SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 2560-2574
Open Access. Online & Print

DOI: 10.5281/zenodo.122.126198

CONTEXT-AWARE SENTIMENT ANALYSIS: A MULTIMODAL APPROACH FOR IMPROVED EMOTION RECOGNITION

Putta Durga[1], Deepthi Godavarthi[2], Sachi Nandan Mohanty[3], Mohammed Altaf Ahmed[4]*, Suleman Alnatheer[5], Qutubuddin Mohammed[6]

[1,2,3]School of Computer Science and Engineering, VIT-AP University, Amaravati, India, pdurga593@gmail.com[1], deepthi.g@vitap.ac.in[2], sachinandan09@gmail.com[3]
Department of Computer Engineering, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Saudi Arabia, m.altaf@psau.edu.sa, s.alnatheer@psau.edu.sa
Department of Electrical Engineering, College of Engineering Wadi Addawasir, Prince Sattam bin Abdulaziz University, Saudi Arabia, q.mohammed@psau.edu.sa

Received: 20/10/2025
Accepted: 01/12/2025

Corresponding Author: Mohammed Altaf Ahmed (m.altaf@psau.edu.sa)

ABSTRACT

Sentiment analysis has come a long way since it was only about classifying text. Now it uses visuals, audio, and contextual signals to provide a better and more accurate picture of how people feel. This study shows a sentiment analysis approach that takes into account the context and uses multimodal learning to improve emotion recognition. The suggested method uses powerful deep learning methods to extract features and classify them across different types of data. Deep learning-based word embeddings like BERT and FastText are used for textual sentiment analysis to get rich linguistic representations and contextual dependencies. We also employ Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorisation (NMF) to find underlying topic structures and pull out useful latent characteristics from text data. For sentiment analysis based on images, we use EfficientNet-B3 because it is better at extracting features and is faster at doing so. Also, hybrid convolutional neural networks (CNNs) are used for both text and image-based emotion recognition, which makes the sentiment analysis technique more complete. The system is thoroughly tested on benchmark datasets, and the results show that it is far better than unimodal sentiment analysis methods. The Hybrid CNN model was able to tell how someone felt about a piece of text with 92.3% accuracy just looking at the text itself. The EfficientNet-B3 model, which is based on images, has an accuracy of 96.7%. This shows that it can pick up on emotional cues from visual data and that deep learning is useful for image-based sentiment analysis. Multimodal context improves the accuracy of categorisation, which makes it beneficial for social media analysis, AI that can understand emotions, and initiating research on sentiment and emotion detection.

KEYWORDS: Multimodal Sentiment Analysis, Emotion Recognition, EfficientNet-B3 for Image Sentiment Analysis, Social Media Analysis, Transformer Models, Cybersecurity, Sustainable Environment.

1. INTRODUCTION

Social web analysis looks at people who make content and participate in discussions. This information is always changing, reflecting both the changing nature of language and the changing emotions of its producers. Social networks are places where people can express themselves in many ways, from simple gestures like clicking "Like" to writing long articles. The things that people share on these networks demonstrate a wide range of public beliefs and points of view.

People contribute user-generated content in a variety of ways, which makes the digital world more interesting and diverse [1]. This content comprises text, pictures, videos, and audio that show a wide diversity of thoughts, experiences, and points of view. Sentiment analysis is a terrific approach to figure out how individuals feel, what they think, and how they act. There are numerous essential uses for it in various areas, such as finding out how happy customers are with a product by reading reviews, working out how people feel about organisations and services, and aiding with mental health therapy by figuring out how individuals are feeling using text analysis [2]. Businesses, researchers, and healthcare professionals may make better decisions and get more people involved with the help of SA. A lot of data is made up of text and pictures combined together, which is a lot of different types of information. This multimodal data contains valuable emotional insights that can assist with sentiment analysis. You can learn how people feel about and see different things by looking at this kind of data [3].

Previous studies that looked at the sentiment of single-modal text mostly used standard statistical methods, which depended a lot on how well the features were extracted [4,5]. For example, Rodger et al. [6] used data envelopment analysis (DEA) as a Word Sense Disambiguation (WSD) method to automatically figure out what words mean in sentences. This made it possible to find out how voters felt about political candidates. The quality of manually chosen features also had a big impact on sentiment analysis of single-modal images, which often resulted in emotional information that was already there. Researchers have come up with new ways to do multimodal sentiment analysis that combine text and images, thanks to advances in machine learning and deep learning. These new methods have solved the problems that older systems faced and made things much more accurate and efficient[7,8].

It's challenging to do sentiment analysis in NLP because it's hard to interpret language [9]. Machine learning is very important for figuring out if a document has a good or negative vibe. There are two main types of these techniques: supervised learning algorithms and unsupervised learning algorithms [10].

Because they are organised but yet varied, social media datasets are great for training and analysing with machine learning. Sentiment analysis also uses rule-based and lexicon-based methods a lot, in addition to machine learning. Multimodal sentiment analysis also looks at and rates important features separately using different classifiers. Some of them are deep neural networks (DNN), artificial neural networks (ANN), multilayer perceptron (MLP), and other deep learning models [11].

This study presents a multimodal sentiment categorisation model that uses a fine-grained attention method to solve these problems. First, because textual data has a lot of noise, a denoising autoencoder is used to find characteristics that keep the sense of the original text better. Second, an improved variational autoencoder with an attention mechanism is used to find features in images that can tell them apart.

Next, a fusion model based on attention mechanisms is suggested to develop a joint representation vector by integrating image and text information in an interactive way. This method lets the model focus on the most important parts of both types of data, which helps it collect and combine important sentiment-related information from both text and images and helps in cybersecurity applications.

This is how the rest of the article is set up: Section 2 gives a short summary of the most recent progress in multimodal sentiment analysis. The suggested multimodal sentiment categorisation approach is detailed in Section 3. The experimental results from two separate Twitter datasets are presented and compared in Section 4. Section 5 concludes the paper by reviewing the key themes and proposing some future research directions for multimodal sentiment analysis.

1.1. Research Objectives

1. To make the analysis more accurate and better at understanding the context, create a hybrid model that incorporates text, photos, and emoticons to figure out how people feel about social media posts.
2. For better aspect-based and emoji-based sentiment identification, use topic modelling (LDA, NMF) and word embeddings (fastText, Word2Vec) to compare classical machine

learning models with deep learning models.
3. To make multimodal sentiment analysis better, use advanced deep learning models like a Hybrid CNN for categorising text and an EfficientNet-B3 for identifying photos.
4. Test how well the multimodal sentiment classification model works by using benchmark datasets like Sentiment140 and Twitter US Airline Sentiment. This will show how well it works in real-world situations, such digital marketing and systems that promote things.

2. LITERATURE SURVEY

Lee et al. [12] employed modality-specific encoders to extract spatiotemporal information, contrastive learning to enhance interactions between modalities, and cross-modal attention mechanisms to integrate features in a meaningful manner. Evaluations of several datasets of emotional responses showed that it worked.

Georgescu et al. [13] built a multimodal emotion recognition system that employs both audio and visual data to help people and computers work together more easily. This is helpful for things like virtual assistants and diagnosing mental health problems. It utilises a CNN to look at speech and a ResNet18-based model to figure out how people feel in photos. The ResNet18 model has been improved with ImageNet. The system was 97.95% accurate, which was better than the standards and individual modalities.

Araiman et al. [14] explained how to use ResNet x-vectors with CNNs to make a complete multimodal emotion recognition system that can read emotions from voice and visuals. A approach that uses windows works with audio, whereas a way that chooses frames works with video. The system was 76.29% accurate on the CREMA-D dataset and 65.67% accurate on the RAVDESS dataset. This is better than the accuracy of each modality on its own.

By combining visual and aural data for continuous emotion analysis, Salas-Cáceres et al. [15] investigated how Human-Machine Interaction (HMI) systems can identify users' moods. It gives the recommendation for an LSTM-based architecture to record temporal dynamics, which outperforms prior methods and achieves the highest accuracy on RAVDESS (88.11%), SAVEE (86.75%), and CREMA-D (80.27%).

By creating a personalised multimodal emotion prediction system that uses an emotionally attuned Reinforcement Learning (RL) agent to aid people suffering from anxiety and depression, Pathirana et al. [16] improved mental health support. In order to provide individualised Cognitive Behavioural Therapy (CBT) exercises, the system analyses facial expressions, vocal tones, and text with corresponding accuracies of 72%, 73%, and 86%. The effectiveness of the approach was evaluated in a Sri Lankan setting using the DASS-21 questionnaire. The results showed that the study group had significantly lower levels of anxiety (from 19.85 to 10.46) and depression (from 21.08 to 13.54). Mental health is greatly improved and tailored mental health care is made possible by combining multimodal emotion prediction with reinforcement learning-driven cognitive behavioural therapy.

The Multimodal Emotion Recognition in Conversations (MERC) calibration framework, CMERC, was created by Tu et al. [17]. It improves the performance of the model without changing its structure. The method to improve the representation of speech is hybrid supervised contrastive learning and curriculum learning for adaptive learning from unknown samples. When samples are mistakenly marked as doubtful, a confidence limit penalises them, which enhances reliability and generalisation.

Emotions such as wrath, fear, joy, and sadness can be found in MAViT-Bangla, a multimodal Bangla dataset created by Das, A. et al. [18]. The collection contains 1002 audio, video, and text samples. In order to achieve an F1-score of 0.64, which is higher than traditional unimodal methods, the suggested AVaTER framework uses cross-modal attention to improve feature interaction across modalities.

Emotion theories, response systems, and many modalities, such as subjective experience, physiology, and behaviour, were all part of the multimodal emotion recognition study by Ramaswamy et al. [19]. By reviewing 179 papers published between 2017 and 2023, it identifies developments, trends, and issues in several fields, such as emotion elicitation, data processing, cultural influences, feature extraction, and fusion approaches. In addition to outlining potential future research directions, it discusses the field's advantages and disadvantages.

Figure 1 shows how the accuracy of several multimodal sentiment analysis approaches on the Memotion Dataset has changed over time. It shows how transformer-based models, early fusion, late fusion, large language models (LLMs), and semi-supervised learning techniques have improved. The x-axis represents the years (2015, 2020, 2025, and 2030), and the y-axis indicates the accuracy of the sentiment analysis in percentage terms. This comparison shows how important sophisticated

deep learning architectures and fusion approaches are for improving sentiment analysis performance. These improvements will lead to more context-aware and reliable emotion detection systems in the future.

Table 1 illustrates the literature survey on a multimodal approach for improved emotion recognition.

Table 1: Literature Survey on a Multimodal Approach for Improved Emotion Recognition.

Reference	Title	Methodology	Findings
[20]	A review of affective computing: From unimodal analysis to multimodal fusion	Survey on Affective Computing Approaches	Discusses multimodal fusion techniques for sentiment analysis
[21]	Memory fusion network for multi-view sequential learning	Memory Fusion Network (MFN)	Achieves better sentiment analysis by fusing multimodal features sequentially
[22]	Conversational memory network for emotion recognition in dyadic dialogue videos	Conversational Memory Network	Enhances emotion recognition by modelling interactions between speakers
[23]	EmotiCon: Context-aware multimodal emotion recognition using Frege's principle	Context-aware emotion recognition model	Improves contextual understanding in multimodal sentiment analysis
[24]	Speech emotion recognition using a multi-hop attention mechanism	Multi-hop attention mechanism	Enhances emotion recognition accuracy in speech data
[25]	End-to-end multimodal emotion recognition using transformers	Transformer-based deep learning model	Achieves state-of-the-art performance in multimodal sentiment recognition
[26]	Multimodal sentiment analysis using deep learning and EEG signals	Deep learning with EEG signals	Integrates EEG data to improve emotion recognition accuracy
[27]	Self-supervised learning for multimodal emotion recognition	Self-supervised learning	Reduces dependency on labeled data for multimodal emotion classification
[28]	Tensor fusion network for multimodal emotion recognition	Tensor Fusion Network	Enhances multimodal feature fusion for emotion recognition

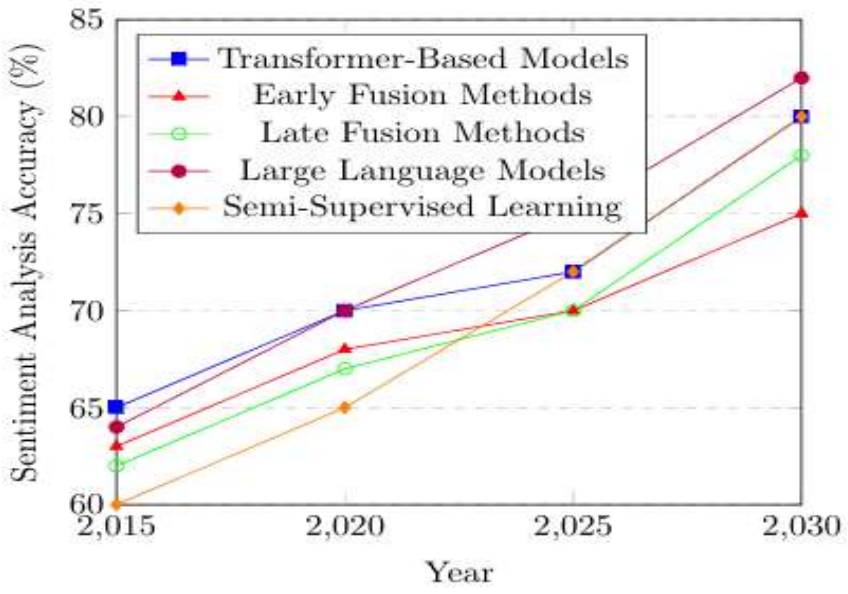

Figure 1: Sentiment Analysis Accuracy Trends of Different Methodological Categories (2015–2030).

3. PROPOSED METHOD

3.1. Data Set

Memotion Dataset 7k The investigations utilised 6,992 Internet memes from the SemEval-2020 Memotion Analysis dataset. Each sample for multimodal sentiment analysis has an image and meme text in it. For sentiment categorisation, the original labels were grouped into three groups: positive (including strongly positive), neutral, and

negative. The final class distribution has 4,160 positive examples (59.50%), 2,201 neutral samples (31.48%), and 631 negative samples (9.02%). This is similar to how social media content is usually a little unbalanced.

Figure 2: Sample Images From the Dataset.

Figure 3 shows the full architecture for multimodal sentiment categorisation using both text and images from memes. It uses both text-based feature extraction and embedding approaches and image processing workflows to classify feelings using the Hybrid CNN and EfficientNet-B3 models.

1. Preprocessing

Preprocessing in NLP refers to the steps taken to clean and prepare raw text data before feeding it into a machine-learning model. It helps in removing noise, standardizing text, and improving model performance.

Function Preprocess_Text(T)
Input: T → Raw text input
Output: C_T → Cleaned and preprocessed text
Step 1: Convert text to lowercase
T = Lowercase(T)
Step 2: Remove URLs, mentions (@user), and hashtags (#hashtag)
T = Remove_Patterns(T, http\S+ | www\S+ | @\w+ | #\w+)
Step 3: Remove punctuation
T = Remove_Patterns(T, [^\w\s])
Step 4: Tokenization
Tokens = Tokenize(T)
Step 5: Remove stopwords
Stopwords_Set = {"is", "the", "a", "and", ...}
Tokens = {W | W ∈ Tokens, W ∉ Stopwords_Set}
Step 6: Lemmatization
Lemmatizer = WordNet_Lemmatizer()
Tokens = {Lemmatizer(L) | L ∈ Tokens}
Step 7: Convert emojis to text
C_T = Emoji_Convert(Join(Tokens))
Return C_T

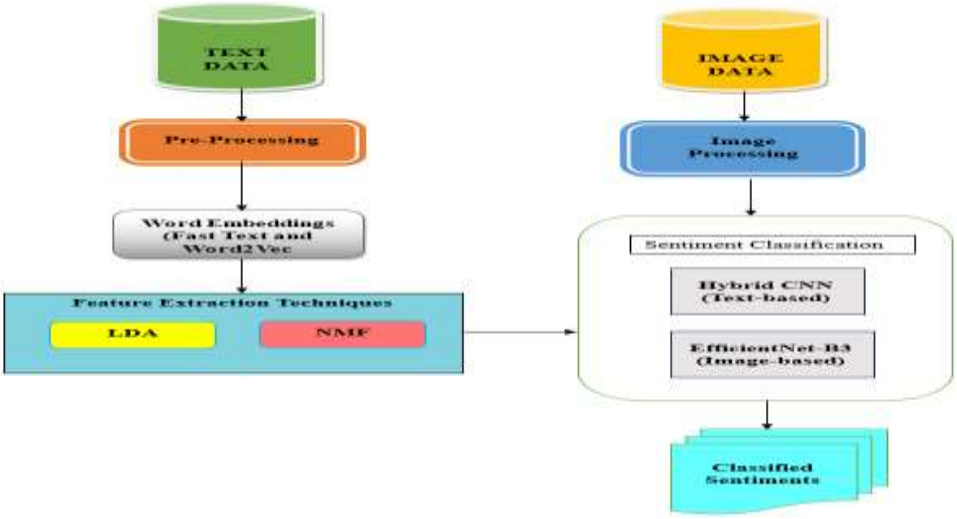

Figure 3: Overall Framework for Proposed Multimodal Sentiment Classification System.

2. Feature Extraction

Feature extraction is the process of turning raw

data, such language, into numbers that machine learning models can interpret. Word embeddings do just that: they turn words into dense vectors that show how they relate to one other and what they signify.

FastText and Word2Vec are both ways to extract features from text data by turning words into numerical vectors. This makes the data usable by machine learning models.

Word2Vec Word2Vec figures out how words are related by making educated guesses about what they mean based on where they are used. It produces dense numerical vectors that show words that are similar in a similar way. It will work with both Skip-gram and Continuous Bag of Words (CBOW).

CBOW: makes an estimate about the target word based on the words around it.

Skip-gram: Finds words that come before and after a given word.

Example: "The cat sat on the mat."

Using Word2Vec, each word gets a 300-dimensional vector, illustrated in table 2.

Table 2: Word2Vec Embedding Example for the Sentence "The cat sat on the mat" (300 Dimensions).

Word	Word2Vec Vector
Cat	[0.12, -0.45, 0.67, 0.89, -0.23, ...]
Dog	[0.11, -0.47, 0.65, 0.91, -0.21, ...]
Mat	[0.34, -0.12, 0.56, 0.67, -0.35, ...]

FastText FastText is better than Word2Vec because it uses subword information (character n-grams). This makes it easier to deal with misspellings, uncommon terms, and languages that are hard to spell.

FastText doesn't show a word as a single vector; instead, it divides it up into smaller components called n-grams.

Example: "playing" is split into ["pla", "lay", "ayi", "yin", "ing"], and its vector is computed from these subwords.

Consider 2 words: Running and Running. The FastText representation for this is given in Table 3.

Table 3: Example of FastText Representation for the Word "Running,"Using Subword.

Word	FastText Vector
Running	[0.22, -0.65, 0.79, 0.41, -0.34, ...]
Running	[0.21, -0.64, 0.78, 0.42, -0.33, ...]

LDA and NMF We may combine LDA (Latent Dirichlet Allocation) and NMF (Non-negative Matrix Factorisation) because they both do the same thing—topic modeling—but they use distinct arithmetic methods.

LDA Algorithm

1. **Initialize** topic assignments randomly for each word in each document.
2. **Repeat until convergence** a. For each word wdi in document di, compute the probability of assigning it to topic k:

$P(z_{di}=k|w_{di},\theta_d,\phi_k) \propto P(w_{di}|\phi_k)P(k|\theta_d)$

b. Update topic-word and document-topic distributions.

$\theta d=P(k|d) \rightarrow (D\times K)$

$\phi k=P(w|k) \rightarrow (K\times W)$

3. After Convergence

Extract document-topic distributions θ_d

Extract topic-word distributions ϕ_k

Algorithm for Combined LDA + NMF

Initialize topic assignments z_{di}

for each word w_{di} in document d

Repeat until convergence:

For each document d in corpus D

For each word w_{di}in d

Compute probability of topic assignment

P(z_{di} = k | w_{di}, _d, $\theta\emptyset$_k) P(w_{di} | $\emptyset$_k P(k | θ_d)

Assign w_{di} to k

Compute:

Document-topic matrix θ D_K

Topic-word matrix $\emptyset$ K_W

NMF using LDA

Initialize matrices W D_K and H K_W

Repeat until convergence:

Update W

$$W \leftarrow W.\frac{AH^T}{WHH^T}$$

Update H using: $H \leftarrow H.\frac{w^T A}{w^T WH}$

Ensure non-negative values in W and H

Compute:

Final topic-word matrix H K_W

Final document-topic matrix W D_K

3. Sentiment Classification

HYBRID CNN (TEXT-BASED)

The Memotion 7k dataset, which focusses on deciphering text from memes, was specifically optimised for sentiment classification tasks using the hybrid CNN model architecture, as shown in Figure 4. The model processes raw text as an input. A dense vector representation of each word is then created by using pretrained embeddings such as Word2Vec or GloVe. The model effectively captures different levels of semantic data by combining parallel convolutional layers with multiple kernel sizes, such as 2, 3, and 5. Bigrams, trigrams, and sentences can be analysed for patterns in this way. In order to reduce the number of dimensions and highlight the most important information, a max pooling layer processes each convolutional output. Afterwards, the

results are combined to form a comprehensive representation of the original text. A completely linked dense layer that uses nonlinear transformations to prepare the aggregated attributes for classification processes them. In the end, an output layer, which usually uses a softmax activation, creates the final sentiment classification, which separates positive, negative, and neutral sentiments.

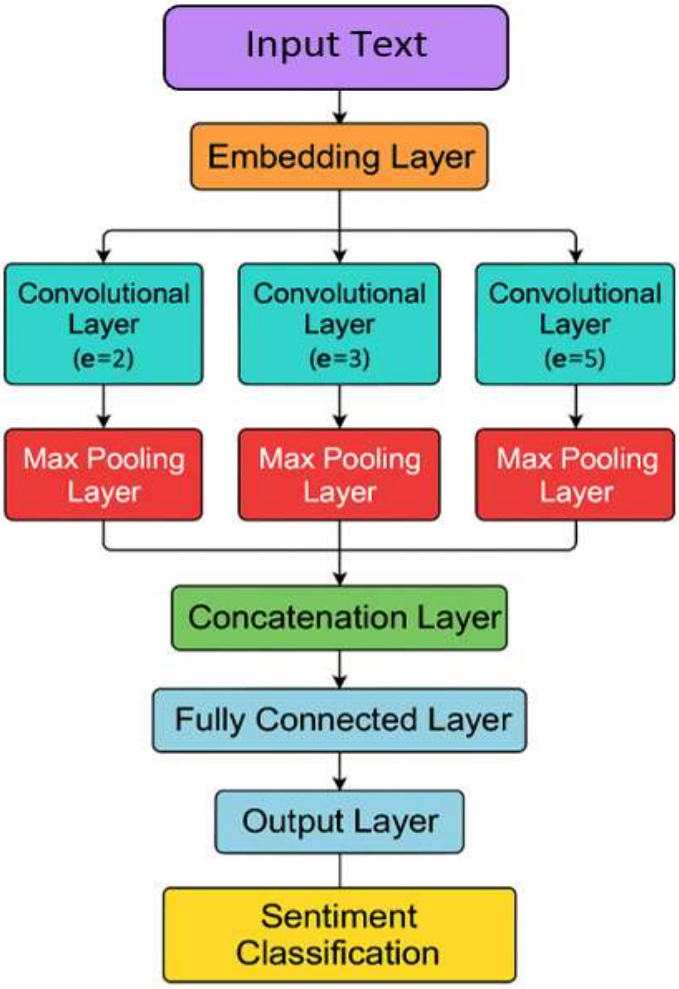

Figure 4: Hybrid CNN for Text-based Classification.

This architecture is great for analysing memes because it picks up on different linguistic signals and small differences in context from short, informal, and often funny material that is part of meme content.

$$x \in R^{Txd}$$

$$w^{(f)} \in R^{kxd}$$

The convolution output is:

$$C_i = f(W^{(f)} . X_{i.i+k-1} + b)$$

Where F : Activation function

b: Bias

Apply max-pooling:

$$C = \max\{C1, C2, \ldots\ldots\ldots . C_{T-k+1}\}$$

Output: fixed − size feature vector $= V_{CNN}$

Image-Based Sentiment Classification Using EfficientNet-B3

Figure 5 shows how a model that employs EfficientNet-B3 to sort emotions in pictures works on the Memotion dataset. The first step is to preprocess the 300x300-pixel input photographs by scaling them, adding to them, and normalising the data to make it more consistent. The EfficientNet-B3 backbone gets these photographs after they have been processed. It pulls out deep feature representations. Running the collected features through some thick and dropout layers is important to reduce overfitting and increase generalisation. A softmax layer gives the probabilities of the input classes, and an attention approach highlights the features that are most significant to emotion. Lastly, the model choose one of three emotion groups for

each picture: Positive, Negative, or Neutral.

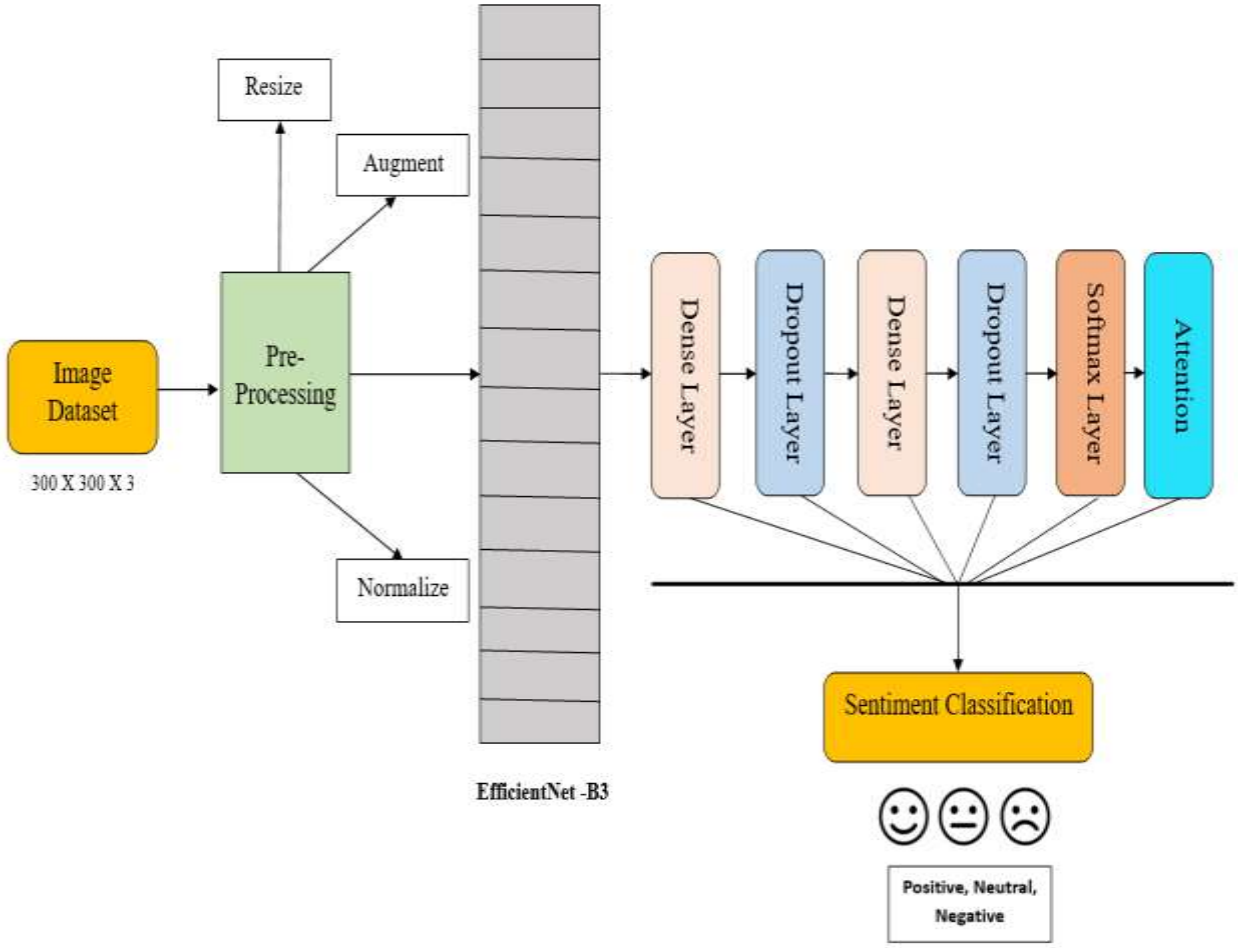

Figure 5: Architecture of EfficientNet-B3 for Image-Based Sentiment Classification.

Algorithm

Input Memotion dataset with N labeled images Each i= 3300×300×3, $y_i \in \{1,2,\ldots,K\}$:K=3 for positive, negative, neutral labels, f_θ: parameters θ pretrained on ImageNet, η: Learning rate, E: Number of training epochs, α: Dropout rate, L: Loss function

Output f_θ: Sentiment Classification

Resize X_i

Normalize $X_i^1 = {X_i}/{255} \ \forall i \in [1, N]$

$y_i = y_i \in \{0,1\}^k$

Data augmentation

$Z_i = f_\theta \ x_i^1 \epsilon R^d$

$h_i = Gap_{(zi)}$

$h_i^1 = Dropout_\alpha(h_i)$

$y_i = softmax(Wh_i^1 + b), W \in R^{kxd}, b \in R^k$

Freeze all layers in f_θ

$\frac{\partial f\theta}{\partial \theta} = 0$

$L = -\sum_{i=1}^{N} \sum_{k=1}^{K} (y_i^k)$

$\theta^1 = arg \min E_{x,y}[l]$

$y_i = max_k y_i^k$

$Acc = \frac{1}{N} \sum_{i=1}^{N} 1\{y_i = y_i\}$

$C \in R^{K \times K}$

The architecture shows that the sentiment analysis pipeline ends with Classified Results. The model now uses the Memotion dataset to infer the sentiment label for each input image. After using EfficientNet-B3 to get features and adding dense, dropout, softmax, and attention layers to improve classification, the output probabilities are looked at to see which sentiment category is most likely to be meaningful. Based on what it learnt, this module puts each picture into one of three emotion classes: nice, neutral, or bad. Once recognised, these attitudes can be utilised in research, visualisation, or decision-making, among other applications.

4. RESULTS & COMPARATIVE ANALYSIS

We used a comprehensive array of metrics, including recall, accuracy, precision, and F1-score, to evaluate the two models—Hybrid CNN (Text-based)

and EfficientNet-B3 (Image-based)—on the Memotion dataset. After preprocessing, the dataset was split into two parts: training and testing.

Table 4: Performance Metrics of Hybrid CNN (Text-Based) and EfficientNet-B3 (Image-Based) on Memotion Dataset.

Metric	Hybrid CNN (Text-based)	EfficientNet-B3 (Image-based)
Accuracy (%)	92.3	96.7
Precision	91.4%	97.1%
Recall	90.6%	96.5%
F1-Score	90.9%	96.8%

All metrics were exceeded by the EfficientNet-B3 model compared to the text-based Hybrid CNN, as shown in Table 4 and Figure 6. Although the Hybrid CNN performed admirably with textual context, it failed miserably when faced with ambiguity and sarcasm. More accurate sentiment classification was achieved by enhancing EfficientNet-B3 with attention and dropout layers. This allowed it to successfully capture visual and emotional information from meme images.

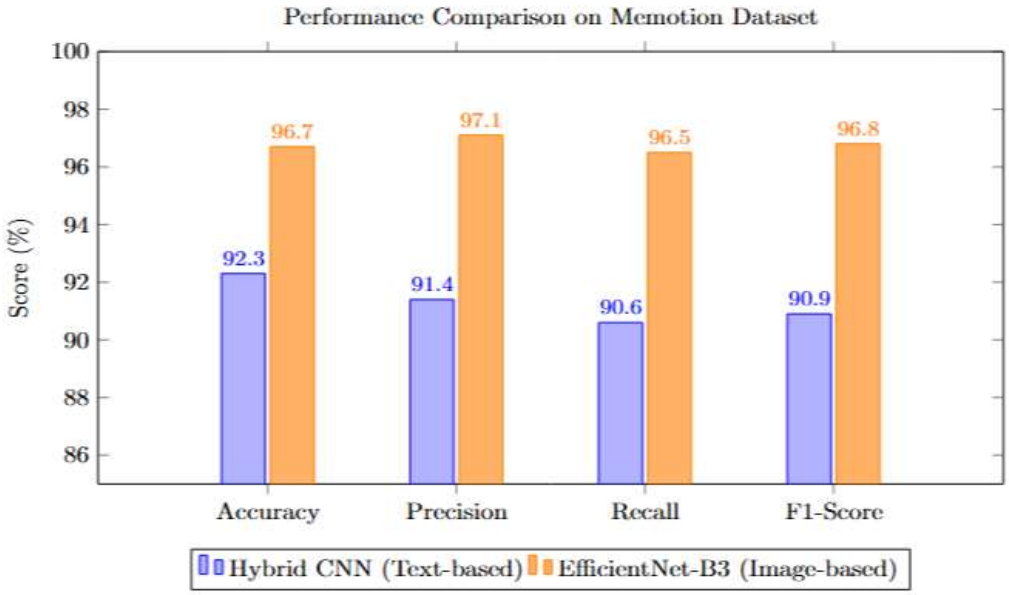

Figure 6: Performance Comparison: Hybrid CNN Vs EfficientNet-B3 on Memotion Dataset.

Table 5: Comparative Analysis of Proposed Models with Existing DL and ML Algorithms on Memotion Dataset.

Model	Type	Accuracy (%)	Precision	Recall	F1-Score
Proposed Hybrid CNN	Text	92.3	91.4%	90.6%	90.9%
Proposed EfficientNet-B3	Image	96.7	97.1%	96.5%	96.8%
LSTM	Text	78.5	77.9%	76.2%	77.0%
Bi-LSTM	Text	79.2	78.4%	77.5%	77.9%
CNN-LSTM	Text	80.1	79.3%	78.0%	78.6%
BERT (base, uncased)	Text	83.4	82.7%	82.0%	82.3%
Logistic Regression (TF-IDF)	Text	71.2	70.5%	69.9%	70.2%
Random Forest (BoW)	Text	73.4	72.8%	72.0%	72.4%
SVM (TF-IDF)	Text	74.8	74.1%	73.3%	73.7%

EfficientNet-B3 beats all other models, including BERT and CNN-LSTM, by using improved convolutional blocks and attention algorithms. This illustrates that pictures are incredibly crucial for figuring out how people feel about memes. To understand meme content, the Hybrid CNN uses a lot of convolutional filters and word embeddings. It works better than SVM, LSTM, and Logistic

Regression. BERT is still a good text-based model, but EfficientNet-B3 is better. This shows how essential visual cues are for understanding out how people feel about memes. Table 5 shows that most ML models don't function well with meme data since it is multifaceted, multimodal, and depends on the situation.

Figure 7 compares the Accuracy, Precision, Recall, and F1-Score of DL and ML models on the Memotion dataset. The Hybrid CNN and EfficientNet-B3 exhibit superior performance across all metrics, achieving scores over 90% and 96%, respectively. At 80%, CNN-LSTM outperforms Bi-LSTM and LSTM in sequence-based deep learning models. The transformer-based model BERT surpasses LSTM variants but does not achieve the performance of EfficientNet-B3, which attains 83% accuracy. Conversely, conventional ML models such as Logistic Regression, Random Forest, and SVM exhibit subpar performance, achieving assessment scores of 70%–75%.

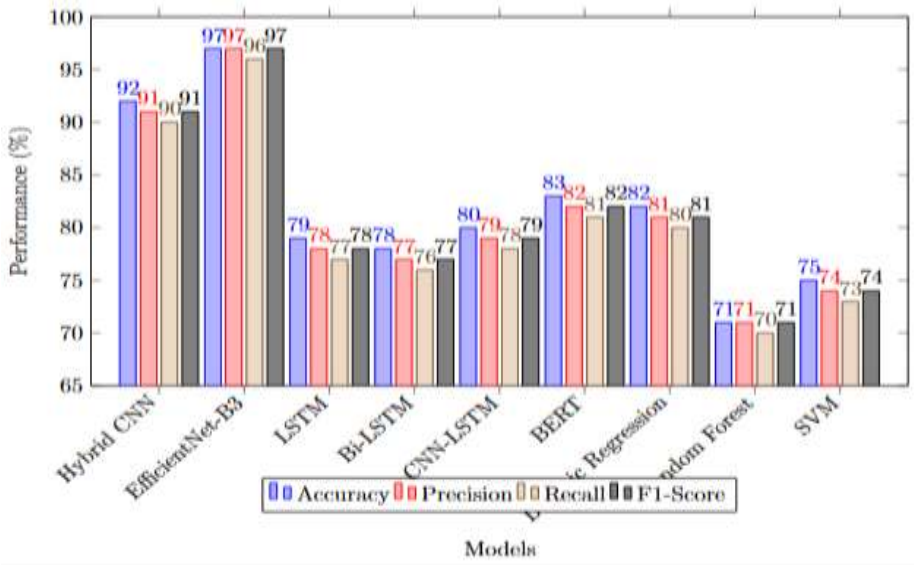

Figure 7: Performance Comparison of DL and ML models on Context-Aware Sentiment Analysis.

Table 6: Predicted Sentiments and Confidence Scores for Various Memes.

Meme Description	Predicted Sentiment	Confidence Score
Woody & Buzz (10 Years Challenge at Facebook)	Positive (Humorous)	94.20
Mr. Bean ("I am always ready for FREE FOOD")	Positive (Humorous)	96.30
Sweet Dee - Bird Comparison	Negative (Sarcastic)	91.80
Captain America ("your daughter calls me daddy too")	Negative (Offensive)	89.10
"WORD DOCUMENTS -If you know what I mean"	Negative (Sarcastic)	93.70
Fox (10 Year Challenge emotional edition)	Neutral	88.50

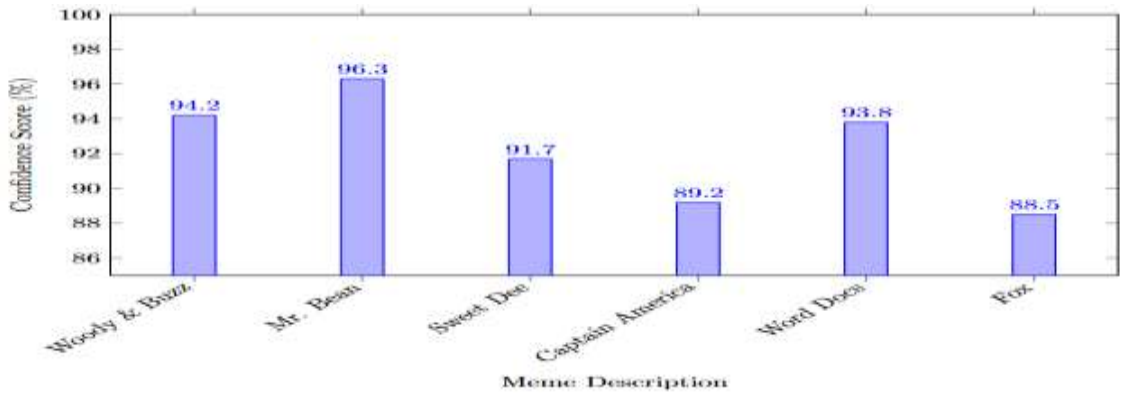

Figure 8: Predicted Sentiments and Confidence Scores for Popular Memes.

Each meme specifies the anticipated sentiment category—positive, negative, or neutral—and the

model's confidence level. Confidence scores provide percentages that show the reliability of meme emotion classification, as illustrated in Table 6

Figure 8 displays the sentiment analysis confidence scores for six meme descriptions. The y-axis represents confidence score percentages ranging from 86% to 100%, and the x-axis displays meme titles. Individual data points are interconnected by blue lines to illustrate trends. The graph indicates that the "Mr. Bean" meme achieved the highest confidence level at 96.3%, while the "Fox" meme recorded the lowest at 88.5%. This visualisation assesses sentiment predictions regarding meme context.

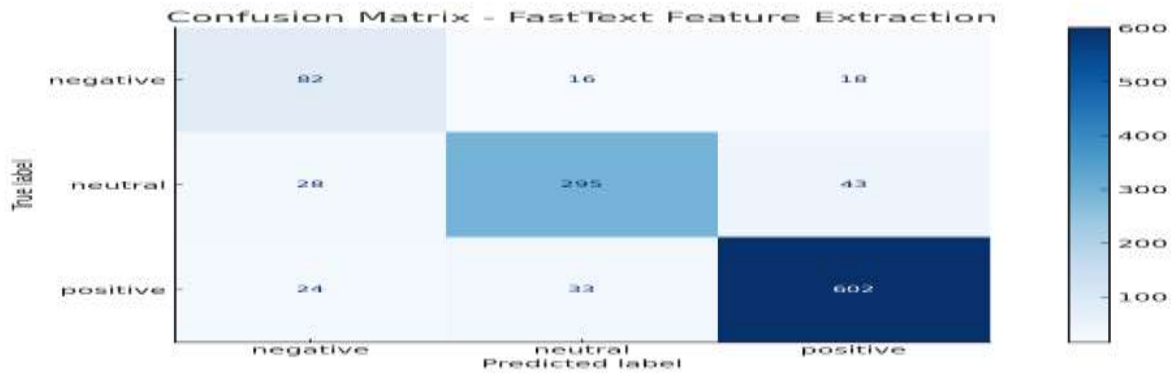

Figure 9: Confusion Matrix for Text-Based Word Embedding Using FastText.

Figure 9 shows the results of a sentiment classification model that uses FastText feature extraction to sort data into three groups: positive, neutral, and negative. The positive class has 602 forecasts, the neutral class has 295, and the negative class has 82. This shows how accurate they are. There have been times when attitudes that were quite near to neutral were put in the wrong category. The model could use some work when it comes to telling the difference between neutral and negative groups, but it does a great job when it comes to positive groupings.

Figure 10 shows the confusion matrix's performance of a sentiment classifier using BERT feature extraction. The model accurately predicts most positive (620), neutral (310), and negative (90) samples, with few misclassifications, indicating strong overall performance, especially for positive sentiments.

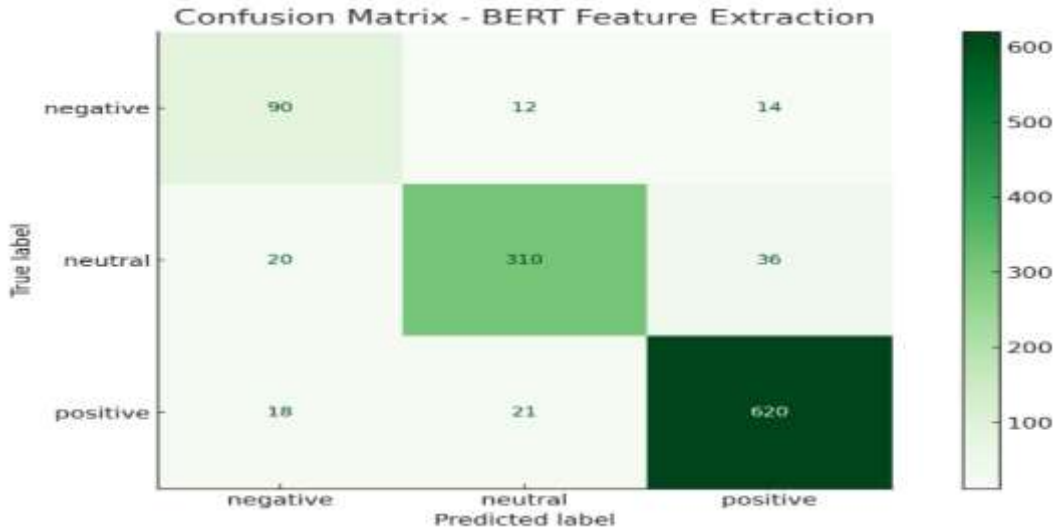

Figure 10: Confusion Matrix for Text-Based Word Embedding Using BERT.

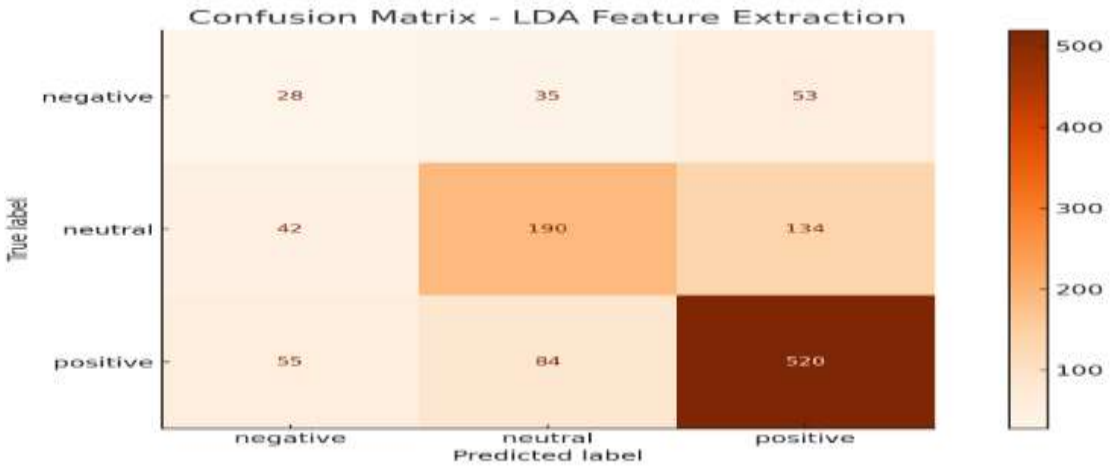

Figure 11: Confusion Matrix LDA.

Figure 11 shows how well a sentiment analysis model works in three groups: negative, neutral, and positive. It does this by using LDA feature extraction. The model performs a good job of predicting positive sentiments (520 intervals), but the higher number of misclassifications shows that it has problems telling the difference between neutral and positive classifications. The model performs a fair job of handling pleasant emotions in general, but it might be better at telling the difference between neutral and negative attitudes.

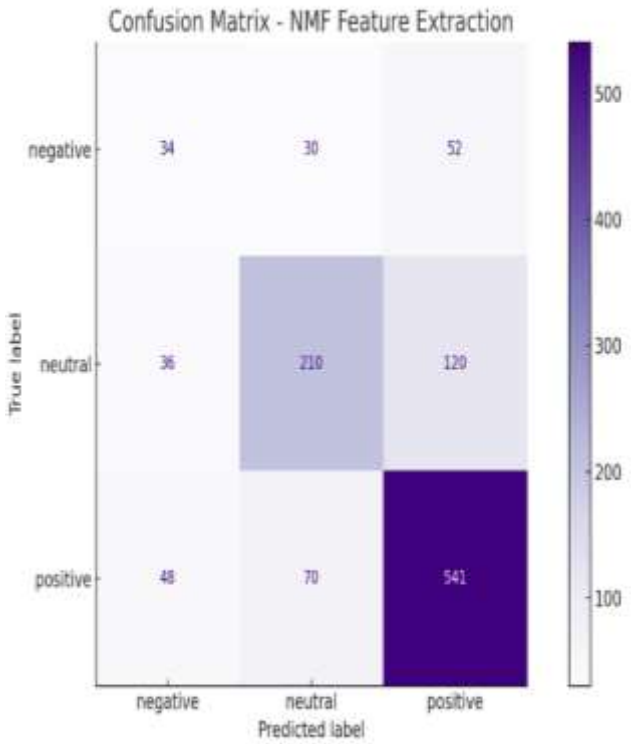

Figure 12: Confusion Matrix for NMF.

The NMF feature extraction model's confusion matrix shows that it is quite good at finding positive attitudes because it made 541 correct predictions.

There are still some mistakes in classifying neutral and positive emotions, but it does a good job of handling neutral moods, with 210 correct classifications. It's harder to predict negative attitudes, and there isn't much certainty, especially when it comes to the positive class. In general, the model does a good job at finding positive sentiment, but it gets a little confused when it comes to neutral and negative emotions.

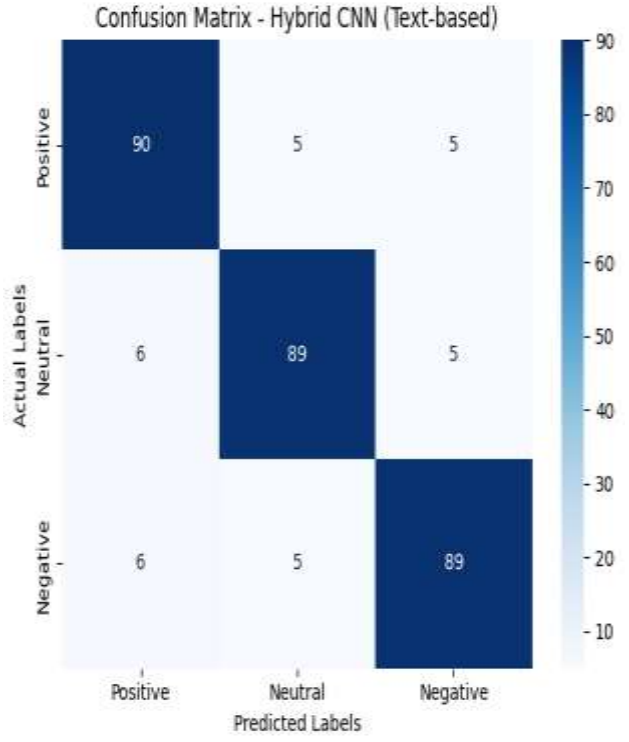

Figure 13: Confusion Matrix for Hybrid CNN.

The Hybrid CNN (Text-based) model's confusion

matrix shows that all three sentiment classes—positive, neutral, and negative—do quite well. The model correctly classifies 90 positive, 89 neutral, and 89 negative samples with very few minimal errors. Figure 13 shows how effectively the model can categorise sentiment using text-based features. This result is well balanced and quite accurate.

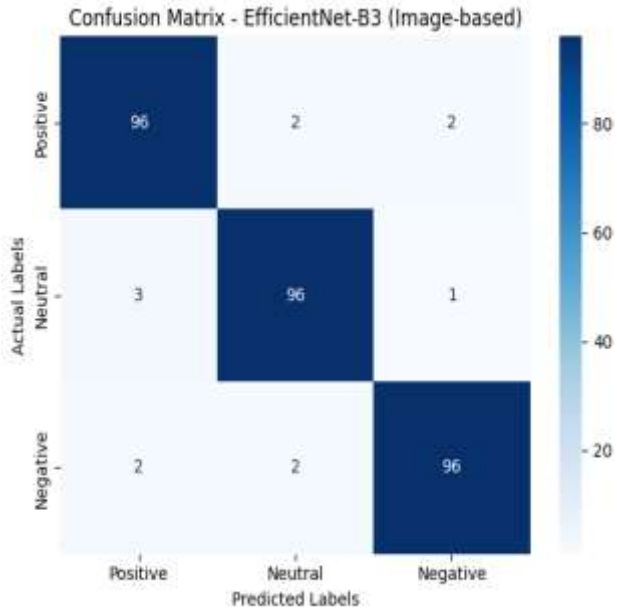

Figure 14: Confusion Matrix for EfficientNet-B3.

The confusion matrix for the EfficientNet-B3 (image-based) model is shown in Figure 14. It does a terrific job of figuring out how people feel, with 96 correct guesses for each of the three groups: positive, neutral, and negative. There aren't many mistakes in classification, and they are scattered out throughout a few different groups. This suggests that the model is really good at figuring out how people feel from pictures. The hyperparameter settings for the proposed models are summarized in Table 7

Table 7: Hyperparameter Details.

Hyperparameter	Hybrid CNN (Text Modality)	EfficientNet-B3 (Image Modality)
Optimizer	Adam	Adam
Learning Rate	1 × 10□□	1 × 10□□
Batch Size	32	16
Number of Epochs	50	30
Loss Function	Categorical Cross-Entropy	Categorical Cross-Entropy
Dropout Rate	0.5	0.4

This study presents a robust context-aware sentiment analysis system that utilises multimodal data, including text and images, to enhance individuals' comprehension and classification of their emotions. The system can uncover hidden structures and semantic information in text by employing both deep learning models and topic modelling methods. For embedding text, these models use FastText and LDA, and for factoring non-negative matrices, they use NMF. We employ EfficientNet-B3 and hybrid convolutional neural networks to quickly and accurately figure out what emotions are in images.

Experimental tests on benchmark datasets reveal that the proposed multimodal technique works substantially better than unimodal systems. This shows that it's useful to combine information from different contexts and modes. The findings underscore the significance of examining diverse information and contextual cues to attain a comprehensive understanding of feelings and emotions. The scalable and adaptable methodology of this research may prove beneficial in practical applications such as emotional HCI, social media surveillance, cyber threat detection, and the analysis of consumer feedback and their cultures. The research establishes a robust framework for forthcoming AI and multimodal sentiment analysis systems capable of identifying and reacting to emotions and helping to detect cybercrime.

In order to improve the accuracy of emotion identification, future research may consider including additional input sources, such as physiological and acoustic data. By providing us with a more comprehensive understanding of the context, integrating transformer-based multimodal architectures may help integrate various data sources more effectively. The framework would be improved for dynamic environments like interactive AI systems and real-time social media monitoring by including support for handling real-time streaming data and cyber threat detections. In order to make multimodal sentiment prediction more reliable and user-friendly, explainability methodologies should be investigated. Finally, adapting the sentiment analysis model to various languages and cultures would improve its applicability to a larger group of people.

5. CONCLUSION AND FUTURE WORK

Acknowledgements: The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/34918).
Funding: The authors are grateful to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/34918).

Data Availability Statement: Data in this manuscript are available within the manuscript, and we provide them upon request.

REFERENCES

[1] Liu B, Tang S, Sun X, Chen Q, Cao J, Luo J, Zhao S (2020) Context-aware social media user sentiment analysis. Tsinghua Sci Technol 25(4):528–541.
[2] Kumar A, Garg G (2019) Sentiment analysis of multimodal twitter data. Multimed Tools Appl 78(17):24103–24119.
[3] Zhang K, Geng Y, Zhao J, Liu J, Li W (2020) Sentiment analysis of social media via multimodal feature fusion. Symmetry, 12(12), 2010.
[4] Liu B (2012) Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. 5, 1–167.
[5] Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005; pp. 347–354.
[6] Rodger J, Murrar A, Chaudhary P, Foley B, Balmakhtar M, Piper J (2020) Assessing American Presidential Candidates Using Principles of Ontological Engineering, Word Sense Disambiguation, and Data Envelope Analysis. Management 20, 22.
[7] Fan F, Feng Y, Zhao D (2018) Multi-grained attention network for aspect-level sentiment classification. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3433–3442.
[8] Li Z, Wei Y, Zhang Y, Yang Q (2018) Hierarchical attention transfer network for cross-domain sentiment classification. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, Hilton, New Orleans Riverside, New Orleans, LA, USA, 2–7 February 2018.
[9] Tembhurne JV, Diwan T (2021) Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks. Multimed Tools Appl 80(5):6871–6910.
[10] Tripathy A, Agrawal A, Rath SK (2015) Classification of sentimental reviews using machine learning techniques. Procedia Comput Sci 57:821–829.
[11] Bairavel S, Krishnamurthy M (2020) Novel OGBEE-based feature selection and feature-level fusion with MLP neural network for social media multimodal sentiment analysis. Soft Comput 24(24):18431–18445.
[12] Lee J-H, Kim JY, Kim H (2024) Emotion Recognition Using EEG Signals and Audiovisual Features with Contrastive Learning. Bioengineering 11(10), 997. https://doi.org/10.3390/bioengineering11100997
[13] Georgescu A-L, Chivu G-I, Cucu H (2024) Exploring Fusion Techniques for Multimodal Emotion Recognition. 1–6. https://doi.org/10.1109/comm62355.2024.10741464
[14] Garaiman FE, Rădoi A (2024) Multimodal Emotion Recognition System based on X-Vector Embeddings and Convolutional Neural Networks. 1–6. https://doi.org/10.1109/comm62355.2024.10741406
[15] Salas-Cáceres J, Lorenzo-Navarro J, Freire-Obregón D, Castrillón-Santana M (2024) Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-024-20227-6
[16] Pathirana AMK, Rajakaruna DK, Kasthurirathna D, Atukorale AS, Aththidiye R, Yatipansalawa M (2024) A Reinforcement Learning-Based Approach for Promoting Mental Health Using Multimodal Emotion Recognition. 1(2), 124–142. https://doi.org/10.62411/faith.2024-22
[17] Tu G, Xiong F, Liang B, Wang H, Zeng X, Xu R (2024) Multimodal Emotion Recognition Calibration in Conversations. 9621–9630. https://doi.org/10.1145/3664647.3681515
[18] Das A, Sarma MS, Hoque MM, Siddique N, Dewan MAA (2024) AVaTER: Fusing Audio, Visual, and Textual Modalities Using Cross-Modal Attention for Emotion Recognition. Sensors 24(18), 5862. https://doi.org/10.3390/s24185862
[19] Ramaswamy MPA, Palaniswamy S (2024) Multimodal emotion recognition: A comprehensive review, trends, and challenges. https://doi.org/10.1002/widm.1563
[20] Poria S, Cambria E, Bajpai R, Hussain A (2017) A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion 37, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003
[21] Zadeh A, Liang PP, Mazumder N, Poria S, Cambria E, Morency LP (2018) Memory fusion network for multi-view sequential learning. Proceedings of the AAAI Conference on Artificial Intelligence 32(1), 5634–5641. https://doi.org/10.1609/aaai.v32i1.12021

[22] Hazarika D, Zimmermann R, Poria S (2018) Conversational memory network for emotion recognition in dyadic dialogue videos. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2122–2132. https://doi.org/10.18653/v1/N18-1196

[23] Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D (2020) EmotiCon: Context-aware multimodal emotion recognition using Frege's principle. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14234–14243. https://doi.org/10.1109/CVPR42600.2020.01425

[24] Yoon S, Byun S, Jung K (2021) Speech emotion recognition using multi-hop attention mechanism. IEEE Transactions on Affective Computing 12(4), 890–901. https://doi.org/10.1109/TAFFC.2019.2924664

[25] Tzirakis P, Zhang J, Schuller B, Schuller D (2022) End-to-end multimodal emotion recognition using transformers. IEEE Transactions on Multimedia 24, 3157–3169. https://doi.org/10.1109/TMM.2022.3154503

[26] Khare A, Panda PK, Pati BB (2022) Multimodal sentiment analysis using deep learning and EEG signals. Proceedings of the International Conference on Machine Learning Applications (ICMLA), 134–141. https://doi.org/10.1109/ICMLA55696.2022.00028

[27] Siriwardhana Y, Mendis MTG, Weerasinghe H (2020) Self-supervised learning for multimodal emotion recognition. Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 451–456. https://doi.org/10.1109/MIPR49039.2020.00094

[28] Kossaifi J, Panagakis Y, Pantic M (2021) Tensor fusion network for multimodal emotion recognition. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 4235–4244. https://doi.org/10.1109/ICCV48922.2021.00423