcid.org/0000-0003-0885-0985>)

Diaa Salama Abdelminaam (<https://orcid.org/0000-0003-0881-3164>)

The authors would like to acknowledge the support of Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R435), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

CONFLICTS OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

PROCESS DATES

September 30, 2024

Received: April 30, 2024, Revision: September 2, 2024, Accepted: September 19, 2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Asif Nawaz (Pakistan, asif.nawaz@uaar.edu.pk)

REFERENCES

- Al-Sammarraie, Y. Q., Khaled, A. Q., Al-Mousa, M. R., & Desouky, S. F. (2022). Image captions and hashtags generation using deep learning approach. In *Proceedings of the 2022 International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI)* (pp. 1-5). IEEE. DOI: 10.1109/EICEEAI56378.2022.10050455
- Alsaity, A., & Orji, R. (2024). Machine learning techniques for emotion detection and sentiment analysis: Current state, challenges, and future directions. *Behaviour & Information Technology*, 43(1), 139–164. DOI: 10.1080/0144929X.2022.2156387
- Bhuvaneshwari, P., Rao, A. N., Robinson, Y. H., & Thippeswamy, M. N. (2022). Sentiment analysis for user reviews using Bi-LSTM self-attention based CNN model. *Multimedia Tools and Applications*, 81(9), 12405–12419. DOI: 10.1007/s11042-022-12410-4
- Chaudhry, H. N., Javed, Y., Kulsoom, F., Mehmood, Z., Khan, Z. I., Shoaib, U., & Janjua, S. H. (2021). Sentiment analysis of before and after elections: Twitter data of us election 2020. *Electronics (Basel)*, 10(17), 2082. DOI: 10.3390/electronics10172082
- El-Affendi, M. A., Alrajhi, K., & Hussain, A. (2021). A novel deep learning-based multilevel parallel attention neural (MPAN) model for multidomain Arabic sentiment analysis. *IEEE Access: Practical Innovations, Open Solutions*, 9, 7508–7518. DOI: 10.1109/ACCESS.2021.3049626
- Gastaldo, P., Zunino, R., Cambria, E., & Decherchi, S. (2013). Combining ELM with random projections. *IEEE Intelligent Systems*, 28(6), 46–48.
- Ghorbanali, A., Sohrabi, M. K., & Yaghmaee, F. (2022). Ensemble transfer learning-based multimodal SA using weighted convolutional neural networks. *Information Processing & Management*, 59(3), 102929. DOI: 10.1016/j.ipm.2022.102929
- Goularte, F. B., da Graça Martins, B. E., da Fonseca Carvalho, P. C. Q., & Won, M. (2024). SentPT: A customized solution for multi-genre sentiment analysis of Portuguese-language texts. *Expert Systems with Applications*, 245, 123075. DOI: 10.1016/j.eswa.2023.123075
- Huang, F., Wei, K., Weng, J., & Li, Z. (2020). Attention-based modality-gated networks for image-text sentiment analysis. [TOMM]. *ACM Transactions on Multimedia Computing Communications and Applications*, 16(3), 1–19. DOI: 10.1145/3388861
- Huang, F., Zhang, X., Zhao, Z., Xu, J., & Li, Z. (2019). Image-text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems*, 167, 26–37. DOI: 10.1016/j.knosys.2019.01.019
- Kanwal, B., Rana, M. R. R., Nawaz, A., & Kiani, A. N. (2024). Benchmarking travelling reviews using opinion mining. *Foundation University Journal of Engineering and Applied*, 4(1), 1–12.
- Khan, L. (2023). Improved multi-lingual sentiment analysis and recognition using deep learning. *Journal of Information Science*, 01655515221137270. DOI: 10.1177/01655515221137270
- Kumar, M. T., Kumar, N., Sha, S. N., Kennedy, E. N., & Ilankadhir, M. (2024). E-commerce strategies in the digital age enhancing customer experience and market reach. In *Proceedings of the 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)* (pp. 1-7). IEEE.
- Li, Y., Ding, H., Lin, Y., Feng, X., & Chang, L. (2024). Multi-level textual-visual alignment and fusion network for multimodal aspect-based sentiment analysis. *Artificial Intelligence Review*, 57(4), 78. DOI: 10.1007/s10462-023-10685-z
- Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025. DOI: 10.18653/v1/D15-1166
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113. DOI: 10.1016/j.asej.2014.04.011
- Meena, Y., Kumar, P., & Sharma, A. (2018,). Product recommendation system using distance measure of product image features. In *Proceedings of the 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1351-1355). IEEE. DOI: 10.1109/ICCONS.2018.8663113

- Mei, Z., Yu, J., Zhang, C., Wu, B., Yao, S., Shi, J., & Wu, Z. (2024). Secure multi-dimensional data retrieval with access control and range query in the cloud. *Information Systems*, 122, 102343. DOI: 10.1016/j.is.2024.102343
- Miah, M. S. U., Kabir, M. M., Sarwar, T. B., Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports*, 14(1), 9603. DOI: 10.1038/s41598-024-60210-7 PMID: 38671064
- Onan, A. (2020). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation*, 33(23), e5909. DOI: 10.1002/cpe.5909
- Rana, M. R. R., Nawaz, A., Ali, T., El-Sherbeeney, A. M., & Ali, W. (2023). A BiLSTM-CF and BiGRU-based deep sentiment analysis model to explore customer reviews for effective recommendations. *Engineering, Technology & Applied Scientific Research*, 13(5), 11739–11746.
- Rana, M. R. R., Rehman, S. U., Nawaz, A., Ali, T., & Ahmed, M. (2021). A conceptual model for decision support systems using aspect based sentiment analysis. *Proc. Rom. Acad. Ser. A-Mathematics Phys. Tech. Sci. Inf. Sci.*, 22(4), 381–390.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- Singh, U., Abhishek, K., & Azad, H. K. (2024). A survey of cutting-edge multimodal sentiment analysis. *ACM Computing Surveys*, 56(9), 1–38. DOI: 10.1145/3652149
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-9). IEEE.
- Wu, S., Fei, H., Ren, Y., Li, B., Li, F., & Ji, D. (2021). High-order pair-wise aspect and opinion terms extraction with edge-enhanced syntactic graph convolution. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 2396–2406. DOI: 10.1109/TASLP.2021.3095672
- Yang, J., She, D., Sun, M., Cheng, M. M., Rosin, P. L., & Wang, L. (2018). Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia*, 20(9), 2513–2525. DOI: 10.1109/TMM.2018.2803520
- Zulqarnain, M., Ghazali, R., Aamir, M., & Hassim, Y. M. M. (2024). An efficient two-state GRU based on feature attention mechanism for SA. *Multimedia Tools and Applications*, 83(1), 3085–3110. DOI: 10.1007/s11042-022-13339-4

Zohair Ahmed, received the degree of BS(CS) from Arid Agriculture University, Pakistan and MS(CS) from International Islamic University, Pakistan in 2014 and 2017, respectively. He is currently pursuing his Ph.D. degree in the school of computer science and engineering at Central South University, Changsha, China. His research interest includes Natural language Processing (NLP), Opinion Mining, Sentiment Analysis, and developing efficient schemes for semantic analysis using classification and similarities measures.

Mohammad Alshinwan received the Ph.D. degree from the School of Computer Engineering, Inje University, Gimhae, Republic of Korea, in 2017. He was an Assistant Professor with the Department of Computer and Information Sciences, Amman Arab University, Jordan. He is currently an Associate Professor with Applied Science Private University, Jordan. His research interests include computer networks, mobile networks, information security, AI, and optimization methods.