Multus Medium Opinion Mining (MMOM): A Novel Hybrid Deep Learning Approach for Comprehensive Product Review Analysis

Bushra Kanwal , Asif Nawaz * , Ghulam Mustafa , Tariq Ali , Muhammad Babar , Basit Qureshi , Anis Koubaa

Posted Date: 6 June 2023

doi: 10.20944/preprints202306.0405.v1

Keywords: Multus Medium; Opinion Mining; Deep Learning; Product Review Analysis; Bidirectional Long Short-Term Memory; Convolutional Neural Network

Article

Multus Medium Opinion Mining (MMOM): A Novel Hybrid Deep Learning Approach for Comprehensive Product Review Analysis

Bushra Kanwal [1], Asif Nawaz [1,*], Ghulam Mustafa [1], Tariq Ali [1], Muhammad Babar [2], Basit Qureshi [3] and Anis Koubaa [2]

[1] University of Institute Information Technology, PMAS-Arid Agriculture University Rawalpindi, Pakistan
[2] Robotics and Internet of Things Lab, Prince Sultan University, Riyadh, Saudi Arabia
[3] College of Computer and Information Systems, Prince Sultan University, Riyadh, Saudi Arabia
* Correspondence: asif.nawaz@uaar.edu.pk

Abstract: Advancements in technology have revolutionized communication on social media platforms, where reviews and comments greatly influence product selection. Current opinion mining methods predominantly focus on textual content, overlooking the rich information within customer-posted images, termed here as Multus-Medium. This research introduces an innovative deep learning approach, Multus Medium Opinion Mining (MMOM), capitalizing on both text and image data for comprehensive product review analysis. MMOM employs an integrated model of Bidirectional Long Short-Term Memory (BiLSTM) and embedded Convolutional Neural Network (CNN), incorporating architectures of GoogleNet and VGGNet, thus enabling efficient extraction and fusion of textual and visual features. This synergistic approach enables data collection, preprocessing, feature extraction, fusion strategy-based feature vector generation, and subsequent product recommendation. Performance evaluation on two diverse real-world datasets name as "flicker8k" and "t4sa". These datasets show substantial improvement over existing methods. MMOM outperforms standard benchmark models, achieving an accuracy, F1 score and ROC over fliker8k are 90.38%, 88.75% and 93.08%, whereas for twitter dataset are 88.54%, 86.34%, and 92.26% respectively, the accuracy of purposed model is 7.34% and 9.54% higher than the other two mentioned techniques. These statistics highlight the robustness and applicability of MMOM across various domains. The compelling results underscore the potential of MMOM, providing a more holistic and precise approach to opinion mining in the era of social media product reviews. This research, product recommendation helps the customer to make purchasing decision. Last but not the least, the purposed scheme can further be expanded in any other sentiment task like hospital recommendation system, crop farming recommendation and medical diagnostic system etc.

Keywords: Multus Medium; Opinion Mining; Deep Learning; Product Review Analysis; Bidirectional Long Short-Term Memory; Convolutional Neural Network

1. Introduction

Electronic commerce is a commercial practice that exchanges products and services online. The emergence of e-commerce has enabled consumers to buy products and services at their discretion, regardless of time and place, without leaving the comfort of their own homes or places of business. With the advent of e-commerce, companies can now serve their clientele around the clock, which often results in more business and happier clients. Review sites, online shopping sites, and weblogs have allowed almost anybody to voice their opinion on products and services. Horrigan [1] surveyed online shoppers in 2008 and found that 81% of respondents had researched the goods using internet resources before purchasing, with 79% expressing confidence that they had made the best possible choice. People can communicate their feelings and experiences by posting their opinions and images online. Emotions include a vast spectrum of human affairs, from liking something to disliking

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

something, being happy to feel sad, and being low to feeling happy [2]. It also helps companies raise their product quality because of these reviews.

Opinions expressed in online reviews have a significant part in the buying decisions of online shoppers. Reviews give consumers an idea of what it's like to purchase from a company by providing feedback on the quality of the items and services they've experienced. Online reviews are crucial since they may significantly affect a company's image and, by extension, its revenue. Customer confidence in a company and the number of consumers that patronize it may increase with positive evaluations, whereas credibility and customer aversion can decrease with unfavorable reviews [3]. Online reviews are more crucial than ever in today's market. With the excess of options available, customers put a premium on hearing about the perspectives and experiences of others before making a purchase. Hence, in today's cutthroat e-commerce environment, survival requires keeping an eye on customer feedback and responding appropriately. Figure 1 states that if someone is concerned about purchasing a headphone, they can look online and read reviews from other customers about the product. After going through these reviews, the person's decision will be influenced by the opinions of other customers [1].

Figure 1. Textual Reviews.

Online reviews are assessments or viewpoints expressed by customers or users of a good, service, or company online. They assist other prospective consumers in making selections by sharing information and feedback. Online reviews can be accessed on many different platforms, including social media sites, specialized review websites, and mobile apps. They often include a rating or score along with a written summary of the reviewer's experience, level of satisfaction, and any advantages or disadvantages they encountered. Online reviews are something that businesses frequently pay attention to because they have a big impact on their reputation and how customers perceive them. Reviews may be broken down into two broad categories: quantitative (like star ratings) and qualitative (like lists of pros and cons) [4]. It can also include Annotation made in the form of words or pictures that classifies how one feels. Such quantitative expressions of emotion are the star ratings seen on many online retailers' sites, as shown in Figure 2. This rating system is used to assess people's opinions. Classification refers to separating one kind of review or criticism into distinct categories. Reviews may be emotionally categorized in several ways, including positive, negative, neutral, and some additional categories, including anger, sadness, happiness, annoyance, etc.

Figure 2. Rating Reviews.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

As a result, customers are now more dependent on product reviews to obtain information and make informed purchasing decisions. Due to the sheer volume of information available, about 32% of consumers have reported confusion, while 30% have expressed frustration. Reading all the reviews can prove to be time-consuming, yet many people prefer to go through a few reviews before opting for a product or making a choice [5,6]. Sentiment analysis (SA) offers useful insights for people and firms in the modern digital age, where enormous amounts of data are produced everyday through online reviews, social media posts, and client feedback. The sentiment or opinion represented in text data, such online reviews, is analyzed and determined using sentiment analysis. SA has become a critical or beneficial tool to extract, evaluate and report the sentiments of people using any social media platform [7].

SA is a sub-field of NLP (Natural Language Processing), which automatically identifies opinions from text data. Its primary objective is to categorize reviews created by users as negative or positive depending upon the author's viewpoint or sentiment on a particular subject. As per Liu, opinion mining or sentiment analysis, as both terms, can be used alternatively in the research or study of behaviors, attitudes, feelings, emotions, evaluation, categorization, opinions, and sentiments towards various aspects of a service or product or a person [8].

Many different ways for opinion mining may be used to provide product recommendations. These approaches rely mainly on the text content that consumers give. Consumers can communicate their feedback using Multus-medium, including text and graphics [9].Online reviews that include both text and photos are referred to as multus-medium reviews. Reviews that include text and images are very important in many fields. They provide people with a more thorough and interesting platform to share their thoughts and experiences. Images offer context and visual proof that writing alone often struggles to convey. Reviewers may provide a more thorough and accurate account of their experiences by incorporating photographs with their comments, empowering future consumers to make more educated selections. Reviews with both images and text are frequently seen as being more sincere and reliable. Since real-world photographs are more difficult to manipulate than text-only reviews, seeing them with written criticism strengthens the review's trustworthiness [10]. Images have the power to stir feelings and forge closer ties with readers. Reviews are more important when consumers use images to describe their experiences because they connect with prospective customers on a deeper level. While typical online evaluations are mostly composed of written descriptions, multus-medium reviews improve the reviewing process by including both textual and visual aspects that complement or supplement the material. As can be seen in Figure 3, users express their thoughts, sentiments, and emotions by posting text and photographs on social media platforms regarding their delayed flights.

Figure 3. Multus-Medium Reviews [9].

The product recommendation to the new customer, many analysts have evaluated customer reviews using traditional machine learning and opinion grouping approaches like deep multimodal attentive fusion, VGG-Net-16, Attention-Based Modality-Gated Networks (AMGN), and others [11]. The Multus-Medium reviews include ample information that must be evaluated before processing. As a result, a sophisticated technique is employed to analyze client feedback from Multus-Medium image data. This research focuses on the Multus-Medium, which extracts the feature from images and text and helps customers recommend the product. Multus-medium reviews have the potential

 doi:10.20944/preprints202306.0405.v1

to place a big impact on e-commerce by improving user engagement, trust, product understanding, and customer service. Multus-medium evaluations give companies a competitive edge, higher conversion rates, and insightful market data. A more knowledgeable and interesting shopping experience benefits the customer, increasing their contentment and self-assurance in their purchase choices. It will significantly benefit individuals to make better purchasing decisions. Also, it will provide a revolutionary help to the organizations to make better decisions regarding their products and policies.

Textual ratings and reviews are the foundation of all the current work on recommender systems. When doing so, however, they ignore the Multus-Medium photos that also contain a rich amount of information consumers provide. Reviews that include both text and images are extremely important in influencing customers' purchasing choices but existing approaches neglect the visual reviews. Images give products and services a real depiction, enabling clients to perceive the true appearance, quality, and characteristics. Customers may better grasp what to anticipate thanks to this visual context, which also helps them make more educated choices. As social evidence, image and text reviews demonstrate that others have had favorable experiences with a good or service. Customers' confidence is increased and their perception of risk is decreased when they can see visual proof of several happy consumers, which increases the possibility that they will make a purchase. Current approaches neglect these important characteristics because of ignorance of visual reviews. The evaluations presented in the style of Multus-Medium include a substantial amount of data, which also has to be considered for the analysis. MMOM (Multus medium opinion mining) is a technique suggested to be used in the study to review product reviews. It offers an improved model for deep learning by using BiLSTM and an embedded CNN that is trained on GoogleNet in addition to VGGNet.

1.1. Research Contribution

Here is the list of main contributions of this research,

- Investigation of Multus-Medium by utilizing the discriminative and unique features in the form of texts and images to get the recommendation of a better product.
- BiLSTM embedded CNN and feature fusion with sentiment analytic model. The proposed MMOM produces a better recommendation.
- Experimental results demonstrate that the accuracy, F1 score and ROC over fliker8k are 90.38%, 88.75% and 93.08%, whereas for twitter dataset are 88.54%, 86.34%, and 92.26% respectively, the accuracy of purposed model is 7.34% and 9.54% higher than the other two mentioned techniques.

In the coming sections of this research, literature review about the techniques and practices involved in analyzing the sentiments available in the shape of text or images on the internet. The proposed methodology is discussed in Section 3. It includes a method to collect and analyze precise data. The proposed approach will help gather accurate information and provide insights into the research subject. Analysis and Results after the data analysis is presented in the fourth section of the study. After investigation, section 5 of the study discusses the results. In the ending part, conclusions, recommendations, and future research guidelines are presented in section 6 of the study.

2. Literature Review

In recent years, opinion mining emerged as a need to understand customer feedback about a product before making any decision about purchasing a specific product or service [12–15]. The following section is about the state-of-the-art work in the field, which is further divided into three sections, i.e., text reviews-based recommendation, image reviews-based recommendation, and multus medium reviews-based recommendation. The detail of each section is discussed below.

 doi:10.20944/preprints202306.0405.v1

2.1. Text Reviews-based Recommendation

Research in the area of analyzing the sentiments of people is being carried out extensively nowadays. In a study cited as [16], panel data of commercial banks in the U.S. were analyzed. Their study employed two measures: "Garcia's sentiment measure," which includes the analysis of positive or negative financial news from the editions of the New York Times, and "News Implied Volatility," a measure of text uncertainty. Their research underscores the adverse impact of the Great Financial Crisis on the investors' sentiments, which resultantly affected declining lending behavior in the U.S., which was a factor in the banking sector becoming unstable. To evaluate the 79% accuracy of their model. Their model required complex training and large dataset. A large dataset and complex training are essential for model development for a number of reasons. First off, a complex model can approximate underlying functions more precisely since it can capture complicated relationships and manage high-dimensional or non-linear data. Additionally, a large dataset's variety of samples enables the model to pick up on a variety of patterns and generalize well to situations that haven't been observed before. Furthermore, complex models are well-suited for jobs involving complex or rich data types since they are excellent at collecting nuanced aspects. Additionally, they are more resilient when dealing with unclear or noisy data. Last but not least, complicated models are fundamentally necessary to attain high performance in some domains or challenging issue types. But it's crucial to take resource constraints into account and balance model complexity and performance.

The work of Luong et al. [17] proposed a deep learning model to categorize the sentiments of people in the reviews from the text. The authors observed that a comprehensive deep learning-based approach incorporating a unified feature set, including embedding words, knowledge of sentiments, rules of sentiment shift, and linguistic and statistical expertise, was not used for sentiment analysis. Their work used RNN (Recurrent Neural Network) with LSTM (Long Short-Term Memory) for sequential processing and a sentence-level sentiment approach to classifying text reviews. Their model Loss the sequential and contextual information. The achieved accuracy of their model was 74.78%. The performance and dependability of models that rely on sequential and contextual information might be affected by the loss of that information. Szegedy et al. [18] describe a research study introducing a deep learning technique for performing sentiment analysis on product reviews using the Twitter dataset. Their proposed method combined TF-IDF weighted GloVe word embeddings and a CNN-LSTM architecture. They performed numerous experiments to analyze and compare the performance of different word embedding schemes having traditional deep neural network architectures. The study's findings revealed that the proposed deep learning architecture performed better than conventional deep learning approaches. Their model achieved 89% accuracy but the drawback of their model was unable to normalize polarity intensity of words. A model's inability to normalize the polarity intensity of words can have a number of negative effects. First off, the model might have trouble faithfully capturing and representing the finer distinctions in mood or emotion indicated by various phrases. As a result, the model might not be able to distinguish sufficiently between very positive, somewhat positive, or neutral sentiments, leading to less accurate sentiment analysis or emotion classification. Second, the performance of downstream activities that rely on sentiment analysis, such opinion mining, social media analysis, or customer sentiment prediction, may be adversely affected by the absence of polarity intensity normalization. Making accurate judgments and decisions in these jobs necessitates having a detailed understanding of sentiment intensity.

2.2. Image Reviews based Recommendation

In the field of image sentiment analysis, the work of [19] addressed the challenge of recognizing high-level abstractions of image data. They concluded image and local region information contain important sentimental details. To achieve this, they proposed a framework that 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 convolutional neural network (CNN) to compute the sentiment scores for each candidate, considering both abjectness and sentiment scores to discover effective regions

 doi:10.20944/preprints202306.0405.v1

automatically. Finally, the CNN outputs from local regions were combined with the whole images to generate the final predictions. Notably, their framework only required image-level labels, significantly reducing the annotation burden typically associated with training in this area. Their model Fusion reports 83.05% accuracy. They required a more dependable recurrent model for improving the accuracy. Due to that the model's capacity to manage sequential data and generate precise predictions by adopting a more reliable recurrent model. This is crucial in situations when precision and dependability are essential, including in autonomous systems, financial forecasting, medical diagnostics and product recommendation. Correct predictions or unreliable behaviour are hazards that can be reduced with a more reliable recurrent model. It can give you more precise information, help you make better decisions, and make your system or application run better as a whole.

Two-stage deep learning architecture for fashion picture recommendations was proposed by [20]. A visually aware feature extractor driven by data using a neural network classifier was utilized. The latter was subsequently used as input for ranking algorithm-based suggestions that were similarity-based. They evaluated the proposed work on the Fashion dataset, which was publicly available. The authors used their approach in combination with other established systems to enhance the robustness and performance of existing content-based recommendation systems. This aimed to better align with specific client styles, among other benefits. Their model was dictionary-based and domain-specific. The model's capacity to comprehend and react to ambiguous or context-dependent questions may be constrained by the dictionary-based method. Language is frequently complicated, and word meanings can change depending on the situation. The model's predictions may be inaccurate and irrelevant if it is unable to correctly analyze and comprehend context. The ability to learn from user interactions and modify replies may not be present in dictionary-based models. This restricts its room for growth and reduces its ability to provide users' queries the individualized attention they deserve. Overall, a dictionary-based and domain-specific language model may be useful in some situations, but it has drawbacks such as knowledge gaps, rigidity, difficulty keeping up with changes, difficulty understanding context, and limited adaptability.

Wing et al. [21] proposed a model that takes images and their features as inputs to provide users with personalized recommendations. Their model used various image features, such as color and shape, to differentiate images into categories. The authors used the mean and standard deviation of image matrices to classify images and calculated the distance matrix between images to distinguish them. To evaluate the 83% accuracy of their model. This model fails to produce multi dimension features. A more thorough representation of the input is made possible by multidimensional features, allowing the model to incorporate subtle differences and subtleties. Without it, the model may have trouble learning from the data and making generalizations, which would lower its performance and accuracy. Without multidimensional characteristics, the model would have trouble separating several classes or categories, which could cause confusion and incorrect categorization. Overall, the model's capacity to learn intricate patterns, generalize successfully, and accurately handle multiple jobs and data types might be greatly hampered by the lack of multidimensional features.

2.3. Multus-Medium Reviews-based Recommendation

Researchers in [22] Performed sentiment analysis on the Weibos posts about the travelers' experiences with commercial air travel. Travel-related occurrences, such as a flight delay, may hurt passengers' moods. They used a multi-task structural arrangement to instantaneously evaluate events and sentiment by modeling the cross-modal linkages to provide more discriminative representations. The authors extracted features from the prescribed dataset. Numerous tests demonstrate that the suggested procedure outperforms the current cutting-edge methods. The accuracy of their proposed model was 89%. Their model was unable to manage out-of-vocabulary words. A model is said to be incapable of managing out-of-vocabulary (OOV) terms if it is unable to recognize or comprehend words that are not contained in its training set or specified vocabulary. This restriction may cause a number of problems. The model may either respond incorrectly or illogically to OOV terms or it may fail to give any useful output at all. This can seriously hurt the model's

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

performance and impair its capacity to comprehend user queries and provide appropriate answers. OOV words are also frequently used in everyday speech, particularly in dynamic fields or when neologisms, slang, or technical terms are involved. If the model is unable to handle OOV terms, it can find it difficult to keep up with changing slang and fail to offer accurate and current information.

The authors proposed a mixed-fusion architecture [23] for The Image-Text Sentiment Analysis, which leverages the inherent correlation and discriminative properties between visual and semantic contents. Their model utilized two unimodal attention models to classify emotions for the text and image modalities effectively. These attention models focus on identifying the essential words and discriminative regions closely associated with the sentiment. To develop a joint sentiment classification, their model employed an intermediate fusion-based multimodal attention model that considered the internal correlation between visual and textual data. Afterward, the three attention models were combined for sentiment prediction using a late fusion approach. Deep Multimodal Attentive Fusion reports 76% accuracy for Twitter data and 87% for Getty photos. Their model was costly and time-consuming to search for opinionated words.

A model's efficiency, scalability, user experience, and capacity to deliver precise and timely opinions can all be hampered by the expensive and time-consuming nature of searching for opinionated words in a model. To overcome these difficulties, more effective ways of incorporating and processing opinionated words would be needed.

The study of [24] proposed an effective model incorporating extensive social information to enhance the effectiveness of multi-model sentiment analysis. A regional attention schema was utilized to highlight emotional areas based on the attended channels. A heterogeneous relation network was developed to create high-quality representations of social images, and the Graph Convolutional Network was expanded to order content data from social contexts. To evaluate their approach, they experimented on the benchmark datasets of Flickr and Getty. Their model achieved a high accuracy rate of 87% in rating photos from these two datasets but some features are lost during extraction. In order to understand and make informed decisions, features are critical attributes or characteristics of the data that capture relevant patterns, relationships, or properties. If features are lost during the extraction process, however, important information may be discarded or overlooked, which can result in a loss of predictive power, reduced accuracy, and limited understanding of the underlying data.

In the work of [25], a novel approach to sentiment analysis classification was introduced. The method employed discriminative feature extraction from both images and text. The multimodal sentiment analysis was performed by exploiting the correlations between image and text modalities. To achieve this, a visual-semantic attention paradigm was utilized to obtain attended visual aspects for each word, which was further used to develop a semantic self-attention model for automated identification of distinguishing features for sentiment categorization. The approach was tested manually, annotated, and without machine labeling of the dataset. Their technique gives 88% accuracy on FlickrW images and 79% on Twitter data. Their model was weight base non-reliable method for feature extraction. The assigned weights may not adequately reflect the underlying value or relevance of the characteristics if the weight-based method employed for feature extraction is unreliable. This may lead to inaccurate or misleading feature rankings and selections. The resulting subpar performance in future analysis or modelling tasks may be caused by the extracted features failing to accurately reflect the underlying patterns or relationships in the data. This can jeopardize the retrieved characteristics' fairness and interpretability as well as the overall reliability of the findings. Employing strong and established methods that appropriately assess the value and contribution of each feature is essential to ensuring reliable feature extraction in order to prevent these problems. In addition to this, Table 1 provides an overview of additional methods that are also focusing the same issue of sentiment analysis.

Table 1. Summarize the table of literature review.

Ref	Proposed Methodology	Dataset	Result Accuracy	Limitation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

[16]	Empirical Analysis	U.S. commercial bank dataset	79%	Required complex training and large dataset.
[17]	RNN-LSTM algorithm	Movie review datasets, Document Understanding Conferences (DUC) datasets	74.78%	Loss of sequential and contextual information
[18]	CNN-LSTM	93000 tweets (tweeter dataset)	89%	Unable to normalize polarity intensity of words
[19]	R-CNN	IAPS, ArtPhoto, Twitter, Flickr, Instagram	83.05%	Required a more dependable recurrent model
[20]	convolutional neural network	Fashion Dataset	87%	Dictionary-based and domain-specific
[21]	Deep feature modal using regression	Amazon product reviews	83%	Fail to produce multi-dimension features
[22]	multi-modal event-aware network	Weibos posts	89%	Unable to manage out-of-vocabulary words
[23]	Image-Text Sentiment Analysis model	Twitter dataset, Getty dataset.	76% accuracy for Twitter data and 87% accuracy for Getty photos.	Costly and time-consuming to search for opinionated words
[24]	AHRM	Flicker8k dataset, Getty dataset.	87%	Features are lost during extraction
[25]	Attention-Based Modality-Gated Networks for Image-Text Sentiment Analysis	Flicker8kW, Twitter	88% accuracy on FlickrW images and gives 79% accuracy on Twitter data	Weight base non-reliable method for feature extraction.

From the above discussion, it has been concluded that most research focuses on text or images and may ignore multus reviews. As multus medium reviews contain more information and help the DSS to make a better decision, and may increase the satisfaction level of customers. In this research, A Multus medium opinion mining (MMOM) approach has been proposed for product recommendation to focus on the Multus-Medium.

3. Research Methodology

This section presents the methodology of the proposed MMOM (Multus-Medium Opinion Mining) framework for product recommendation. It consists of several stages to ensure comprehensive and accurate recommendations. The methodology includes the following steps.

- The methodology begins with aspect and opinion extraction, followed by aspect refinement and identification. Data collection is conducted from Twitter and Flickr8k datasets, providing a diverse range of reviews and images related to products.
- Feature extraction involves semantic attention for text-based feature extraction and visual attention models for image-based feature extraction. Text-based features are extracted using BiLSTM, employing techniques like data preprocessing, lowercasing, stemming, and tokenization.
- Simultaneously, visual attention models utilize GoogleNet and VGGNet architectures to extract low-level features from product images. These pre-trained models capture visual characteristics such as product quality and price.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

- The fusion of extracted features combines the results from semantic and visual attention models, creating a unified feature vector set. These fusion harnesses the strengths of both textual and visual information for more accurate recommendations.
- Reinforcement learning is applied to further enhance the recommendation system. It models the recommendation process as a Markov decision process, continuously adapting recommendations based on user feedback and behavior.

Next subsection provides details of each of the aforementioned methodology phases.

3.1. Data Set Description

Two large-scale social Multus medium datasets Twitter [26] and flicker8k [27], have been used in this study. Table 2 presents the details of the dataset. Each dataset includes classes labeled as positive, negative, or neutral. As in [28] for Twitter, Twitter API[1] has been used to obtain a substantial volume of Tweets. Tweets with a picture in addition to English text are kept as to 3.4 million, and 4 million photographs were obtained. Tweets that did not contain any static images, videos, and/or animated GIFs) were eliminated. Moreover, tweets not written in English and may have fewer than five words were also discarded. As far as the Flickr8k[2] dataset is concerned, there are 8092 JPEG (Joint Photographic Experts Group) photos overall in Flickr8k, of different sizes and forms. 2000 photographs were used for testing while 6092 are used for training. Additionally, it includes text files with a total of 40460 captions, five captions per image.

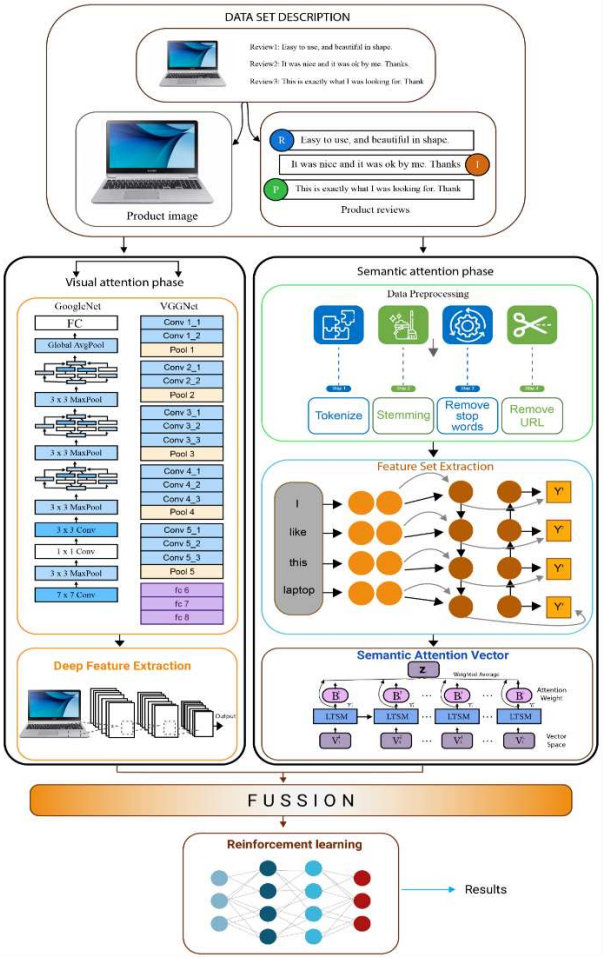


Figure 4. Proposed MMOM for Product Recommendation.

[1] http://www.t4sa.it/dataset/t4sa_all_imgs.tar

[2] https://www.kaggle.com/datasets/adityajn105/flickr8k

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

Table 2. Statistics of the dataset.

Dataset	Positive	Neutral	Negative	Sum
Twitter (t4sa)	372,904	444,287	156,862	974,053
Flicker8k	3,999	1,711	2,382	8,092

3.2. Semantic Attention Phase for Feature Extraction

To improve the entire customer experience and guide business decisions, feature extraction from reviews is essential since it identifies the salient features or characteristics of a good or service that customers mention in their reviews. We extract the feature to reduce the computational complexity. In this stream, text-based features were extracted using BiLSTM.

3.2.1. Data Preprocessing

Data processing aims to shape the raw data into a comprehensive format (as depicted in Figure 5). It includes lowercasing, stemming, tokenization, standardization, eliminating insignificant substance, and literal interpretation of text reviews.

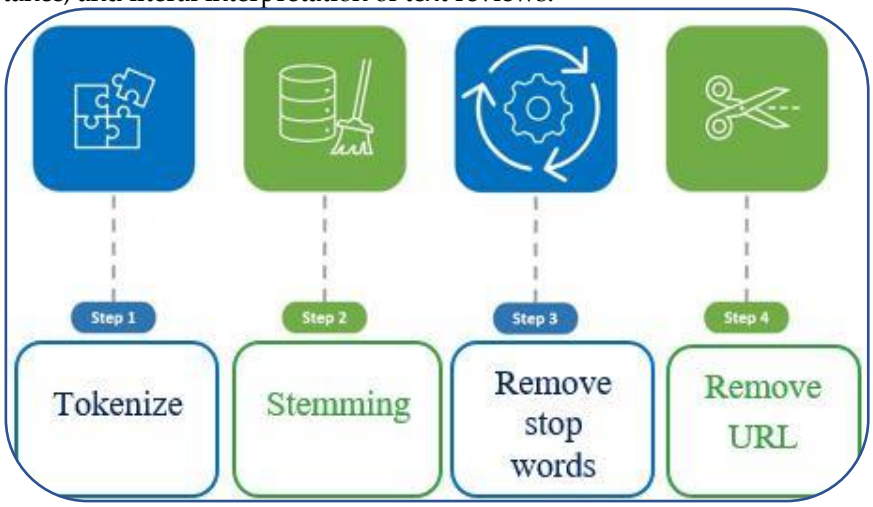


Figure 5. Data Preprocessing.

Algorithm 1 expressed the preprocessing phase. The algorithm provides detailed descriptions of each step. These processes are designed to address issues such as changing the meaning of a sentence from positive to negative due to the inclusion of words like "NOT." For example, the phrase "the car is not comfortable" would be modified through these processes to eliminate the negative connotation associated with the word "NOT."

Algorithm1. Data Preprocessing	
1.	Input: DS_i : A data set of reviews
2.	Output: W^{all} : All preprocessed words
3.	Initialize $W^{all} \leftarrow \Phi$
4.	For Each DS ε DS_i
5.	String St ⟵ tokenize (DS_i)
6.	St' ⟵ lowercase (DS_i)
7.	End For
8.	For Each S_i ε St
9.	St'' ⟵ normalize (S_i)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

10.	St''' ← stem (St'')
11.	St''''' ← transliteration (St'''')
12.	End For
13.	Return W^{all}

In this phase extraction of features from the text has been performed that helps the MMOM to efficiently categorize the contents of data.

3.2.2. Feature Set Extraction

Most studies have employed syntactic structure [29], language patterns [30], and dependency relation [31] to accomplish this goal. Incredible as these techniques are but they may produce poor results due to complex review structure. To provide a unified set of features, a BiLSTM model is used in this research. Due to its dependency parsing strategy, it produces far better results as compared to traditional approaches. Optimal patterns for discovering aspects (features) and opinion terms have been developed using BiLSTM. The issue of the bidirectional situation has also been addressed using a "forward/reverse" based programming method, with a predefined tag such as "neg" being used to indicate the direction of the reverse operation. The feature set extraction is graphically represented in Figure 6. This adjustment accounts for most phrases, for example, in the statement "I like this laptop," numerous adjectives that aim at several nous are also dealt with. BiLSTM provides result positive after the implementation of the phrase [32] [33]. The general equation of BiLSTM for understanding the output is as:

$$y_i = BiLSTM(\alpha_t, \beta\{t-1\}) \quad (1)$$

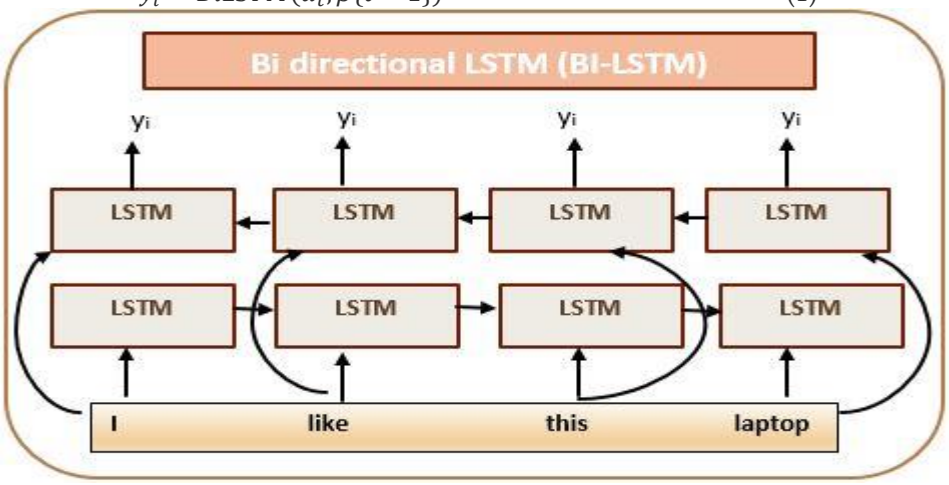


Figure 6. Explore the feature set using Bilstm.

Text semantic Attention (Text Vectors Formation)

Certain words in the text are crucial while conveying the author's intended tone. From previous studies, it has been observed that the semantic attention mechanism is beneficial for numerous assignments and tasks during natural language processing [34,35]. In this work, an end-to-end process using a semantic attention model for sentiment classification has been developed that directly emphasizes the most crucial words.

Let $R = \{R1, R2, \ldots, Ri, \ldots, Rn\}$ denotes a set of n number of text-based reviews. For each review Ri , the words are plotted into a vector space using an embedding matrix $VI = \{V_i^1, V_i^2, \ldots, V_i^j \ldots V_i^L\}$ as follows:

$$VI = RiMe,\ Vi \in \mathbb{R}^{LXE} \quad (2)$$

where L is the length of the text, E represents the dimension of the word embeddingV_i^j, and Me is the parameter matrix. The symbol $\mathbb{R}$ represents the real numbers. The BiLSTM is then used to learn the more abstract level feature.$Yi = \{y_i^1, y_i^2, \ldots, y_i^j \ldots y_i^L\}$:

$$Yi = fv(Zi, \Theta v), Yi \in \mathbb{R}^{LXB} \quad (3)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

where θv represents BiLSTM parameters and B represents BiLSTM cell size. The importance of each word y_i in the sentiment classification is weighted and given an attention score β_i^j analogous to visual attention, which is mentioned below:

$$\beta_i^j = \exp(e_i^j)/\sum^L{}_{j=1}\exp(e_i^j) \quad (4)$$

where

$$e_j^i = \sigma\left(My_i^j + b\right) \quad (5)$$

How well the word $\{y_i^j\}$ fits the emotion is quantified by the normalized attention score. It is necessary to learn two parameters, the weight matrix M and the bias term b. The activation function σ (.) is typically a $tanh$ function since it is nonlinear. To normalize focus across all the phrases $\{y_i^j\}$ 1≤j≤L, we employ β_i^j. The weighted average of the word features can be used to derive the attended semantic features:

$$Z_i^K = \sum_{1\le j\le L}\beta_i^j y_i^j, Zi\in\mathbb{R}^B \quad (6)$$

We refer to the full procedure for creating the textual features that are attended to as follows:

$$Z^k = f_a\left(Yi, {\Theta_a}^{(t)}\right), Z_x\in\mathbb{R}^B \quad (7)$$

where $\Theta_a^{(t)}$ is the weight parameters that consist M and b in Eq (5). Emotion-related text features are also better represented by the attended semantic feature mapping Zi.

3.3. Visual Attention Model for Image-based Feature Extraction

This phase involves extracting distinct low-level features from images using two renowned CNN architectures, namely GoogleNet [36] and VGGNet [37].

3.3.1. Feature Extraction using GoogleNet and VGGNet

Initially, a CNN architecture is employed independently for feature extraction, and then the two are integrated into a fully linked and classified layer. Several picture features may be included in the merged set. Specifically, the suggested framework makes use of the two most recent and cutting-edge deep CNN architectures: GoogleNet [38] and Visual Geometry Group Network (VGGNet) [39] is adopted as feature extractors for the classification of product recommendations. Firstly, these architectures are pre-trained on a diverse collection of generic product photos, and image feature extraction is performed. Each adopted CNN architecture's rudimentary design is laid forth below.

GoogLeNet

GoogleNet (Inception-v1) is a deep convolutional neural network architecture developed by Google researchers in 2014 for image categorization. GoogleNet's usage of the Inception module, which enables efficient use of computing resources while retaining high accuracy, is one of its defining characteristics [40]. Figure 7 is a detailed illustration of the inception module. GoogLeNet highlights the connection between neighboring picture pixels. In this research, GoogleNet is used to manage the dataset's dynamic characteristics.

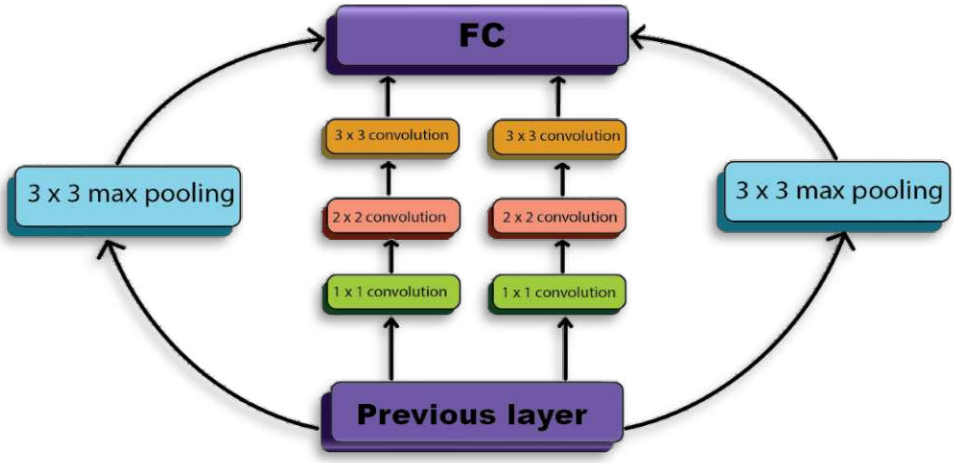


Figure 7. GoogLeNet Internal Structure.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

GoogLeNet employs 1x1 convolutions to increase the network's depth and breadth while reducing its size without compromising performance. 1x1 convolution may help reduce the model size, lowering the overfitting problem. The inception block calculates cross-channel correlations through a 1x1 convolution. Next, cross-spatial and cross-channel correlations are performed using a 3x3 filter. Using an inception module with many convolution filters, GoogleNet extracts numerous layers of features from the same input. The results of these 1x1, 2x2, and 3x3 convolutions are then combined. GoogLeNet eliminates many parameters by using average pooling instead of Fully Connected layers. It merely comprises four million parameters.

VGGNet

In 2014, researchers at the University of Oxford proposed the VGGNet (Visual Geometry Group Network) deep convolutional neural network design. The primary function of VGGNet in image extraction is to extract high-level, task-relevant characteristics from pictures [41]. Figure 8 displays the VGGNet's internal structure. In this research, VGGNet is used to manage the static characteristics of the dataset.

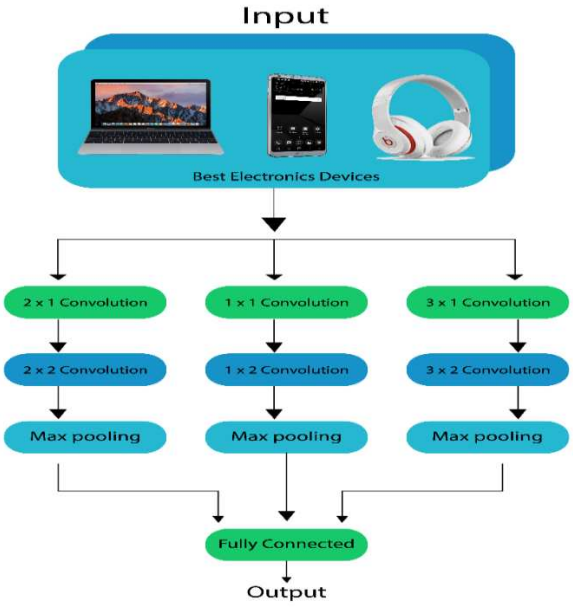


Figure 8. VGGNet Internal Structure.

The VGGNet architecture is a 3x3 filter stack. Utilizing 3x3 filters instead of enormous filters significantly reduces the number of parameters. In one of the configurations, 1x1 convolution filters with a fixed 1-pixel convolution stride are used to modify the input channels linearly. After convolution procedures, spatial padding is employed to retain spatial resolution. Spatial pooling is done using five max-pooling layers with a 2x2 pixel window and a 2-pixel stride. A sequence of convolutional layers, three Fully Connected layers, and a soft-max layer compose the architecture. VGGNet is highly appealing because of its uniform architecture.

Embedding GoogLeNet and VGGNet

The dataset's product image properties are obtained using pre-trained GoogLeNet and VGGNet Convolutional Neural Networks. Early network layers extract attributes such as product quality, price, and caption. The subsequent layers extract additional properties from the created feature maps by the previous levels. Depending on the number of filters and filter size in each layer, various parameters are computed for each layer. GoogLeNet and VGGNet operate with fewer parameters than other ConvNets due to their inception modules. To illustrate the sequential outcomes of a product image through the proposed network, we provide an example, as shown in Figure 9. Upon being processed, the image can be viewed as a three-dimensional matrix. Convolution is performed using a 3x3 matrix filter on the input image, resultantly getting converted into a newer matrix.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

Subsequently, all relevant information is aggregated before extracting product attributes to enhance computational efficiency.

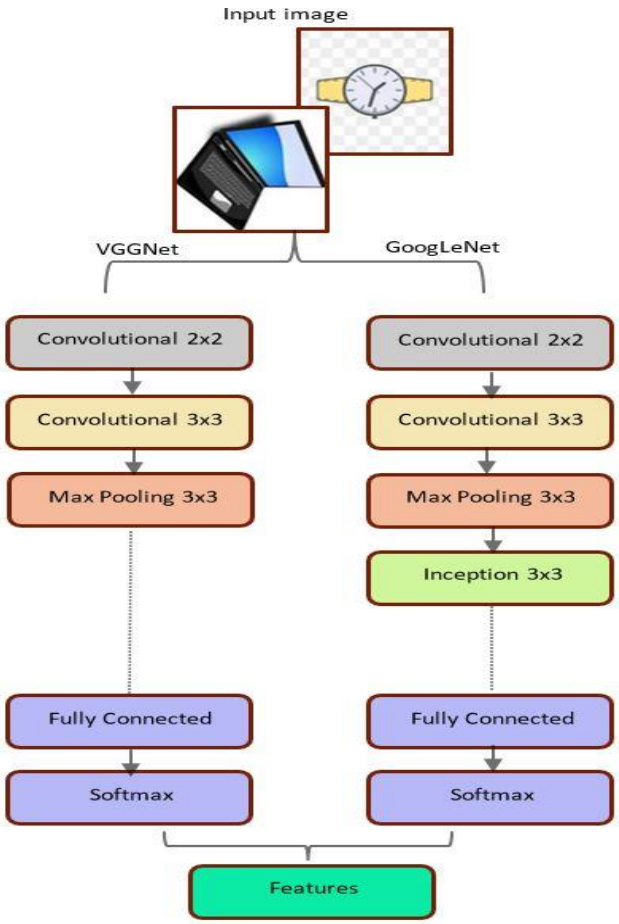


Figure 9. Embedded GoogLeNet and VGGNet Internal Structure.

3.3.2. Deep Feature Extraction

The suggested framework uses the shared characteristics of two different CNN architectures (GoogleNet and VGGNet). Two distinct CNN architectures are learned from example data. This allows the architecture to pick up the generic features from various data sets without additional training. To classify cancerous and healthy cells using average pooling classification, the fully connected layer combines the number of characteristics retrieved independently from the corresponding CNN architecture.

3.3.3. Applying Fusion

Fusion refers to the procedure of amalgamating the results obtained from multiple features that are extracted from text and images. Sentiment analysis in [42,43] used both text and images, with fusion used to integrate the results from textual characteristics and visual features. To conclude, we merged the results of the visual attention model and the semantic attention model. To make a unified feature vector set for learning. When classifying multimodal sentiment, MMOM is offered as a fusion. The model computes the final sentiment score using the following equation [23].

$$Z_i = \begin{cases} \frac{1}{1+\alpha+\sigma}\left(Z_i^{(m)} + \alpha Z_i^{(v)} + \sigma Z_i^{(t)}\right) & input: (V_i, T_i) \\ Z_i^{(v)} & input: (V_i) \\ Z_i^{(t)} & input: (T_i) \end{cases}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

where α and δ are hyper-parameters that determine how much emphasis should be placed on certain classifiers. When the Multus medium input is restricted to an image (Vi), the sentiment score of the visual attention model (Z (v) i) is directly utilized to compute the final sentiment score (Zi).

3.3.4. Applying Reinforcement Learning

Two main approaches for modeling traditional recommender systems are collaborative filtering and content-based strategies. Collaborative filtering-based systems rely on the "user-item interaction matrix" that stores records of users' past interactions with items to generate recommendations. Collaborative filtering algorithms can be constructed using memory-based or model-based techniques. Memory-based techniques determine the most similar user to a new user and suggest their preferred items, but they cannot quantify variation or bias. In contrast, model-based approaches construct a generative model on top of the user-item interaction matrix to predict the preferences of new users. Still, they can be subject to model bias and volatility. Collaborative filtering methods rely heavily on user-item interactions to generate recommendations, which can be problematic when data on users or products is limited. There may need to be more data for new customers or products to predict their preferences accurately. Collaborative filtering approaches are also vulnerable to sparsity.

Content-based recommendation systems consider user-item interactions, user preferences, and additional product-related characteristics, such as popularity, description, and purchase history. These user and product characteristics are fed into a model that functions similarly to a standard machine learning model with error optimization. This model includes more descriptive information about the product but has higher bias and lower variance than other modeling techniques. The goal of content-based systems is to suggest the most suitable product to the consumer to generate positive product reviews.

With the Markov property included in reinforcement learning models, recommendation systems are constructed effectively. The issue of reinforcement learning may be phrased with the product as the state, the action as the next best product to be suggested, and the reward as user satisfaction/ conversion or review. Using a reinforcement learning model that balances exploration and exploitation, we discover an additional benefit. In MMOM, fusion combines the findings of textual characteristics and visual features to create the final dataset for reinforcement learning. Customers are provided with product suggestions using reinforcement learning. In addition to suggesting the product that consumers may find most beneficial, the algorithm will also promote random items, generating new interest in them, as seen in Figure 10. The Reinforcement Learning model will also be continually learning, which means that when the user's interests change, the model's product recommendations will also change, so making the model resilient.

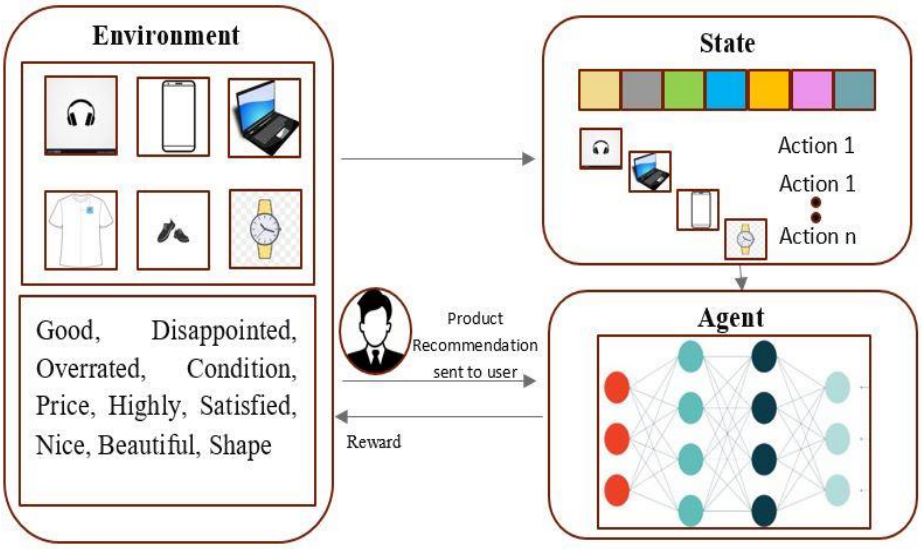


Figure 10. Reinforcement Learning Structure.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

4. Experimental Results

Examining the results of experiments while assessing the success, efficiency, and effectiveness of the proposed method is the focus of this subsection. The suggested System's precision and efficacy have been examined through experiments. It has been decided to put the proposed System through its paces using five benchmark datasets selected from previously published publications. An extensive experimental evaluation confirmed that the propositioned strategy outperformed contemporary and up-to-date alternatives. Each experiment's specifics are described in the subsection.

4.1. Performance Evaluation Measure

Precision, recall, F1 score, and ultimate accuracy are utilized as standard benchmarks to assess the effectiveness of the Multus-medium sentiment analysis. The suggested measure's numerical formulas are mentioned in Equations (8), (9), (10), and (11), where (TP, TN, FN, and FP) stand for (True Positive), (False Negative), and (False Positive), respectively.

$$precision = \frac{TP}{TP + FP} \quad (8)$$

$$recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1\ score = \frac{2.Precicion * Recall}{Precision + Recall} \quad (10)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

Whereas it has been common practice to use a Receiver Operating Characteristic (ROC) Curve to assess a classifier's efficiency. (11) Shows the formula to for performance indicator.

$$ROC = \frac{CP(\frac{i}{positve})}{CP(\frac{i}{negative})} \quad (12)$$

Conditional probabilities, where l is a class label, are calculated using the formula CP (i/l). The results of a classification are shown on a ROC curve [44] from most positive to most negative

4.2. Outcome of the Experiment

The effectiveness of the suggested method is demonstrated by contrasting it with preexisting models. However, the discussion primarily focuses on presenting the results without providing a deep technical discussion or sufficient insights from the results. To address this, we will delve into the underlying reasons and implications of the findings in greater detail.

4.2.1. Baselines

To determine how well the proposed model functions, we utilized reference models and descriptive data presented in the table below as benchmarks. However, it is important to provide a more comprehensive analysis and insights regarding the performance of the proposed model compared to the baselines.

- AHRM [24]: They use a multi-modal (positive, negative) analysis of social images with the help of an attention-based heterogeneous relational model (AHRM).
- DMAF [23]: They proposed a new method of multimodal attentive fusion (DMAF) analysis for the sentiment. This technique can collect supplementary and no redundant information for more accurate sentiment categorization by automatically drawing attention to locations and words that are associated with affection.
- AMGN [25]: They proposed a new approach called "Attention-Based Modality-Gated Networks" (AMGN) to leverage the correlation between the modalities of images and texts, intending to extract discriminative features required for multimodal sentiment analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

In Table 3, the MMOM model consistently achieved higher F1 scores and accuracies compared to the other models on both the Flicker8k and Twitter datasets. This indicates that the proposed MMOM model demonstrated superior performance in sentiment analysis.

Table 3. Experimental results of baselines.

Datasets	Models	F1 score	Accuracy
Flicker8k dataset	AHRM	87.1%	87.5%
	DMAF	85%	85.9%
	MMOM	88.75%	90.38%
Twitter(t4sa)	DMAF	76.9%	76.3%
	AMGN	79.1%	79%
	MMOM	86.34%	88.54%

For the Flicker8k dataset:

- AHRM model achieved an F1 score of 87.1% and an accuracy of 87.5%.
- DMAF model achieved an F1 score of 85% and an accuracy of 85.9%.
- MMOM model achieved the highest F1 score of 88.75% and the highest accuracy of 90.38%.

For the Twitter (t4sa) dataset:

- DMAF model achieved an F1 score of 76.9% and an accuracy of 76.3%.
- AMGN model achieved an F1 score of 79.1% and an accuracy of 79%.
- MMOM model achieved an F1 score of 86.34% and an accuracy of 88.54%.

Here are some possible reasons for the superior performance of the MMOM model:

Leveraging embedded GoogLeNet and VGGNet with BiLSTM: The MMOM model utilized embedded GoogLeNet and VGGNet with BiLSTM (Bidirectional Long Short-Term Memory) as part of its architecture. This combination of models and techniques might have helped capture more comprehensive and relevant features from the input data, leading to improved sentiment analysis performance.

Attention-based mechanism: The MMOM model might have employed an attention-based mechanism that focuses on relevant aspects or regions of the input data, allowing it to better extract discriminative features for sentiment analysis. This attention mechanism could have contributed to the model's superior performance compared to the other attention-based models like AHRM and DMAF.

Aspect term position prediction and semantic similarity: The MMOM model may have incorporated the prediction of aspect terms' position and considered semantic similarity between aspects in sentiment analysis. By taking into account these additional factors, the MMOM model could have gained a better understanding of the sentiment expressed towards specific aspects, resulting in more accurate sentiment analysis.

Integration of external knowledge: The MMOM model might have integrated external knowledge from pre-trained language representation models like GoogLeNet and VGGNet. This incorporation of external knowledge could have provided supplementary information that merged with semantic features, leading to enhanced performance in sentiment analysis tasks.

Overall, the combination of embedded GoogLeNet and VGGNet, the attention-based mechanism, aspect term position prediction, semantic similarity consideration, and the integration of external knowledge could have contributed to the superior performance of the proposed MMOM model compared to the other models mentioned in the table. Further detailed analysis and experiments can help provide a more comprehensive understanding of the specific advantages and strengths of the MMOM model.

The Multus-medium approach was utilized to perform opinion mining of product reviews, whereas results are depicted in Table 3. The proposed model exhibited strong performance compared to other state-of-the-art methods regarding the accuracy and F1 score, while other models also demonstrated competitive performance. The F1 score and Accuracy of our proposed model is much better than the above-mentioned baseline models. The reason behind this better result is GoogleNet

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

with VGGNet embedding through which a merged feature-embedded CNN and integrated supplementary information.

The main results indicate that the performance of GoogleNet embedding can be improved by incorporating external knowledge, which leads to positive outcomes. However, more than using syntactic knowledge is required to enhance aspect-based Multus-medium sentiment analysis. Instead, a combination of external knowledge from pre-trained language representation models like GoogleNet and VGGNet and additional knowledge signals can create auxiliary information that merges with semantic features and enhances performance.

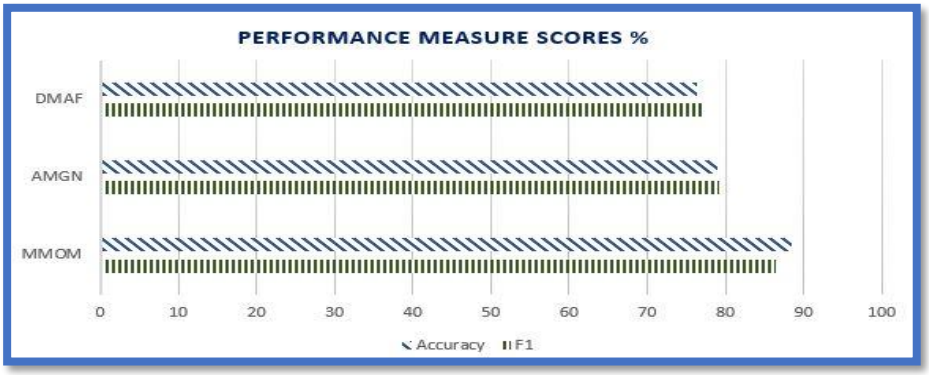


Figure 11. Performance of the purposed model on the Twitter Dataset.

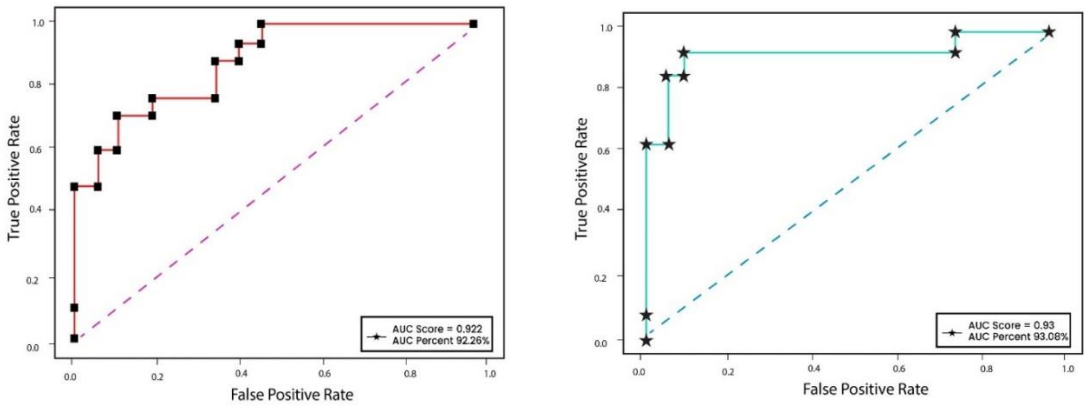
Figure 12. Performance of the purposed model on Flicker8k Dataset.

We conducted another experiment to compare our proposed MMOM model with two baseline methods using the flicker8k and Twitter datasets. Our model, which utilizes embedded GoogLeNet and VGGNet with BiLSTM, outperformed other attention-based models in sentiment analysis. Furthermore, our proposed approach improved the F1 score and accuracy over non-transformer-based baselines. These results indicate that predicting aspect terms' position and semantic similarity between aspects significantly impact sentiment analysis. Figures 11 and 12 visualize the comparison of accuracy and F1 score of the Twitter and flicker8k datasets, respectively.

4.2.2. Performance of Purposed Research

Figure 13 displays the True Positive Rate (TPR) and False Positive Rate (FPR) for the study. Each dataset's TPR and FPR values are shown on a ROC curve. The area covered by the roc curve is 0.92 for dataset 13a and 0.93 for dataset 13b.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

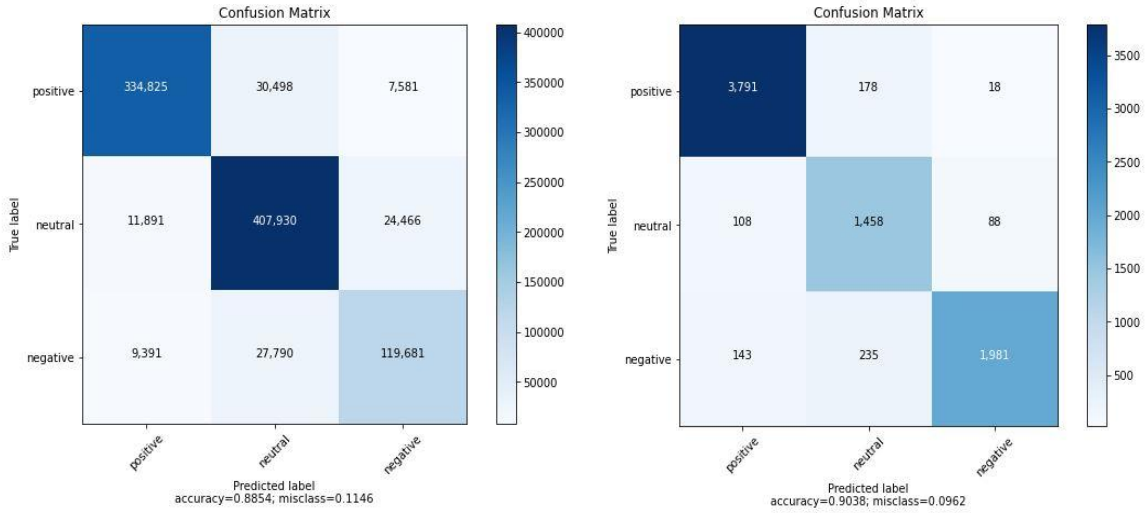


(**a**) (**b**)

Figure 13. ROC curve on each dataset. (**a**) Twitter. (**b**) Flicker8k.

To further assess the effectiveness of the proposed technique, a ROC curve was used to analyze True Positive Rate (TPR) and False Positive Rate (FPR) for each dataset (Figure 13). The proposed classification approach demonstrated an overall correct prediction rate of 88.54% and 90.38% on the Twitter and flicker8k datasets, respectively. However, it is crucial to provide a deeper technical discussion and offer insights into these performance measures and their implications.

A performance-based matrix, a confusion matrix, has been created to demonstrate the effectiveness of the proposed technique using TP, TN, FP, and FN Figures 14 and 15.



(**a**) (**b**)

Figure 14. (**a**) Confusion Matrix on the Twitter dataset, (**b**) Confusion Matrix on Flicker8k dataset.

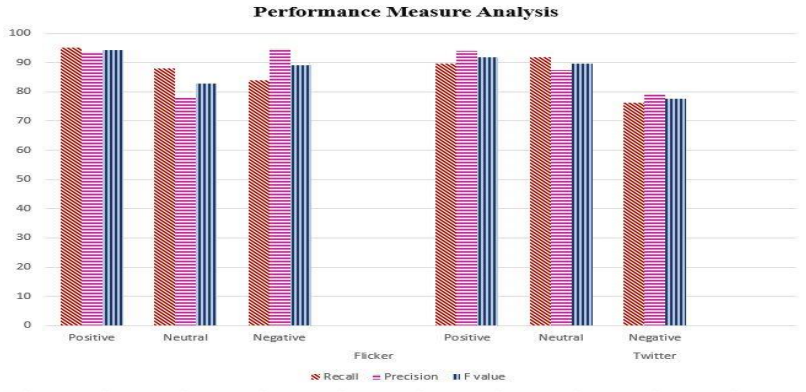


Figure 15. Performance measure analysis on each dataset.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0405.v1

It display two different combinations of predicted and actual values. The correct prediction rate for the proposed system is 88.54% in Figure 14 and 90.38% in Figure 15. These results indicate that the proposed classification approach performs well and produces better outcomes. Additionally, precision, recall, F1 measure, and accuracy have been employed to determine the proposed model's effectiveness. The flicker8k dataset's macro results demonstrate 88.88% precision, 89.04% recall, 88.75% F value, and 90.38% accuracy, supporting the proposed work's superior efficiency. The proposed work's results on the Twitter dataset show 86.80% precision, 85.97% recall, 86.34% F value, and 88.54% accuracy, as shown in Table 4.

A confusion matrix and performance-based matrix were employed to demonstrate the effectiveness of the proposed technique using TP, TN, FP, and FN. Figures 14 and 15 display different combinations of predicted and actual values. The proposed system achieved 88.54% precision, 85.97% recall, 86.34% F1 value, and 88.54% accuracy on the Twitter dataset (Table 4). Similarly, on the flicker8k dataset, the proposed system demonstrated 88.88% precision, 89.04% recall, 88.75% F1 value, and 90.38% accuracy. However, it is important to provide a deeper technical discussion and elaborate on the implications of these performance measures.

Table 4 summarizes the performance of the proposed model in terms of recall, precision, and F value for each sentiment class within the Flicker8k and Twitter datasets. These results provide insights into how well the model can accurately classify text into different sentiment categories, indicating its effectiveness in sentiment analysis tasks on these specific datasets.

Table 4. Proposed Model output.

Dataset	Class	Recall	Precision	F value
Flicker8k	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
Twitter	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

For the Flicker8k dataset:

- Positive class: The proposed model achieved a recall of 95.08%, indicating that it correctly identified a high percentage of positive instances. The precision for the positive class was 93.79%, indicating a high accuracy in predicting positive instances. The F value for the positive class was 94.43%, reflecting a good balance between precision and recall.
- Neutral class: The model achieved a recall of 88.15% and a precision of 77.93% for the neutral class. The F value for the neutral class was 82.72%.
- Negative class: The proposed model achieved a recall of 83.89% and a precision of 94.92% for the negative class. The F value for the negative class was 89.11%.
- Average: The average recall across all classes was 89.04%, the average precision was 88.88%, and the average F value was 88.75%.

For the Twitter dataset:

- Positive class: The proposed model achieved a recall of 89.79% and a precision of 94.02% for the positive class. The F value for the positive class was 91.86%.
- Neutral class: The model achieved a recall of 91.82% and a precision of 87.5% for the neutral class. The F value for the neutral class was 89.61%.
- Negative class: The proposed model achieved a recall of 76.3% and a precision of 78.9% for the negative class. The F value for the negative class was 77.57%.
- Average: The average recall across all classes was 85.97%, the average precision was 86.80%, and the average F value was 86.34%.

These results demonstrate the performance of the proposed model for different sentiment classes within each dataset. The model achieved high recall and precision values for positive instances in

 doi:10.20944/preprints202306.0405.v1

both datasets. It also showed varying levels of performance for the neutral and negative classes. The average performance across all classes indicates the overall effectiveness of the proposed model in sentiment analysis for both the Flicker8k and Twitter datasets.

Based on the results, it can be inferred that the proposed system achieves greater accuracy and lower loss during both the training and validation phases. This indicates that the proposed approach is capable of accurate classification with minimal error. However, it is necessary to provide a more in-depth technical discussion to explain the reasons behind these findings and discuss their implications.

Figure 15 provides a detailed analysis of performance measures, including Precision, Recall, F1 score, and Accuracy, across various datasets. However, it is recommended to provide more insights and discuss the implications of these performance measures in greater detail.

4.3. Discussion

The experimental analysis aims to improve the comprehension of semantic vectors and refine BiLSTM for sentiment classification. The study utilizes GoogleNet and VGGNet for deep feature extraction and integrates external knowledge into the MMOM model to aid in auxiliary information, ultimately improving sentiment analysis performance. The research indicates that additional knowledge can yield better results. However, integrating knowledge into the model is complicated due to the diverse embedding space, which includes image features, text words, and distance vector entities. A consistent vector space is essential for the model to learn the embedding of GoogleNet and VGGNet knowledge with BiLSTM.

Moreover, sentiment analysis performance depends on contextual knowledge, and inappropriate knowledge incorporation may negatively affect performance. The proposed Multus Medium-based opinion mining injects BiLSTM semantic vectors that may introduce noise and alter the original context of the input vector. Therefore, an efficient solution is required to manage excessive noise that does not change the actual context of the input sentence.

5. Conclusions and Future Work

Product reviews and comments posted online have a significant impact on purchasing decisions and product sales. They also play a role in quality improvement and recommendation systems for new users. In this study, a deep learning approach is applied to conduct sentiment analysis on product reviews. The proposed method utilizes BiLSTM and embedded CNN, incorporating GoogleNet and VGGNet, to develop an enhanced deep learning model. The model includes several critical steps such as data preprocessing, feature set extraction, semantic attention vector, BiLSTM for text analysis, CNN with VGGNet and GoogleNet for image analysis, deep feature extraction for images, fusion, and reinforcement learning.

A Multus Medium Opinion Mining (MMOM) method is proposed, which utilizes both textual and image features to provide better product recommendations based on unique and discriminative features. The MMOM model combines BiLSTM embedded CNN and feature fusion for sentiment analysis, outperforming other models in terms of recommendation accuracy. Experimental results demonstrate the high accuracy, F1 score, and ROC values achieved by the MMOM model. On the Flicker8k dataset, the model achieves an accuracy of 90.38%, an F1 score of 88.75%, and an ROC of 93.08%. On the Twitter dataset, the model achieves an accuracy of 88.54%, an F1 score of 86.34%, and an ROC of 92.26%. These results indicate a significant improvement over the other mentioned techniques, with an accuracy advantage of 7.34% and 9.54%.

The study highlights the significant impact of online product reviews and comments on purchasing decisions and product sales. By conducting sentiment analysis on these reviews, businesses can gain valuable insights into customer sentiments and preferences, enabling them to make informed decisions to improve product quality, marketing strategies, and customer satisfaction. the proposed Multus-Medium approach, particularly the MMOM model, offers a comprehensive solution for sentiment analysis and recommendation systems by leveraging both textual and image features. The experimental results validate its superiority over existing techniques,

providing businesses with valuable insights for informed decision-making and assisting customers in making better purchasing decisions. Future work could focus on adopting more effective decision-making techniques to further improve accuracy. Furthermore, the proposed scheme can be expanded to other sentiment-related tasks such as hospital recommendation systems, crop farming recommendation, and medical diagnostic systems. This research contributes to assisting customers in making informed purchasing decisions.

Funding: The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Horrigan, J. Online shopping, Pew Internet & American Life project. 2018 Washington, DC Available at:< http://www. pewinternet. org/Reports/2008/Online-Shopping/01-Summary-of-Findings. aspx>[Accessed 8/8/2014].
2. Lei, Z.; Yin, D.; Mitra, S.; Zhang, H. Swayed by the reviews: Disentangling the effects of average ratings and individual reviews in online word-of-mouth. *Production and Operations Management* **2022**, 31, 2393-2411.
3. Nawaz, A.; Awan, A. A.; Ali, T.; Rana, M. R. R. Product's behaviour recommendations using free text: an aspect based sentiment analysis approach. *Cluster Computing* **2020**, 23, 1267-1279.
4. Rana, M.R.R.; Nawaz, A.; Iqbal, J. A survey on sentiment classification algorithms, challenges and applications. Acta Universitatis Sapientiae, Informatica 2018, 10, 58-72.
5. Al-Abbadi, L.; Bader, D.; Mohammad, A.; Al-Quran, A.; Aldaihani, F.; Al-Hawary, S.; Alathamneh, F. The effect of online consumer reviews on purchasing intention through product mental image. *International Journal of Data and Network Science* **2022**, 6, 1519-1530.
6. Kurdi, B.; Alshurideh, M.; Akour, I.; Alzoubi, H.; Obeidat, B.; Alhamad, A. The role of digital marketing channels on consumer buying decisions through eWOM in the Jordanian markets. *International Journal of Data and Network Science* **2020**, 6, 1175-1186.
7. Bhuvaneshwari, P.; Rao, A. N.; Robinson, Y. H.; Thippeswamy, M. N. Sentiment analysis for user reviews using Bi-LSTM self-attention based CNN model. *Multimedia Tools and Applications* **2020**, 81, 12405-12419.
8. Rana, M. R. R.; Rehman, S. U.; Nawaz, A.; Ali, T.; Imran, A.; Alzahrani, A.; Almuhaimeed, A. Aspect-Based Sentiment Analysis for Social Multimedia: A Hybrid Computational Framework. *Computer Systems Science & Engineering* **2023**, 46.
9. Khan, A. Improved multi-lingual sentiment analysis and recognition using deep learning. *Journal of Information Science* **2023**, 01655515221137270.
10. Chaudhry, N. N.; Yasir J.; Farzana K.; Zahid M.; Zafar I. K.; Umar, S.; Sadaf H.; J. Sentiment analysis of before and after elections: Twitter data of us election 2020. *Electronics* **2021**, 10, 2082.
11. El-Affendi, M. A.; Khawla A.; Amir H. A novel deep learning-based multilevel parallel attention neural (MPAN) model for multidomain arabic sentiment analysis. *IEEE Access* **2021**, 9, 7508-7518.
12. Gastaldo, P.; Zunino, R.; Cambria, E.; Decherchi, S. Combining ELM with random projections. *IEEE intelligent systems* **2013**, 28, 46-48.
13. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal* **2014**, 5, 1093-1113.
14. Nawaz, A.; Ali, T.; Hafeez, Y.; Rehman, S. U.; Rashid, M. R. Mining public opinion: a sentiment based forecasting for democratic elections of Pakistan. *Spatial Information Research* **2022**, 1-13.
15. Schuckert, M.; X. Liu; R. Law. Hospitality and tourism online reviews: Recent trends and future directions. *Journal of Travel & Tourism Marketing* **2015**, 32, 608-621.
16. Agoraki, M.; Aslanidis, N.; Kouretas, G. P. US banks' lending, financial stability, and text-based sentiment analysis, *Journal of Economic Behavior & Organization* **2022**, 197, 73-90.
17. Abdi, A.; Shamsuddin, S. M.; Hasan, S.; Piran, J. Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. *Information Processing & Management* **2019**, 56, 1245-1259.
18. Onan, A.; Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience* **2021**, 33, e5909.
19. Yang, J.; She, D.; Sun, M.; Cheng, M. M.; Rosin, P. L.; Wang, L. Visual sentiment prediction based on automatic discovery of affective regions. *IEEE Transactions on Multimedia* **2020**, 20, 2513-2525.
20. Tuinhof, H.; Pirker. C.; Haltmeier, H. Image-based fashion product recommendation with deep learning. *In international conference on machine learning, optimization, and data science* **2018**.
21. Meena, Y.; Kumar, P.; Sharma, A. Product recommendation system using distance measure of product image features. *in 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)* **2018**

22. Wang, L.; Guo, W.; Yao, X.; Zhang, Y.; Yang, J. Multimodal Event-Aware Network for Sentiment Analysis in Tourism. *IEEE MultiMedia* **2021**, 28, 49-58.
23. Huang, F.; Zhang, X.; Zhao, Z.; Xu, J.; Li, Z. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowledge-Based Systems* **2019**, 167, 26-37.
24. Xu, J.; Li, Z.; Huang, F.; Li, C.; Philip, S. Y. Social image sentiment analysis by exploiting multimodal content and heterogeneous relations. *IEEE Transactions on Industrial Informatics* **2020**, 17, 2974-2982.
25. Huang, F.; Wei, K.; Weng, J.; Li, Z. Attention-based modality-gated networks for image-text sentiment analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications* **2020**, 16, 1-19.
26. Ghorbanali, A.; Sohrabi, M. K.; Yaghmaee, F. Ensemble transfer learning-based multimodal sentiment analysis using weighted convolutional neural networks. *Information Processing & Management* **2022**, 59, 102929.
27. AL-Sammarraie, Y. Q.; Khaled, A. Q.; AL-Mousa, M. R.; Desouky, S. F. Image Captions and Hashtags Generation Using Deep Learning Approach. *In 2022 International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI)* **2022**, 1-5.
28. Gaspar, A.; Alexandre. L. A. A multimodal approach to image sentiment analysis. in Intelligent Data Engineering and Automated Learning–IDEAL 2019. *20th International Conference* **2019**, 14–16.
29. Wu, S.; Fei, H.; Ren, Y.; Li, B.; Li, F.; Ji, D. High-order pair-wise aspect and opinion terms extraction with edge-enhanced syntactic graph convolution. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2021**, 29, 2396-2406.
30. Ahmed, Z.; Duan, J.; Wu, F.; Wang, J; EFCA: An extended formal concept analysis method for aspect extraction in healthcare informatics. *In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* **2021**, 1241-1244
31. Rana, M. R. R.; Rehman, S. U.; Nawaz, A.; Ali, T.; Ahmed, M. A Conceptual Model for Decision Support Systems Using Aspect Based Sentiment Analysis. *Proceedings of The Romanian Academy Series A-Mathematics Physics Technical Sciences Information Science* **2021**, 22, 381-390.
32. Bhende, M.; Thakare, A.; Pant, B.; Singhal, P.; Shinde, S.; Dugbakie, B. N. Integrating Multiclass Light Weighted BiLSTM Model for Classifying Negative Emotions. *Computational Intelligence and Neuroscience* **2022**.
33. Xiao, Y.; Li, C.; Thürer, M.; Liu, Y.; Qu, T. User preference mining based on fine-grained sentiment analysis. *Journal of Retailing and Consumer* Services **2022**, 68, 103013.
34. Luong, M.T.; Pham, H; Manning, C. D; Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 **2015**.
35. Wang, Y.; Huang, M.; Zhu, X.; Zhao, L. Attention-based LSTM for aspect-level sentiment classification. *In Proceedings of the 2016 conference on empirical methods in natural language processing* **2016**, 606-615.
36. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D. Going deeper with convolutions. *In Proceedings of the IEEE conference on computer vision and pattern recognition* **2015**, 1-9.
37. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 **2014**.
38. Yu, Z.; Dong, Y.; Cheng, J. Research on Face Recognition Classification Based on Improved GoogleNet. *Security and Communication Networks* **2022**, 1-6.
39. Goswami, A. D.; Bhavekar, G. S.; Chafle, P. V. Electrocardiogram signal classification using VGGNet: a neural network based classification model. *International Journal of Information Technology* **2023**, 15, 119-128.
40. Fan, Y.; Wang, Y.; Yu, H.; Liu, B. Movie recommendation based on visual features of trailers. In Innovative Mobile and Internet Services in Ubiquitous Computing: Proceedings of the 11th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS-2017) **2018**, 242-253.
41. Andreeva, E.; Ignatov, D. I.; Grachev, A.; Savchenko, A. V. Extraction of visual features for recommendation of products via deep learning in Analysis of Images, *Social Networks and Texts: 7th International Conference, AIST 2018, Moscow, Russia* **2048**, 5–7.
42. Wang, M.; Cao, D.; Li, L.; Li, S.; Ji, R. Microblog sentiment analysis based on cross-media bag-of-words model. *In Proceedings of international conference on internet multimedia computing and service* **2014**, 76-80.
43. Cao, D.; Ji, R.; Lin, D.; Li, S. A cross-media public sentiment analysis system for microblog. *Multimedia Systems* **2016**, 22, 479-486.
44. Sokolova, M.; Lapalme, G. Classification of opinions with non-affective adverbs and adjectives. *in Proceedings of the International Conference RANLP-2009* **2009**.