DESIGN OF FRAMEWORK FOR SENTIMENT ANALYSIS USING DEEP LEARNING

A Thesis Submitted
in the Partial Fulfilment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

by

ANANYA PANDEY

(2K21/PHDIT/08)

Under the Supervision of
Prof. DINESH KUMAR VISHWAKARMA
Delhi Technological University

Department of Information Technology

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

August, 2024

ACKNOWLEDGEMENTS

I am profoundly grateful to my esteemed PhD supervisor, **Prof. Dinesh Kumar Vishwakarma**, whose exceptional guidance and unwavering support have been instrumental in completing this thesis. His exemplary discipline, unyielding focus, and relentless work ethic have inspired me and set a high standard for academic excellence. I am fortunate to have benefited from his expertise and commendable commitment throughout this challenging yet rewarding journey. I thank **Mrs. Sushma Vishwakarma**, **Diya** and **Advika** for ensuring I always had a home away from home.

No man is complete without his family, who silently toil behind the scenes to help him fight for his dreams. I thank my parents, **Mr. Rajeev Pandey** and **Mrs. Shobha Pandey,** for being the best parents one could hope for. I thank my sibling **Anushka Pandey, Rudransh Agnihotri** and **Uditanshu Pandey**, for doing what siblings do best, i.e., be loving in their own fun way.

Finishing PhD is a highly challenging journey, and the seniors who helped navigate this path need a special mention. To this end, I am grateful for the support of my PhD seniors, **Dr. Ankit Yadav, Mrs. Anusha Chabbra.**

As I went through the most challenging phase of my PhD, my juniors ensured that I never gave up and always came back stronger. I am grateful to **Ashish Bajaj, Abhishek Verma** and **Bhavana Verma** for all the light-hearted conversations. I would like to express my heartfelt gratitude to my dear friend and roommate, **Sangeeta**, for being an unwavering pillar of support throughout my PhD journey.

I extend my heartfelt appreciation to the state-of-the-art research lab established by my supervisor. Equipped with cutting-edge NVIDIA GPUs, it was pivotal in facilitating the success of the computationally expensive deep learning-based research experiments throughout my PhD.

Last but not least, I thank God for giving me the persistence and strength to show up at my lab each day and work through the ups and downs of this PhD journey.

Ananya Pandey
2K21/PHDIT/08

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I Ananya Pandey hereby certify that the work which is being presented in the thesis entitled "Design of Framework for Sentiment Analysis using Deep Learning" in partial fulfilment of the requirements for the award of the Degree of Doctor of Philosophy, submitted in the Department of Information Technology, Delhi Technological University is an authentic record of my own work carried out during the period from 02/08/2021 to 08/08/2024 under the supervision of Prof. Dinesh Kumar Vishwakarma.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Ananya

Candidate's Signature

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that Ananya Pandey (2K21/PHDIT/08) has carried out their research work presented in this thesis entitled “Design of Framework for Sentiment Analysis using Deep Learning” for the award of Doctor of Philosophy from the Department of Information Technology, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Signature:

Name of the Supervisor: Prof. Dinesh Kumar Vishwakarma

Designation: Professor**;** Head of Department (Information Technology)

Address: Rohini, Delhi

Date: 10/08/2024

ABSTRACT

Sentiment analysis is a computational technique that analyses the subjective information conveyed within a given expression. This encompasses appraisals, opinions, attitudes or emotions towards a particular subject, individual, or entity. Conventional sentiment analysis solely considers the text modality and derives sentiment by identifying the semantic relationship between words within a sentence. Despite this, certain expressions, such as exaggeration, sarcasm and humour, pose a challenge for automated detection when conveyed only through text. Multimodal sentiment analysis incorporates various forms of data, such as visual and acoustic cues, in addition to text. By utilising fusion analysis, this approach can more precisely determine the implied sentiment polarity, which includes positive, neutral, and negative sentiments. Thus, the recent advancements in deep learning have boosted the domain of multimodal sentiment analysis to new heights. The research community has also shown significant interest in this topic due to its potential for both practical application and educational research. In light of this fact, this research aims to present a thorough analysis of recent ground-breaking research studies conducted in the field of sentiment analysis using diverse modalities. Furthermore, this thesis dives into a discussion of the multiple categories of multimodal data, diverse domains in which multimodal sentiment analysis can be applied, challenges associated with multimodal sentiment analysis, and suggests different frameworks for analysing sentiments using visual-caption pairs and videos. The ultimate goal of this investigation is to indicate the success of deep learning architectures in tackling the complexities associated with multimodal data analysis.

People are becoming accustomed to posting images, captions and audios on social media platforms to express their opinions. For our subsequent strategy, we conducted a comprehensive assessment and examination of the performance of several multimodal sentiment analysis models across a range of modalities. However, most recent multimodal strategies concatenate features from the visual, caption & audio modalities with the help of pre-trained deep learning models containing millions of trainable parameters without adding a dedicated attention module, ultimately leading to less desirable results. Motivated by this observation, we have proposed a novel model VABDC-Net, VyAnG-Net that integrates the attention module with the conventional state-of-the-art models to extract the most relevant contextual information from these diverse modalities. The experimental results show that our suggested approaches can generate ground-breaking outcomes when applied to publicly available multimodal datasets, specifically Twitter-2015, Twitter-2017, MUStARD, and

MUStARD++. The experimental results demonstrate that the proposed model attains much superior accuracy scores on all of these datasets and exhibits much higher efficiency compared to conventional approaches in predicting sentiment in multimodal data.

Also, this investigation aims to employ Target-Dependent Multimodal Sentiment Analysis to identify the level of sentiment associated with every target (aspect) stated within a multimodal post consisting of a visual-caption pair. Despite the recent advancements in multimodal sentiment recognition, there has been a lack of explicit incorporation of emotional clues from the visual modality. The challenge at hand is to proficiently obtain visual and emotional clues and subsequently synchronize them with the textual content. In light of this fact, this thesis also presents a novel approach called the Visual-to-Emotional-Caption Translation Network (VECT-Net) technique to effectively acquire visual sentiment clues by analyzing facial expressions. Additionally, it effectively aligns and blends the obtained emotional clues with the target attribute of the caption mode.

Additionally, a novel contrastive learning-based multimodal architecture have been introduced to predict emoticons using the Multimodal-Twitter Emoticon dataset acquired from Twitter. This proposed model employs the joint training of dual-branch encoder along with the contrastive learning to accurately map text and images into a common latent space. Our key finding is that by integrating the principle of contrastive learning with that of the other two branches yields superior results. The experimental results demonstrate that our suggested methodology surpasses existing multimodal approaches in terms of accuracy and robustness.

In conclusion, this thesis presents substantial discoveries and identifies potential areas for future research on the subject of sentiment analysis utilizing multi-modal data.

LIST OF PUBLICATIONS

Publications Arising from Research Work in the Thesis

SCIE Journal Papers

- **A. Pandey** and D. K. Vishwakarma, “VABDC-Net: A framework for Visual-Caption Sentiment Recognition via spatio-depth visual attention and bi-directional caption processing,” ***Knowledge-Based Systems***, vol. 269, June. 2023, doi: https://doi.org/10.1016/j.knosys.2023.110515.
- **A. Pandey** and D. K. Vishwakarma, “Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey,” ***Applied Soft Computing***, vol. 152, November. 2024, doi: https://doi.org/10.1016/j.asoc.2023.111206.
- **A. Pandey** and D. Kumar Vishwakarma, “Target-Dependent Multimodal Sentiment Recognition Via a Visual-to-Emotional-Caption Translation Network using Visual-Caption Pairs.” Under Minor Revision in ***Signal, Image and Video Processing*** (Pub: Springer). https://doi.org/10.48550/arXiv.2408.10248.
- **A. Pandey** and D. K. Vishwakarma, “Contrastive Learning-based Multi-Modal Architecture for Emoticon Prediction by Employing Image-Text Pairs.” Under Review in ***Cognitive Computation*** (Pub: Springer). https://doi.org/10.48550/arXiv.2408.02571.
- **A. Pandey** and D. K. Vishwakarma, “VyAnG-Net: A Novel Multi-Modal Sarcasm Recognition Model by Uncovering Visual, Acoustic and Glossary Features.” Under Minor Revision in ***Intelligent Data Analysis*** (Pub: IOS Press). https://doi.org/10.48550/arXiv.2408.10246

Conference Papers

- **A. Pandey** and D. K. Vishwakarma, “Multimodal Sarcasm Detection (MSD) in Videos using Deep Learning Models,” in ***2023 IEEE International Conference in Advances in Power, Signal, and Information Technology (APSIT)***, Institute of Electrical and Electronics Engineers (IEEE), Jun. 2023, pp. 811-814. doi: 10.1109/APSIT58554.2023.10201731.
- **A. Pandey** and D. K. Vishwakarma, “Attention-based Model for Multi-modal sentiment recognition using Text-Image Pairs,” in *2023* ***4th International Conference on Innovative Trends in Information Technology (ICITIIT)***, Institute of Electrical and Electronics Engineers (IEEE), March. 2023, pp. 1–5. doi: 10.1109/ICITIIT57246.2023.10068626.

Publications Arising from Research Work Outside the Thesis

SCIE Journal Papers

- S. Aggarwal**, A. Pandey**, D. K. Vishwakarma, "Modelling Visual Semantics via Image Captioning to extract Enhanced Multi-Level Cross-Modal Semantic Incongruity Representation with Attention for Multimodal Sarcasm Detection." Under Review in ***Neural Computing and Applications***, (Pub: Springer). https://doi.org/10.48550/arXiv.2408.02595.

Conference Papers

- Pankaj Gupta, **A. Pandey**, and D. K. Vishwakarma, "Attention-free based dual-encoder mechanism for Aspect-based Multimodal Sentiment Recognition," in *2023* ***IEEE International Conference in Advances in Power, Signal, and Information Technology (APSIT)***, Institute of Electrical and Electronics Engineers (IEEE), Jun. 2023, pp. 534-539. doi: 10.1109/APSIT58554.2023.10201711.
- S. Aggarwal, **A. Pandey**, and D. K. Vishwakarma, "Multimodal sarcasm recognition by fusing textual, visual and acoustic content via multi-headed attention for video dataset," in ***2023 World Conference on Communication & Computing (WCONF)***, Institute of Electrical and Electronics Engineers (IEEE), July. 2023, pp. 1-5. doi: 10.1109/WCONF58270.2023.10235179.
- S. Kapoor, S. Gulati, S. Verma, **A. Pandey**, and D. K. Vishwakarma, "Multimodal Architecture for Sentiment Recognition via employing Multiple Modalities," in ***2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)***, Institute of Electrical and Electronics Engineers (IEEE), May. 2024, pp. 435-439. doi: 10.1109/InCACCT61598.2024.10551131.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Sentiment analysis, also known as opinion mining, is the computational study of people's opinions, sentiments, emotions, and attitudes. Using deep learning techniques, sentiment analysis has made significant advancements, allowing for more accurate and nuanced understanding of text data. Deep learning models, particularly those based on neural networks like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, have proven highly effective in capturing the complexities of human language. These models excel at understanding context, handling ambiguity, and recognizing patterns in large datasets. By leveraging vast amounts of labelled training data, deep learning models can learn to distinguish subtle differences in sentiment, enabling applications such as customer feedback analysis, social media monitoring, and market research. The power of deep learning in sentiment analysis lies in its ability to automatically extract features and representations from raw text, reducing the need for manual feature engineering and improving the overall accuracy and robustness of sentiment predictions.

1.1 Growing Popularity of Social Media Platforms

The last decade has witnessed a tremendous rise in social media platforms. An extensive online presence has become a normal part of daily human lives. The number of active users on social media has grown tremendously, from just over 2.5 million active users at the beginning of 2017 to almost 5 billion active users by the end of 2024 [1] as illustrated from **Figure 1.1**. One of the most significant developments of this century is the data revolution. Furthermore, it is worth noting that over the past two decades, the volume of data has experienced exponential growth. With the widespread availability of affordable mobile devices and high-speed Internet access, the global community has effectively become a closely connected and accessible network. data on the Internet has led to the emergence of an entirely new field known as multimodal data analysis. The importance of social media is discussed as follows:

- Social media platforms facilitate the connection and interaction between individuals.
- Social media serves as a medium for disseminating information, exchanging ideas, and expressing opinions.
- Social media also appeals to a significant percentage of individuals who passively consume information. Users generate and distribute multimedia content, as well as

access and investigate material shared by other members of the online community, organizations, groups, etc.

- Social media exerts a significant influence on the mental state of individuals.

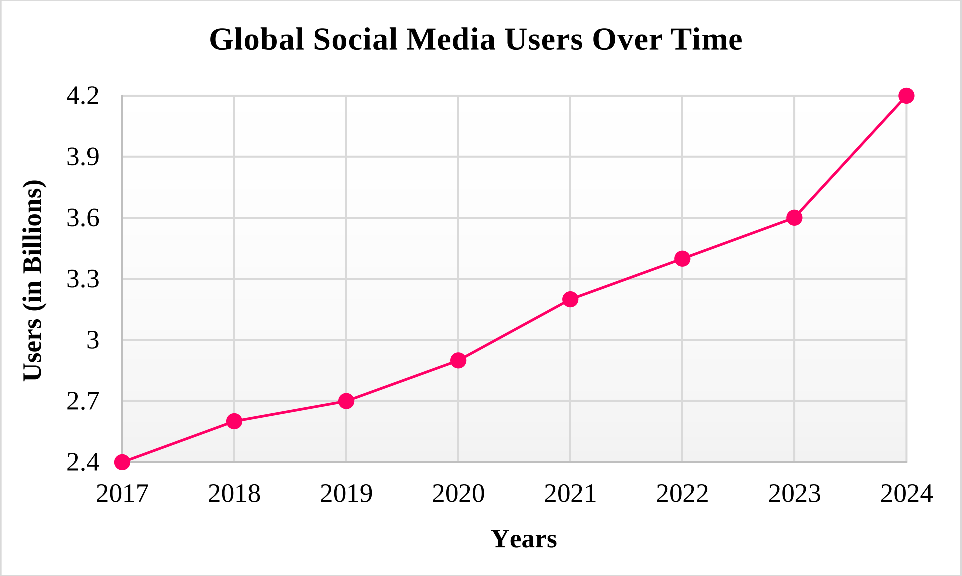


Figure 1.1 The worldwide record of active users on social media platforms, quantified in billions.

With the proliferation of social media platforms, e-commerce websites, video blogs, etc., individuals have the ability to engage in the buying and selling of various goods and services. Typically, individuals provide their feedback on a service, product, or any particular topic, in the form of textual reviews. **Figure 1.2** illustrates the engagement rates on major social media platforms (Facebook, Instagram, Twitter, and TikTok) from 2017 to 2024.

Key observations:

- Facebook: Indicates a consistent yet marginal upward trend in engagement rates over the years.
- Instagram: Illustrates a steady and substantial increase, indicating a progressive user engagement.
- Twitter: Exhibits a progressive rise in engagement rates, though with a less pronounced slope compared to Instagram.
- TikTok: Experiences a significant surge in user engagement beginning in 2019, indicating a rapid growth in its popularity.

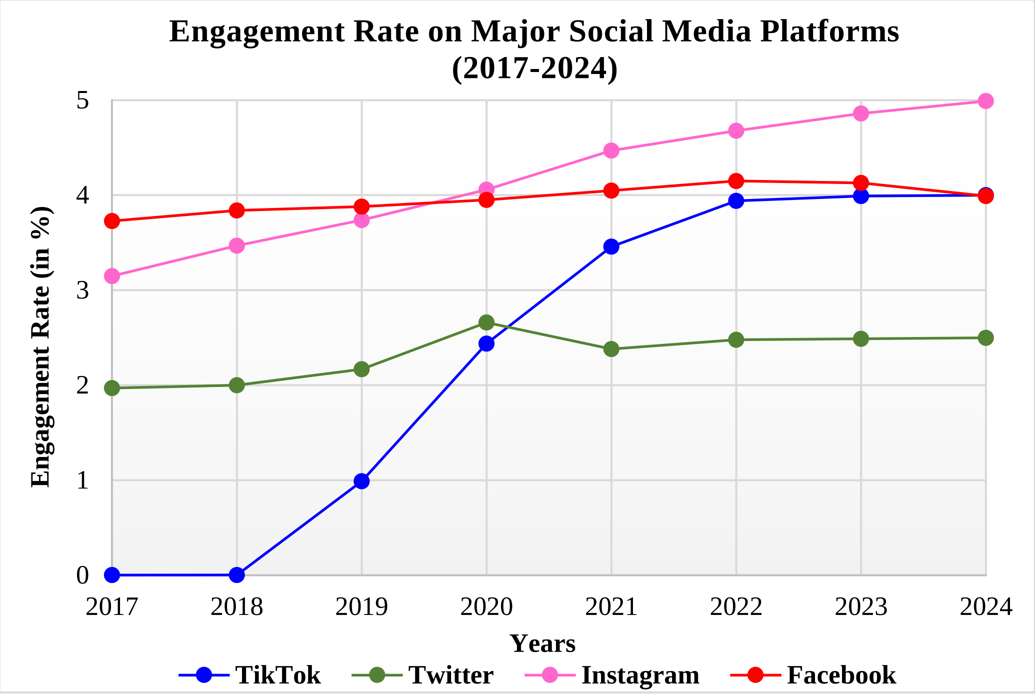


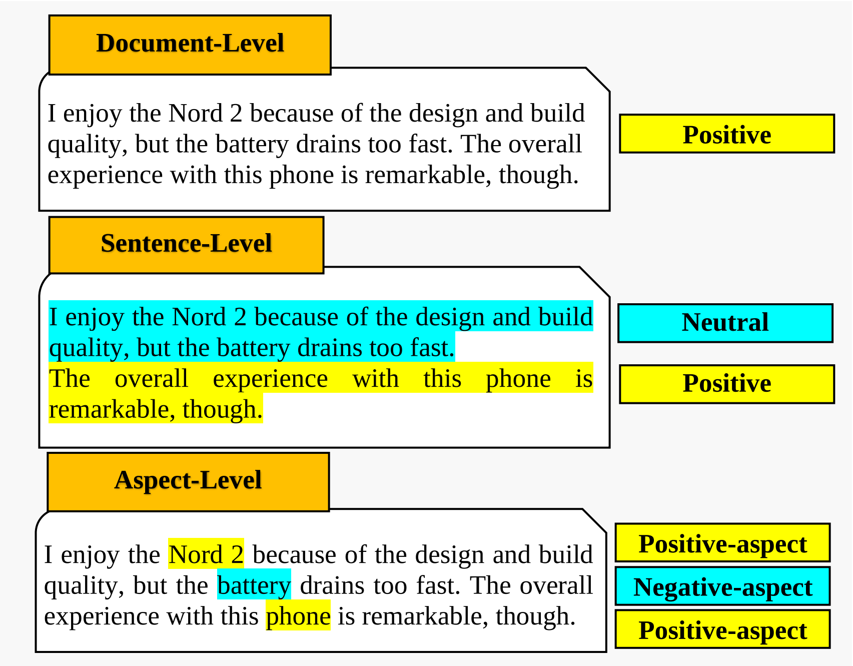
Figure 1.2 Graph illustrating the engagement rates of users on prominent social media platforms in percentage.

The availability of high-speed internet enabled the user to submit their reviews not only in textual form but also in acoustic and visual formats. The proliferation of multimedia data on the internet has led to the emergence of an entirely new field known as multimodal data analysis. As the engagement rate increases on different social media platforms, there is a corresponding increase in the use of multimodal data for reviewing or providing feedback on products.

1.2 Sentiment Recognition

Affective computing is an integrative field that traverses psychology, cognitive sciences, and computer science. For many years, sentiment analysis has become a vigorous domain of exploration in affective computing [2], [3], [4]. The achievement of any company or item straightforwardly relies upon its customer because if the customer likes the product, it is a success; if not, you certainly need to improve the product by making some changes. All in all, here the question emerges how might you know whether your item is successful or not? Well, for that, you want to investigate or analyse your clients. One of the critical qualities of analysing your clients is examining their feelings. Here, the concept of sentiment analysis comes to play. Thus, sentiment analysis can be defined as perceiving and identifying opinions from numerous information sources. **Figure 1.3** shows multiple degree of sentiment analysis in text modality.

Customers and manufacturers seek to understand the "customer's opinion" regarding respective products or services, which is the primary motivation behind sentiment analysis.



I enjoy the Nord 2 because of the design and build quality, but the battery drains too fast. The overall experience with this phone is remarkable, though.

I enjoy the Nord 2 because of the design and build quality, but the battery drains too fast.
The overall experience with this phone is remarkable, though.

I enjoy the Nord 2 because of the design and build quality, but the battery drains too fast. The overall experience with this phone is remarkable, though.

Figure 1.3 Different levels of sentiment analysis that rely on textual-based information.

Consequently, sentiment analysis has enjoyed significant business and scholarly attention. However, some phrases, such as exaggeration, sarcasm, and humour, present a problem for automatic recognition when they are solely communicated via text. In such instances, the incorporation of multiple modalities, such as visual and acoustic signals, can more precisely determine the implied sentiment polarity. The focus of this discussion pertains to the primary goal of conducting multimodal sentiment analysis.

1.3 Multimodal Sentiment Recognition

In today's era, individuals are regularly bombarded with huge amounts of multimodal data across diverse social media platforms. Such data is often presented in the form of comments that incorporate images, text, audio, videos, and emoticons. In light of these facts, the categorization of various modalities might be grouped, as shown in **Figure 1.4**. The task of identifying implicit emotions from these comments is quite complex. Although unimodal sentiment analysis has the capability to identify an individual's attitudes towards a particular subject, it can be challenging to distinguish between explicit sentiment and implicit emotions such as sarcasm, humour, or exaggeration. Hence, in order to achieve a more accurate prediction of the emotion, the utilisation of multimodal data processing would be a more advantageous approach. Hence, the ultimate goal of this study is to indicate the success of deep learning architectures in tackling the complexities associated with multimodal sentiment recognition by discussing the diverse domains in which multimodal sentiment recognition can be applied, challenges associated with it, and suggests different frameworks for analysing sentiments using visual-caption pairs and videos.

People are currently working on multimodal sentiment recognition, which is one of the most exciting and challenging active research areas. Pandey et al. [5], Gupta et al. [6], Pandey et al. [7], Aggarwal et al. [8] are a few recent publications that have established their relevance in this area. The model utilised in these research studies has integrated various modalities as input to estimate sentiment. A crisp idea of it is demonstrated in **Figure 1.10**.

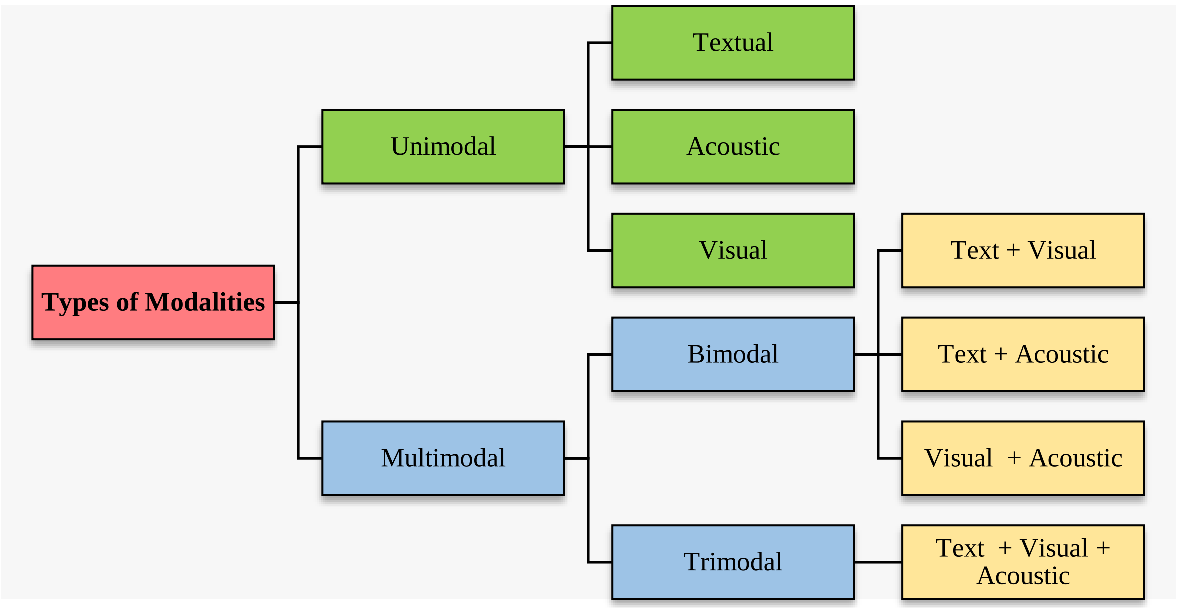


Figure 1.4 A taxonomy of all kinds of modalities.

1.4 Fusion Strategies to Blend Multiple Modalities

For integrating multiple data sources, there are eight fusion techniques that can be utilized, as depicted in **Figure 1.5**. Let us examine each one with greater depth.

1.4.1 Early Fusion (EF)

It is typically performed in three steps as outlined below:

- Firstly, each modality's features are extracted using a different methodology, such as exploiting TF-IDF for text, using of OpenSmile toolkit for acoustic, and utilizing a neural network for visual features.
- A vector is then created from the multiple modalities as a result of combining all the extracted features.

- The combined feature vector is subsequently fed into one of the classification models to make the prediction.

Morency et al. [9] was the first attempt to implement tri-modal sentiment analysis by using a Hidden Markov model after the concatenation of the feature vectors. Perez et al. [10] developed a method that merges all multimodal features into a single feature matrix, and then the final obtained matrix is used by an SVM for classification. Similarly, Park et al. [11] makes use of radial basis function kernels along with SVM for prediction and classification. Zadeh et al. [12] aggregate the different modes of data into a unique feature matrix and transfer it to the Bi-LSTM, while the authors of [13] after converting, transfer the final hidden state to the fully connected layers for sentiment analysis. However, both methods achieve almost similar results.

Disadvantage- The implementation of this method is much more complex and tedious because it involves the unification of those features which are diverse in nature.

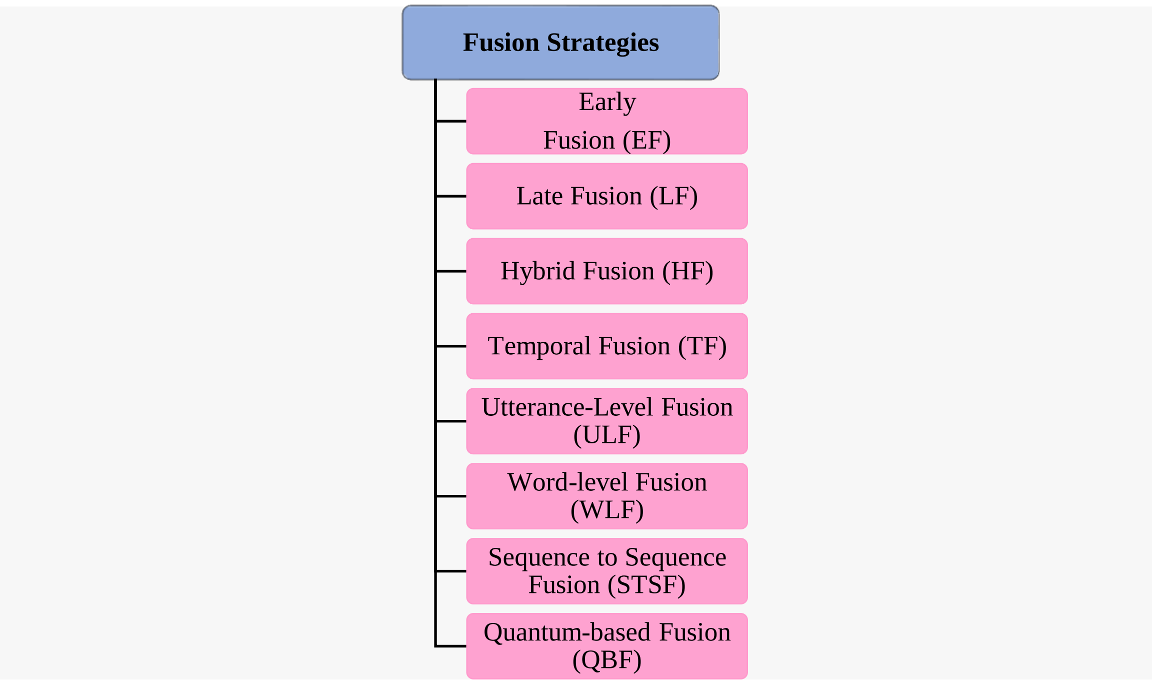


Figure 1.5 Different Fusion strategies for MSR to integrate diverse modalities such as text, images, videos, and emoticons.

1.4.2 Late Fusion (LF)

This technique is again carried out in 3 phases, as drafted below:

- The very first step is similar to the early fusion, namely, extracting features from different modalities.
- Once the feature vectors from multiple modalities have been generated, three of the separate vectors are fed into the associated classification model.

- An ultimate output result is obtained by integrating the three results from the previous step.

In [14] a neural network is trained for each of the separate multimodal data content, and then the decision is made based on performing majority voting on the result of each modality network for the final MSR. This study also utilizes the averaging method for the final sentiment scores. The method in [13] involves first training three different LSTMs for three multimedia pieces of information, then aggregating the last hidden layer of the three LSTMs. The aggregated hidden layers are transferred to two fully connected layers in the final step for obtaining the actual prediction score. The framework in [15] implemented both early and late fusion techniques, but better results were achieved using Late-RMNN. Both early and late fusion strategy are illustrated in **Figure 1.6**.

Advantages-

- It abolishes the need to combine diverse features.
- For the final prediction, each modality can use a separate classification model independently.

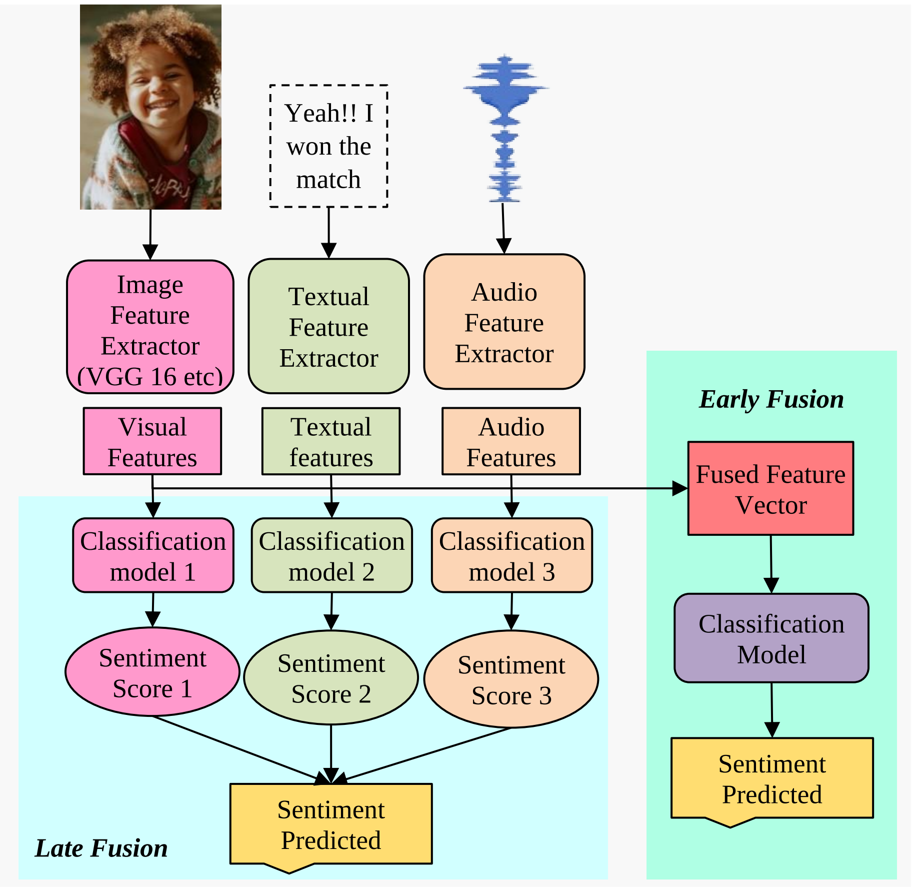


Figure 1.6 Early and Late fusion strategy.

1.4.3 Hybrid Fusion (HF)

It is a process of fusing the features obtained from late and early fusion, as its name suggests. It is viewed as a two-step process, as illustrated below:

- The fusion of two modalities is carried out at the **EF** first.
- After that, **LF** is regulated on the previous results and the remaining modality.

Kumar et al. [16] uses the VGG model to learn the feature representations of image-related data and the Glove word embedding approach for learning the text-based features. Finally, they use the hybrid fusion to predict emotions using BalanceNet (a neural network) classification model. Kumar et al. [17] also, make use of the hybrid fusion technique for fine-grained analysis.

1.4.4 Temporal Fusion (TF)

This fusion technique is particularly beneficial when dealing with view-specific and cross-view dynamics. The LSTM and attention blocks are the two critical components of TF's basic design. In the LSTM block, at each time step t features from each modality are fed to separate LSTM blocks to model view-specific dynamics. The outputs from the several LSTMs are then passed to the attention block, which focuses on the relevant features while ignoring the less significant ones to model cross-view dynamics. All the research studies incorporating this fusion technique differ in how they employ the attention mechanism. **Figure 1.7** depicts the overall structure of TF where y_l^{ts}, y_a^{ts} and y_v^{ts} are the features for textual, acoustic, and visual data at time step ts. The architectures used in [18], [19], [20] make use of TF for merging multiple modalities. The [10] is a recent publication that employs TF for sentiment prediction and emotion recognition. This study proposes a new LSTM-based model named temporal convolution multimodal LSTM, including a gating mechanism for integrating the representation of all modalities and the temporal correlations between them for final prediction.

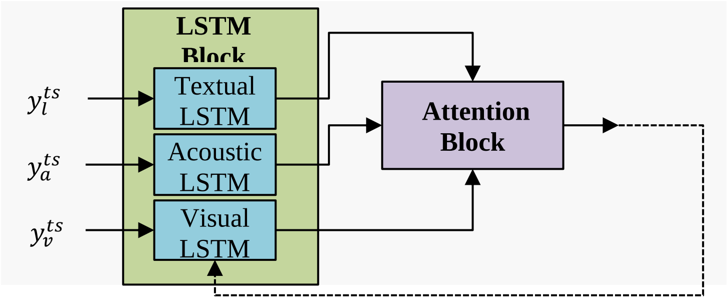


Figure 1.7 General structure of temporal-based fusion approach.

1.4.5 Utterance-level Fusion (ULF)

Unlike TF, this fusion method works on the entire utterance and employs view-specific and cross-view dynamics instead of working on each time step. This fusion model has two modules, namely embedding and fusion network. The output of the embedding module is the embeddings of all the modalities for an utterance. View-specific dynamics are employed in the embedding module. The embeddings of all modalities are then sent to the fusion module, which concatenates all of the embeddings into a single vector representation and thereby models cross-view dynamics. **Figure 1.8** shows the general structure of ULF, where **T, A,** and **V** define textual, acoustic, and visual modalities. [21], [22], [23], [24], [25] are a few cutting-edge studies that employ ULF.

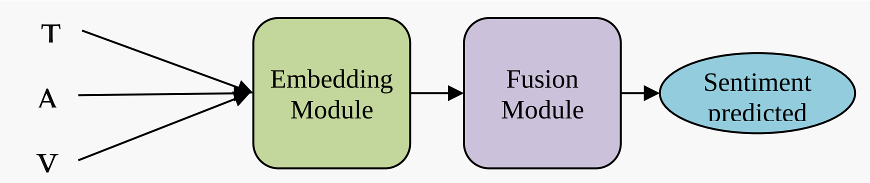


Figure 1.8 General structure of utterance-level fusion strategy.

1.4.6 Word-level Fusion (WLF)

In WLF, every word in a sequence is merged with non-verbal modalities, i.e., acoustic and visual, to learn the variation vectors. In Zhang et al. [26] firstly, the correlation between the image and text is examined with the help of the attention network. A model named IDLSTM, a sentimental inner-class dependency enhancement model, is proposed in this study. The IDLSTM learns the inner dependent relationships between the query and words in the caption and achieves the final sentimental prediction using an image–text pair as the query.

1.4.7 Sequence-to-Sequence Fusion (STSF)

For the first time, Google released STS models for machine translation. Before it, the translation had been working in a pretty naïve fashion. Inspired by the notion behind these models, a multimodal cyclic translation network (MCTN) is proposed by [27] to learn the combined representation of multiple modalities by translation between them and utilizing only source modality as an input. The key benefit of this network is that it does not require all modalities to learn the composite representation for sentiment prediction during testing time. Because for final sentiment prediction, only the data from the source modality at test time is required after the translation model has been trained with paired multimodal data. This ensures that the model is unaffected by changes or missing data in the other modalities. The MCTN network is depicted in **Figure 1.9** for better understanding and visualization. Similarly, Tsai et al. [28] uses the STSF strategy for sentiment prediction.

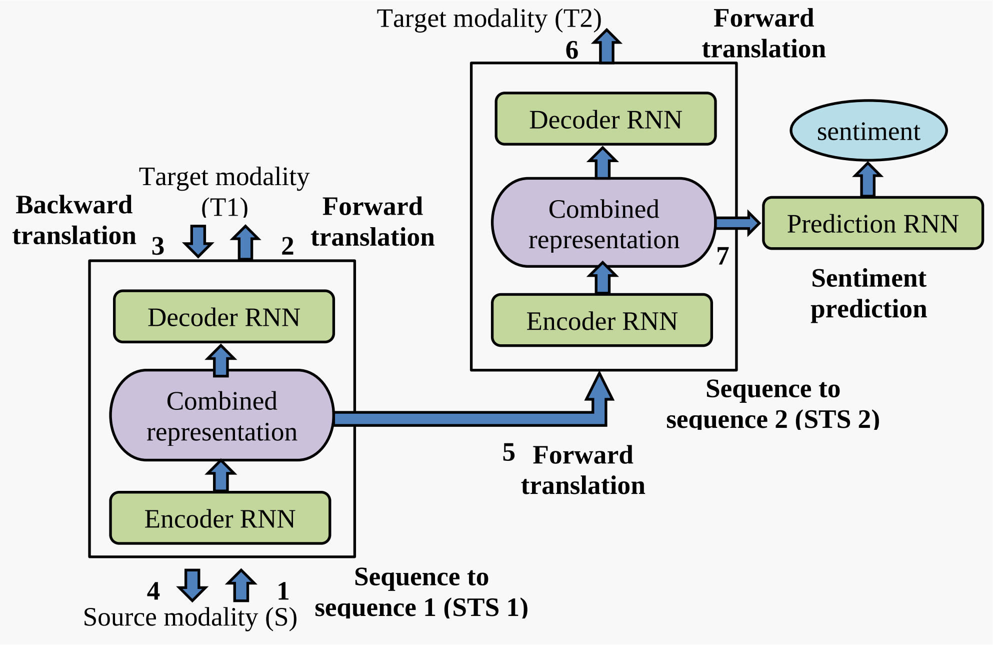


Figure 1.9 MCTN architecture which involves sequence-to-sequence fusion strategy to merge multiple modalities [27].

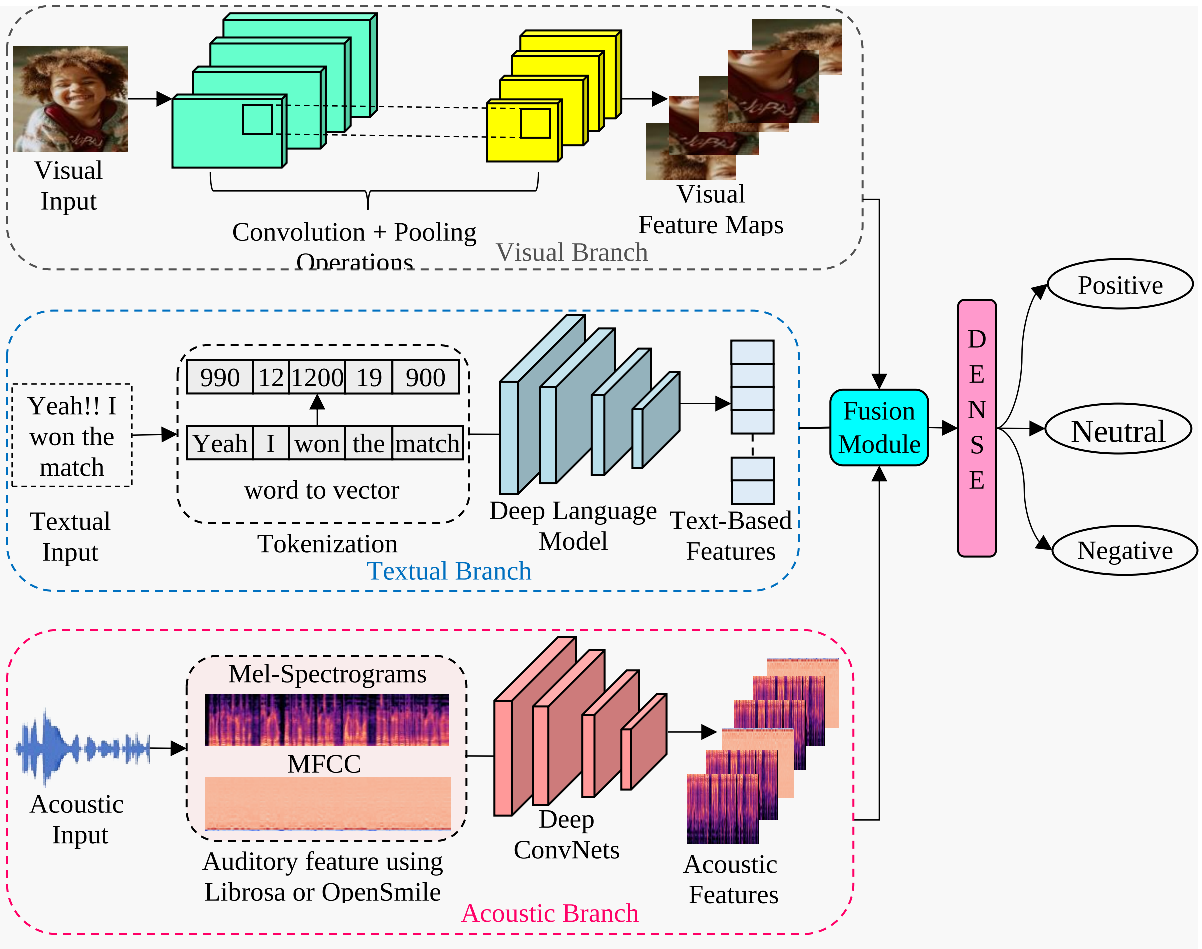


Figure 1.10 Multimodal sentiment analysis (MSA) involves the analysis of textual, visual, and auditory information in order to better understand the emotional content or sentiment being expressed by the speaker.

1.5 Motivation

Currently, numerous social media channels, such as Facebook, are available for individuals to articulate their emotions. As of January 2023, a survey from the Statista research department revealed that Facebook had the highest number of active users globally, totaling 2.89 billion. Multiple research studies analyse user sentiments in order to gain a deeper understanding of a certain service or product. Sentiment analysis has various applications. Over 50% of the present research has not yet been applied practically. Therefore, let's explore the various domains where this field might be utilized effectively. **Figure 1.11** depicts a comprehensive range of possible applications in this subject. Given the widespread usage of sentiment prediction in diverse sectors including several modalities, we were encouraged to explore the field of multimodal sentiment analysis.

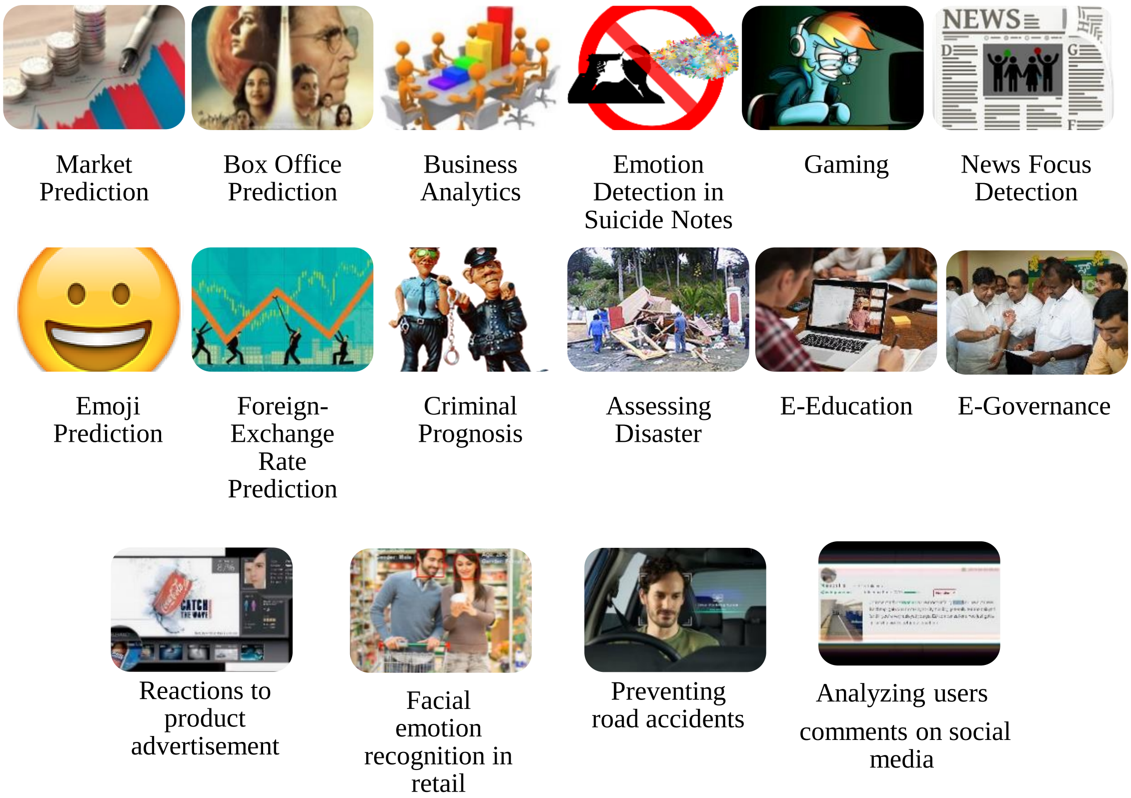

Market Prediction
Box Office Prediction
Business Analytics
Emotion Detection in Suicide Notes
Gaming
News Focus Detection
Emoji Prediction
Foreign-Exchange Rate Prediction
Criminal Prognosis
Assessing Disaster
E-Education
E-Governance
Reactions to product advertisement
Facial emotion recognition in retail
Preventing road accidents
Analyzing users comments on social media

Figure 1.11 A diverse range of domains in which the application of sentiment analysis can prove to be efficacious.

1.6 Significance of Study

Multimodal sentiment analysis, which integrates information from multiple modalities such as text, audio, and visual data, has become increasingly significant for several reasons:

- *Enhanced Accuracy:* By combining different types of data, multimodal sentiment analysis can provide a more comprehensive understanding of sentiment. For example, while text can convey the explicit content of a message, audio can reveal tone and pitch, and visual data can capture facial expressions and gestures. This leads to a more accurate and nuanced sentiment assessment.

- *Contextual Understanding:* Multimodal sentiment analysis allows for a better grasp of the context in which sentiments are expressed. For instance, a sarcastic remark might be detected through a combination of the textual content and the speaker's tone of voice or facial expression, something that might be missed if only one modality is considered.

- *Robustness to Ambiguity:* Sentiment expressed in text alone can be ambiguous or lack clarity. Combining text with other modalities can help disambiguate such sentiments. For example, a statement that appears neutral in text might be understood as positive or negative when considering the speaker's facial expressions or tone.

- *Applications in Real-world Scenarios:* Multimodal sentiment analysis is particularly valuable in applications such as social media monitoring, customer service, and human-computer interaction. It enables more effective sentiment detection in video content, live streams, and face-to-face communications, where relying on a single modality would be insufficient.

- *Improvement in Human-Computer Interaction:* In areas like virtual assistants and chatbots, multimodal sentiment analysis allows for more empathetic and context-aware interactions. For instance, a virtual assistant can better understand and respond to user emotions by analysing voice tone and facial expressions in addition to the spoken words.

- *Rich Data Utilization:* With the proliferation of multimedia content on platforms like YouTube, Instagram, and TikTok, there is a wealth of data available across different modalities. Multimodal sentiment analysis leverages this rich data to derive deeper insights that are not possible through unimodal analysis.

In summary, the significance of multimodal sentiment analysis lies in its ability to provide a more holistic, accurate, and context-aware understanding of sentiments, making it an essential tool in various fields such as social media analytics, customer service, and human-computer interaction.

1.7 Sources of Research Works Studied

In this thesis, we analysed the leading journals and conferences of recent years sourced from the following databases:

- Elsevier
- Association for Computing Machinery Digital Library
- IEEE Xplore
- Springer Link
- Association for Computational Linguistics Anthology

In addition to the above journals, IEEE, Springer, Elsevier, ACM, and ACL Anthology, we screened over 150 research papers on multimodal sentiment analysis, primarily focused on deep learning techniques. Furthermore, around 119 papers have been selected from the top journals and conferences based on the high impact factor. The research articles from these databases were searched using multiple keywords and synonyms such as, multimodal sentiment analysis, sentiment analysis across numerous modalities, multimodal opinion mining, sentiment analysis in videos, sentiment analysis in image and text pairs, and sentiment analysis in videos, audio, and text. A breakdown of the distribution of the articles across all top journals and conferences is displayed in **Table 1.1**.

Table 1.1 Distribution of peer-reviewed articles for multimodal sentiment analysis.

S. No	Name of Journal/Conference	Publisher Name	Conference/Journal/ Workshop	Count
1	IEEE Transactions of affective computing	IEEE	Journal	4
2	Intelligent system	IEEE	Journal	3
3	IEEE Transactions on Audio, speech, and language processing	IEEE	Journal	3
4	IEEE Transactions on Multimedia	IEEE	Journal	3
5	IEEE Transactions on Games	IEEE	Journal	1
6	Signal Processing Letters	IEEE	Journal	2
7	IEEE Transactions on Industrial Informatics	IEEE	Journal	1
8	IEEE Transactions on Computational Social Systems	IEEE	Journal	1

S. No	Name of Journal/Conference	Publisher Name	Conference/Journal/ Workshop	Count
9	CVPR (IEEE Conference on Computer Vision and Pattern Recognition)	IEEE	Conference	1
10	ICIP (IEEE International Conference on Image Processing)	IEEE	Conference	3
11	IJCNN (International Joint Conference on Neural Networks)	IEEE	Conference	4
12	Big data	IEEE	Conference	2
13	ICBDA (IEEE International Conference on Big Data Analysis)	IEEE	Conference	1
14	ICME (IEEE International Conference on Multimedia and Expo)	IEEE	Conference	3
15	ICCVW (IEEE/CVF International Conference on Computer Vision Workshop)	IEEE	Workshop	2
16	ICDM (IEEE International Conference on Data Mining)	IEEE	Conference	4
17	ICCCNT (International Conference on Computing, Communication and Networking Technologies)	IEEE	Conference	1
18	ICASSP (IEEE International Conference on Acoustics, Speech and Signal Processing)	IEEE	Conference	3
19	BCD (IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering)	IEEE	Conference	1
20	IEEE Global Engineering Education Conference (EDUCON)	IEEE	Conference	1
21	Neural Processing letters	Springer	Journal	1
22	Applied Intelligence	Springer	Journal	1
23	Journal of Big Data	Springer	Journal	1
24	Cognitive Computation	Springer	Journal	1
25	World wide web (WWW)	Springer	Journal	2
26	Advances in Intelligent Systems and Computing	Springer	Journal	1
26	Multimedia tools and application	Springer	Journal	5
27	International Conference on Artificial Intelligence and Speech Technology	Springer	Conference	1
28	CDSMLA	Springer	Conference	1
29	Applied soft computing	Elsevier	Journal	1
30	Neurocomputing	Elsevier	Journal	2
31	Information processing and management	Elsevier	Journal	6
32	Knowledge-based system	Elsevier	Journal	8
33	Information and management	Elsevier	Journal	1

S. No	Name of Journal/Conference	Publisher Name	Conference/Journal/ Workshop	Count
34	Expert systems with application	Elsevier	Journal	3
35	Cognitive systems research	Elsevier	Journal	1
36	Image and vision computing	Elsevier	Journal	1
37	Decision support system	Elsevier	Journal	1
38	Computers and geoscience	Elsevier	Journal	1
39	Disaster risk reduction	Elsevier	Journal	2
40	Information Fusion	Elsevier	Journal	3
41	Theoretical Computer Science	Elsevier	Journal	1
42	Transactions on Interactive Intelligent Systems	ACM	Journal	1
43	Advances in Neural Information Processing Systems	ACM	Journal	2
44	Conference on multimedia	ACM	Conference	5
45	ICIMCS (International Conference on Internet Multimedia Computing and Service)	ACM	Conference	1
46	IJCAI (International Joint Conference on Artificial Intelligence)	ACM	Conference	1
47	ACM International Conference on Multimodal Interaction	ACM	Conference	2
48	International Conference on Language Resources and Evaluation	ACL	Conference	1
50	Annual meeting of the Association for Computational Linguistics	ACL	Conference	7
51	Challenge-HML (Second Grand Challenge and Workshop on Multimodal Language)	ACL	Workshop	2
52	Conference on Empirical Methods in Natural Language Processing	ACL	Conference	1
53	Arxiv	-	-	5
54	IEEE Access	IEEE	Journal	1
55	International Journal of Information Systems and Management	INDERS CIENCE	Journal	1
55	Total			119

This study examines multimodal sentiment analysis related research articles spanning 2018 to 2024. **Figure 1.12** shows the total number of papers corresponding to all these years.

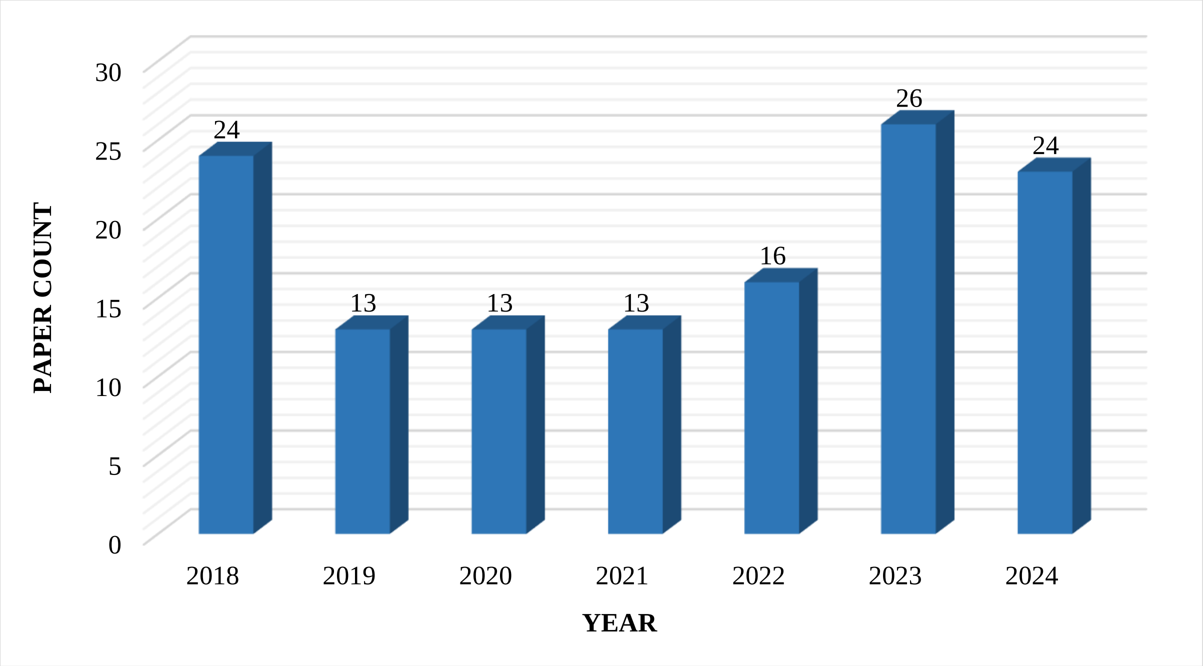


Figure 1.12 Yearly distribution of cited papers.

1.8 Overview of Chapters

The remaining section of the document is structured in the following manner.

- **Chapter 2:** Literature Review of the existing state-of-the-art methods for detecting sentiment analysis using unimodal and multimodal data.
- **Chapter 3:** Elaborate on the proposed model for “sentiment analysis based on image-text pairs”.
- **Chapter 4:** Elaborate on the proposed model for Target-dependent or Entity based multimodal sentiment analysis.
- **Chapter 5:** Elaborate on the proposed model for “emoticon prediction” based on image-text pairs.
- **Chapter 6:** Present the proposed model for predicting “sentiment from videos” by utilizing its visual, audio and textual (subtitles) information.
- **Chapter 7:** Present the conclusion of the research work done and possible future direction.

The **Figure 1.13** provides a detailed representation of how the different chapters of my PhD thesis are interconnected and aligned with the central research theme. It demonstrates the logical flow and relationship of each chapter to the overarching PhD title, highlighting how each section contributes to the broader research objective. This visual aids in understanding the cohesive structure of the thesis and the role of each chapter in addressing the research problem.

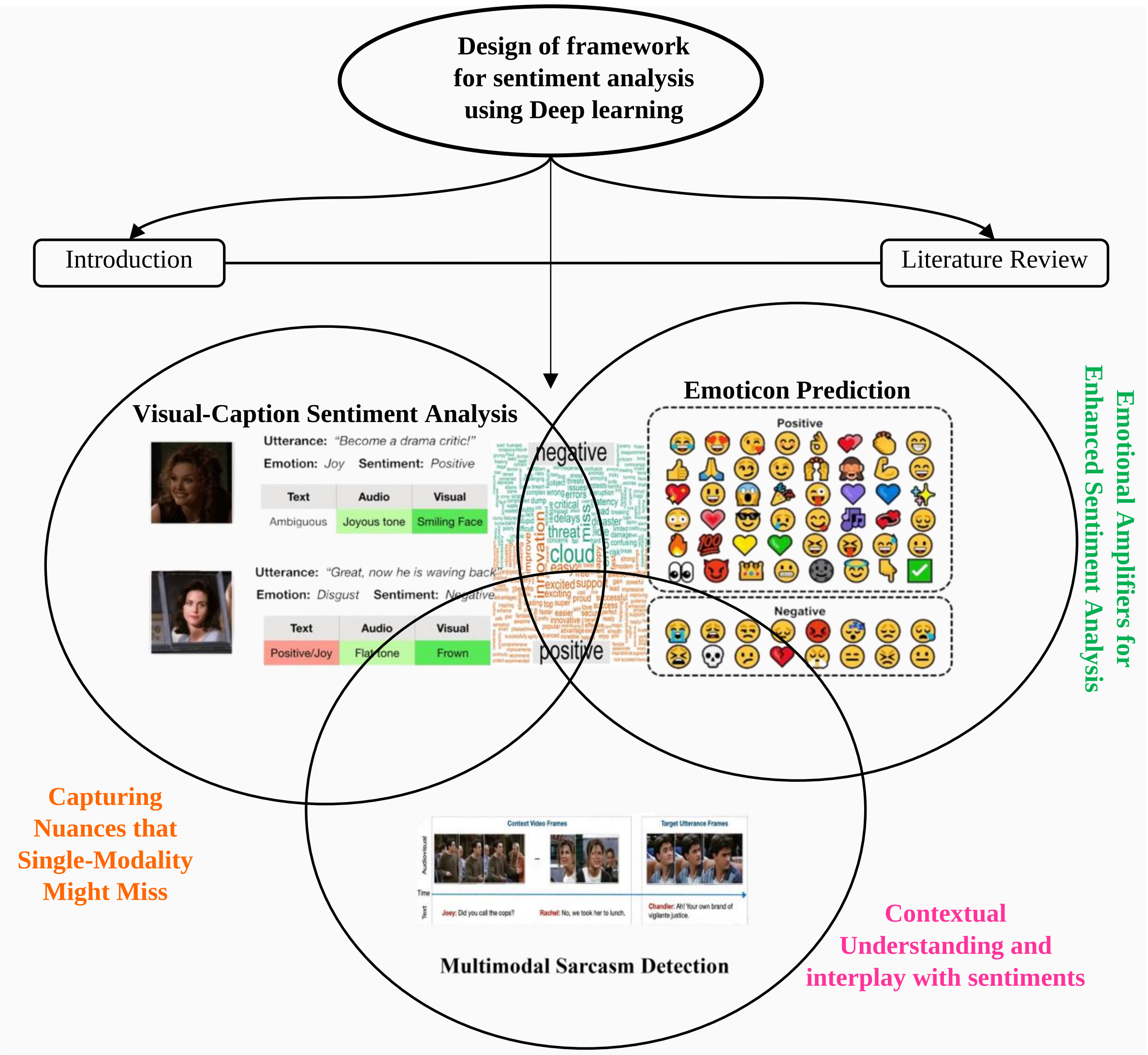

Design of framework for sentiment analysis using Deep learning
Introduction
Literature Review
Visual-Caption Sentiment Analysis
Utterance: "Become a drama critic!"
Emotion: Joy Sentiment: Positive
Text
Audio
Visual
Ambiguous
Joyous tone
Smiling Face
Utterance: "Great, now he is waving back"
Emotion: Disgust Sentiment: Negative
Text
Audio
Visual
Positive/Joy
Flat tone
Frown
negative
positive
Emoticon Prediction
Positive
Negative
Emotional Amplifiers for Enhanced Sentiment Analysis
Capturing Nuances that Single-Modality Might Miss
Contextual Understanding and interplay with sentiments
Context Video Frames
Target Utterance Frames
Audiovisual
Time
Text
Joey: Did you call the cops?
Rachel: No, we took her to lunch.
Chandler: Ah! Your own brand of vigilante justice.
Multimodal Sarcasm Detection

Figure 1.13 Relationships between the various chapters of my PhD thesis and their alignment with the central research title.

Chapter 2: Literature Review

Although there has been notable advancement in predicting sentiment from text and vision separately, there is a dearth of research on simultaneously predicting sentiment from multimodal input. We have surveyed a number of papers on sentiment analysis using multiple modalities and have chosen to contribute to this area.

This section delves into the fundamental context of multimodal sentiment analysis, with a particular focus on combining visual and text modalities. This section also contains a compilation of pre-validated unimodal and multimodal techniques for sentiment identification.

2.1 Unimodal Sentiment Recognition

This section highlights prior studies on sentiment recognition that solely employed one modality.

2.1.1 Text-Based Sentiment Recognition (TBSR)

The method of predicting the sentiment polarity for any script-based data is known as Text-Based Sentiment Recognition (TBSR). Based on script-based data, sentiment recognition may be divided into three categories: document, sentence, and aspect-based. The document-based method [29], [30] determines if the overall sentiment of an evaluation report appears to be positive, negative, or neutral. On the contrary, [31] sentence-based analysis determines the sentiment polarity at the phrase level. As the name implies, aspect-based study [32], [33], [34] focuses on the kind of aspect that relates to the attributes or features in a text-based evaluation. This type of sentiment recognition system is called "fine-grained," which denotes a comprehensive evaluation of text-based data. It tries to identify the category type before classifying sentiments for that particular category. Three approaches are often utilized for the TBSR task: lexicon, machine-learning, and deep-learning-based methods. Traditionally, lexicon-based techniques use sentimental terms and rules such as sentiment inversion to determine sentiment for script-based data. Vu et al. [35] used a lexicon-based approach named SentiWordNet to identify the opinion of the text. Pang et al. [36] used machine-learning algorithms such as Naïve Bayes for sentiment prediction. Barbosa et al. [37] introduced a two-step sentiment categorization strategy for tweets, with internet labels serving as training data. Apart from these studies, ample research has recently been done, including a unique multi-task learning gating technique [38] and meta-based learning [39] for aspect-based sentiment analysis. Cambria et al. [40] uses the neuro-symbolic approach to create reliable symbolic representations that transform natural language into a kind of protolanguage and, as a result, extract polarities from the text in a perfectly easily understandable and explainable form. Cambria et al. [41] provides a morphological-aware concept parser that efficiently extracts emotional multi-word phrases from the text. The same concept can be used for various languages and modalities. MetaPro [42] is the first metaphor processing technique, which identifies metaphors in a sentence on a token level, paraphrases the recognized metaphors into their literal equivalents, and explains metaphoric multi-word formulations to predict the sentiment polarity. Prompt-based classification is getting increasing attention nowadays. Concerning this, an empirical study [43] finds that pre-trained models are biased in sentiment analysis and emotion detection tasks regarding class labels, emotional tag choices, prompt templates, and combinations of words of affective lexicons.

Excellent performance of deep learning models for script-based data motivated [44] to use ConvNet for TBSR. Tang et al. [45], ConvNet and LSTM were first utilized to generate tokens for the script-based data. Then a GRU was employed to encode phrase meanings and their underlying relationships for the final prediction. For more accurate analysis, [46] incorporates an attention module to consider the essential details and dismiss the rest. Both [45], [46] proposed document-based sentiment classification methods. Wang et al. [47] utilized attention mechanism along with LSTM architecture for aspect-based analysis.

Pre-trained language models based on transformers have gained popularity due to their ease of construction, efficient language representation, and comprehension abilities. For example, Dai et al. [48] employed RoBERTa to predict sentiment on the aspect level. Recently, various pre-trained BERT models have been utilized to extract relevant features from the textual content, such as [49] used the RoBERTa model to retrieve the most pertinent information from the textual modality to perform the task of news image captioning. In Tan et al. [50], a unique hybrid architecture combining a transformer-based and sequential-based model, RoBERTa-LSTM, was designed to analyze sentiment. This study used the robustly tuned BERT model to transform phrases into meaningful word embedding, whereas the LSTM approach efficiently captures the long-term contextual semantics. In Revathy et al. [51], machine learning algorithms with transfer learning and the BERT model analyzed the lyrical aspects crucial for identifying four important human emotions - happy, angry, relaxed, and sad. Out of all, BERT outperforms and achieves an overall accuracy of 92%. The most effective technological advancements in computer vision and natural language processing architecture and training methodologies are identified in [52] by statistically analyzing appropriate state-of-the-art methods.

2.1.2 Vision-Based Sentiment Recognition (VBSR)

Apart from writing, people are progressively eager to offer their thoughts with the help of images. As a result, VBSR is a critical field of study. Several studies have been undertaken to analyze sentiment using visual data. Images in VBSR are fed into a model to estimate emotion. Deep learning models have proved their superiority for visual data throughout the history of computer vision, and all research in VBSR has predominantly used deep learning models. In Joo et al. [53], a deep learning architecture has been utilized to understand the intent of the images. Jindal et al. [54] have used pre-trained ConvNet architecture [55] with the concept of transfer learning. The results of this study have been greatly enhanced by domain-specific fine-tuning. Wu et al. [56] provided an approach to decrease noise from the dataset using ANP sentiments and picture tags. Then CaffeNet was trained using softmax and the Euclidean loss function on the improved dataset for VBSR. The attention mechanism [57] has recently gained popularity since it concentrates on the relevant aspects while ignoring the irrelevant ones. Motivated by this, [58] incorporated an attention module and proposed a model named CECCN, which learns the correlation between color and content for predicting sentiment from images. A deep neural network architecture RA-DLNet has been designed in [59] to identify emotion in visual data. This strategy concentrates on sentiment-rich, locally relevant image regions by employing residual-based attention. To reduce the load of annotations, [60] solve the VBSR problem using a weakly supervised dual branch architecture. The first branch recognizes a sentiment-specific soft map, whereas the second examines comprehensive and regional information. The sentiment detection and classification branches are then combined and fed into a single deep framework to optimize the network from start to finish and, based on a similar strategy [61], proposed WSDEN model for the task of image sentiment analysis. Recently, Meena et al. [62] utilized different transfer learning models to recognize sentiment from the visual data on CK+, FER2013, and JAFFE datasets.

2.1.3 Acoustic Sentiment Recognition (ASR)

In addition to the script and vision-based data, researchers are increasingly interested in dealing with audio. Speech is preferable in some situations because it delivers more expressive signs of a speaker's feelings. Acoustic sentiment identification uses an audio dataset to estimate the speaker's sentiment, and the parameters employed to predict sentiment vary according to the level of utterances. In Negi et al. [63], depression was detected using audio samples. Luitel et al. [64] has utilized a multilingual dataset of audio to analyse sentiment.

TBSR, VBSR, and ASR use different approaches and achieve good results. However, while all the above methods process only a single modality, most social media platforms generate multi-modal content.

2.2 Multimodal Sentiment Recognition (MSR)

Sentiment analysis employing diverse modalities, commonly referred to as multimodal sentiment recognition, is a methodology that integrates information gathered from numerous sources to achieve a more precise understanding and interpretation of an individual's feelings and emotions. This approach captures a more extensive emotional context by employing the distinctive characteristics of each modality—acoustic for pitch and tone, text for explicit emotions, and visual information for body language and facial expressions. The various modalities in combination are displayed below:

- ***Text and Audio Analysis***

  Description**:** Combines textual data with audio data to enhance sentiment detection. Techniques**:** Joint modelling approaches using multi-task learning, feature fusion, and hybrid deep learning models that process both text and audio inputs. In the study referenced as [22], Bi-LSTM model have been employed to extract features from both textual and auditory input. Furthermore, a transformer-based model is utilized as a classifier for the final sentiment prediction.

- ***Text and Visual Analysis***

  Description**:** Integrates textual and visual information for a more comprehensive sentiment analysis.

  Techniques**:** Multimodal embedding techniques, image-caption models, and cross-modal attention mechanisms to jointly learn from text and images. Chen et al. [65] perform visual-caption sentiment recognition. In this proposed approach, the text data was encoded into a numeric representation using the word2vec algorithm. These vectors were then used as input to a text convolutional neural network (CNN). Similarly, visual features were extracted using an image CNN. Finally, an early fusion strategy was employed to combine the retrieved features for the purpose of making the final prediction. Similarly, Xu et al. [15] employed the Glove word embedding technique in conjunction with the text-CNN for textual modality and VGG-16 for visual feature extraction. Both the early and late fusion strategies have been employed in [15] to obtain the results.

- ***Audio and Visual Analysis***

Description: Combines audio and visual data to capture sentiment expressed through speech and facial expressions simultaneously.

Techniques: Multimodal feature extraction, synchronized processing of audio and visual streams using CNNs and RNNs, and attention mechanisms. Deshmukh et al. [66] performed facial emotion analysis by integrating images and acoustic information from the FER-2013 dataset.

- ***Text, Audio, and Visual Analysis***

  Description: Utilizes all three modalities to achieve the most comprehensive sentiment analysis.

  Techniques: Advanced multimodal deep learning frameworks, hierarchical fusion methods, and end-to-end models that learn from text, audio, and visual data simultaneously. Huddar et al. [23] utilized a text-CNN model for processing textual data and a 3D-CNN model for extracting visual features. Next, a Bidirectional Long Short-Term Memory (Bi-LSTM) model, combined with an utterance-level fusion strategy, was employed to make predictions. [67], [68], [69], [70] are several additional research studies conducted in the field of sentiment recognition using video datasets.

Researchers have established that the combined influence of multi-modal content mainly determines sentiment scores more accurately [71]. The same image with different words might elicit conflicting emotions. Therefore, a single modality is insufficient to predict sentiment; hence, it is necessary to analyze multi-modal data. MSR takes advantage of features from multiple modalities to analyze the sentiment. In visual-caption sentiment recognition, visual and caption modalities are processed to extract the features, and then both features are merged for sentiment polarity categorization. Xu et al. [72] introduced a model called MultiSentiNet, highlighting crucial visual scene and object properties to emphasize essential language phrases based on attention. Taking into account two separate modalities can influence and augment one another. To categorize multi-modal sentiment, [73] created a co-memory network that models the joint impacts of visual and textual content. Zhao et al. [74] developed a visual-textual consistency-driven strategy that employs visual cues, linguistic characteristics, social features, and visual-textual similarities. Poria et al. [68] utilized a framework based on LSTM to simulate the relationship among utterances of the videos for MSR. Wang et al. [75] used the Bag-Of-Words (BOW) model to predict the microblog's sentiments containing visual and textual data. You et al. [76] suggested a cross-modality regression technique incorporating visual and textual information for joint sentiment prediction. Xu et al. [77] designed a network based on an attention mechanism named HSAN, and visual captions were used as semantic information to analyse multi-modal sentiment. Motivated by aspect-level sentiment analysis employing textual content. Xu et al. [78] proposed a novel and challenging objective, aspect-based MSR. To express MSR on an aspect level, a recent study [79] employed the BART model to get text and aspect embeddings while using faster R-CNN for visual feature extraction. In addition to the aforementioned research papers, **Table 2.1** provides a summary of several more multimodal research studies based on various parameters.

Table 2.1 Summarization of multimodal research studies for sentiment prediction based on various parameters.

Ref.	Year	Techniques Used for Feature Extraction	Fusion Approach Involved	Dataset	Modalities	Results (Accuracy-A, F1 score-F1, Recall-R, Precision-P, Time-T, Loss-L, Macro F1 score-MF1, Mean absolute error-MAE, Correlation-C)
[15]	2017	Text-Glove and CNN Visual-CNN (based on VGG16)	EF & LF	MSVA-single MSVA-multiple	Image & text	Accuracy On MSVA-single using Late RMNN: **74.85 (for 2-class), and 67.09 (for 3-class)** Accuracy On MSVA-multiple using Late RMNN for two classes: **90.38** On MSVA-multiple using Early RMNN for three classes: **67.94**
[80]	2017	Text-Glove & CNN Visual-CNN (Select-Additive Learning (SAL))	TF	MOSI, YouTube, MOUD	videos (visual, text, and audio)	Accuracy using SAL-CNN (with the datasets): **0.73** On YouTube using SAL-CNN (across the datasets): **0.667** On MOUD using SAL-CNN (across the datasets): **0.574**
[67]	2017	Text-word2vec and CNN Visual-3D-CNN Audio-OpenSmile	ULF	CMU-MOSI	Videos (text, audio, and visual)	Accuracy features from AT fusion + CATF-LSTM: **81.30%**
[65]	2017	Text-word2vec & CNN Visual-CNN	EF	VSO, MVSO-EN (English language)	Image & text	On VSO using Deep fusion: **P-0.830, R-0.857, F1-0.844, A-0.847** On MVSO-EN using Deep fusion: **P-0.740, R-0.730, F1-0.735, A-0.737**
[68]	2017	Text-CNN, Visual-3D-CNN, Audio-OpenSmile	ULF	MOSI, MOUD,	Videos (text, audio,	On MOSI: **A:80.3%** On MOUD: **A:68.1%** On IEMOCAP: **A: 76.1%**

Ref.	Year	Techniques Used for Feature Extraction	Fusion Approach Involved	Dataset	Modalities	Results (Accuracy-A, F1 score-F1, Recall-R, Precision-P, Time-T, Loss-L, Macro F1 score-MF1, Mean absolute error-MAE, Correlation-C)
				and IEMOCAP	and visual)	
[81]	2018	Text-Bag-Of-Words Visual-OpenFacetoolkit Audio- librosa and CNN	EF & LF	Movie review and Music review video dataset	Text, audio, and visual	On cross-domain 1 (BoAW & BoVW & BoNG) using Early fusion: **Unweighted average recall-81.0%**
[82]	2018	Text-CNN, Visual-3D-CNN, Audio-OpenSmile (Contextual LSTM for final classification)	HF	Collected tweets by crawling Sina Weibo	Text, image, and emojis	Overall accuracy using WS-MDL: **A-0.695**
[83]	2018	Adjective-Noun Pairs (ANPs) response-based method	LF	Twitter 1269 and Twitter 603	Text and Image	Without Late Fusion**: P- 0.793, R-0.842, F1- 0.816, A- 0.751, T- 1.704** With Late fusion: **P- 0.804, R: 0.864, F1: 0.833, A- 0.772, T- 2.883**
[69]	2018	Bi-GRU	ULF	CMU-MOSEI and CMU-MOSI	Videos (text, audio, and visual)	On CMU-MOSEI: **79.80%** On CMU-MOSI: **A- 82.31%**
[70]	2019	Text, audio, and visual features- (Different variants of RNN) GRNN, LRNN, LGRNN, and UGRNN	EF (Attention Based)	CMU-MOSI	Text, audio, and visual	Using LRNN: **A-78.05%, L- 0.015** Using GLRNN: **A-78.05%, L-0.016**
[84]	2019	Text-GloVe & CNN Visual-CNN Audio-COVAREP software	HF	CMU-MOSI, CMU-MOSEI	Text and audio	On CMU-MOSI: **A-73.6%, F1-73.5%** On CMU-MOSI: **A-71.2%, F1-71.1%**

Ref.	Year	Techniques Used for Feature Extraction	Fusion Approach Involved	Dataset	Modalities	Results (Accuracy-A, F1 score-F1, Recall-R, Precision-P, Time-T, Loss-L, Macro F1 score-MF1, Mean absolute error-MAE, Correlation-C)
[85]	2019	Text- SentiWordNet & Gradient Boosting (SWN + GB) Visual- SentiBank	LF	Flickr 8k dataset, STS-Gold dataset,	Text and Image	Image module: **77.63%** Text module: **84.62%** Multimodal (proposed): **A- 91.32%**
[86]	2019	Text-LSTM Visual-Deep CNN	LF (Attention-Based)	Getty Images, Twitter, Flickr-w, and Flickr-m	Text and Image	On Getty Images: **P- 0.882, R- 0.851 F1-0.866, A- 0.869** On Twitter: **P- 0.778, R- 0.760, F1-0.769, A- 0.763** On Flickr-w: **P- 0.855, R- 0.845, F1-0.850, A- 0.859** On Flickr-m: **P- 0.882, R- 0.870, F1-0.876, A- 0.880**
[87]	2019	Text-LSTM Visual-Deep CNN	HF	Getty Images, Flickr	Text and Image	On Getty Images: **P- 0.871, R-0.854, F1:0.862, A- 0.865** On Flickr: **P- 0.847, R- 0.850, F1-0.848, A- 0.849** On Flickr-ML: **P- 0.880, R- 0.869, F1-0.874, A-0.878** On Flickr-IML: **P- 0.825, R- 0.833, F1- 0.829, A-0.831**
[88]	2020	Text- LSTM Visual- ResNet	HF	Twitter-14, Twitter-15, and Twitter- 17	Text and Image	On Twitter-15: **A-73.38%, MF1-67.37%** On Twitter-17: **A-67.83%,MF1-64.22%**
[89]	2020	Cross-modal BERT	TF	CMU-MOSEI, and CMU-MOSI	Videos (text, audio, and visual)	CM-BERT (T+A): **A7-44.9, A2-84.5, F1- 84.5, MAE-0.729, C- 0.791**
[22]	2020	Bi-LSTM for feature extraction from each modality and	ULF	CMU-MOSI, MELD, IEMOCAP	Videos (text, audio,	On CMU-MOSI: **A-82.71** On MELD: **A- 67.04 (**on text and audio**)** On MELD (Emotion prediction): **A-61.95 (**on text and audio**)**

Ref.	Year	Techniques Used for Feature Extraction	Fusion Approach Involved	Dataset	Modalities	Results (Accuracy-A, F1 score-F1, Recall-R, Precision-P, Time-T, Loss-L, Macro F1 score-MF1, Mean absolute error-MAE, Correlation-C)
		Transformers for actual prediction			and visual)	On IEMOCAP (Emotion prediction): **A-60.81**
[20]	2020	GRU	TF	CMU-MOSI and CMU-MOSEI	Videos (text, audio, and visual)	On CMU-MOSI: **A2- 81.19, F1-80.10** On CMU-MOSEI: **A2- 82.10, F1-80.0, MAE- 0.59**
[90]	2020	Bi-GRU feature extraction and Combination of contextual model, self-attention & cross interaction with a gated mechanism for final prediction	TF	CMU-MOSI and CMU-MOSEI	Opinionated Videos (text, audio, and visual)	On CMU-MOSI: **A-83.91, F1-81.17** On CMU-MOSEI: **A- 81.14, F1-78.53**
[21]	2020	Text-Glove & RNN Visual-3D-CNN Audio-librosa Library	ULF	CMU-MOSEI	Videos (text, audio, and visual)	On CMU-MOSEI: **A- 82.40%**
[91]	2021	Text-Bi-GRU Visual-VGG-19 Graph convolution network for sentiment classification	LF	Getty Images, Flickr	Text and Image	On Getty Images: **F1- 0.875, A- 0.871** On Flickr: **F1- 0.884, A- 0.878**
[18]	2021	Text-CNN Audio-LSTM Visual-CNN LSTM with Gating mechanism for	TF	CMU-MOSI, IEMOCAP, and CMU-MOSEI	Videos (text, audio, and visual)	On CMU-MOSI: **A2 -81.7, A7-35.4, F1- 81.8, MAE- 0.903, C-0.672** On CMU-MOSEI: **A2- 81.4, A7-50.6, F1- 81.6, MAE- 0.606, C- 0.673** On IEMOCAP for angry emotion got the cutting-edge result: **89.0, F1- 88.6**

Ref.	Year	Techniques Used for Feature Extraction	Fusion Approach Involved	Dataset	Modalities	Results (Accuracy-A, F1 score-F1, Recall-R, Precision-P, Time-T, Loss-L, Macro F1 score-MF1, Mean absolute error-MAE, Correlation-C)
		final sentiment prediction				
[92]	2021	Visual-VGG Text-Attention network	EF	MVSA-single, MVSA-multiple, and TumEmo	Text and Image	On MVSA-single: **0.7298, F1- 0.7298** On MVSA-multiple: **A- 0.7236, F1- 0.7230** On TumEmo: **0.6646, F1- 0.6339**
[93]	2021	Transformer	TF	MOSEI	Videos (text, audio, and visual)	On MOSEI in an unaligned setting: **A7- 51.5, A2- 81.8, F1- 81.8, MAE- 0.597, C- 0.671** On MOSEI in aligned setting: **A7- 51.0, A2- 82.2, F1- 82.4, MAE- 0.603, C- 0.662**
[23]	2021	Text-CNN Visual-3D-CNN Audio-OpenSmile and Bi-LSTM with attention for the final classification	ULF	IEMOCA, CMU-MOSI	Videos (text, audio, and visual)	On IEMOCAP using Bi-LSTM with attention for all three modalities: **A- 80.38%** On CMU-MOSI using Bi-LSTM with attention for all the three modalities: **A- 80.18%**
[26]	2021	Text-Glove & CNN Visual-CNN LSTM for polarity prediction	WLF	Flickr, Getty Images, and Twitter	Text and Image	On Flickr: **P-0.841, R- 0.836, F1- 0.834, A- 0.842** On Getty Images: P-0.832, R- 0.791, **F1- 0.810, A- 0.806** On Twitter at least 5-agree: **P-0.880, R-** 0.866**, F1- 0.862, A- 0.863** On Twitter at least 4-agree: **P-0.831,** R- 0.812**, F1- 0.821, A- 0.771** On Twitter at least 3-agree: **P-0.793,** R- 0.776**, F1- 0.788, A- 0.742**
[94]	2021	Visual-encoder (50-layer Residual network) Text-Bi-LSTM	HF	MASAD	Text and Image	On MASAD: **A- 95.63, F1- 95.09**

Ref.	Year	Techniques Used for Feature Extraction	Fusion Approach Involved	Dataset	Modalities	Results (Accuracy-A, F1 score-F1, Recall-R, Precision-P, Time-T, Loss-L, Macro F1 score-MF1, Mean absolute error-MAE, Correlation-C)
[95]	2021	Visual- ResNet, DenseNet Text-CNN gradient Boosting machine learning for final sentiment prediction	LF	B-T4SA	Text and Image	On B-T4SA using AutoML-based Fusion: **A- 95.19%**
[96]	2021	Transformer	ULF	CMU-MOSI	Videos (text, audio, and visual)	On CMU-MOSI: MAE- 0.644, **C- 0.842, A2- 89.33, F1- 89.31, A7- 47.52**

Although previous research endeavours have made a significant impact in the integration of images, words, and audios, additional investigation is required to enhance the overall outcomes. This inspired us to develop a model integrating meaningful information from multiple modalities.

2.3 Research Gaps

- ❖ **RG1:** Prediction of Emoji using multiple modalities is less explored.
- ❖ **RG2:** Recent multimodal strategies are computationally expensive.
- ❖ **RG3:** There is a lack of effective sarcasm detection approaches using multimodal content.
- ❖ **RG4:** Few research studies have been conducted for aspect-based multimodal sentiment analysis.
- ❖ **RG5:** A limited number of application-based research studies have been conducted in multimodal sentiment analysis, such as box-office prediction, news focus detection, etc.
- ❖ **RG6:** No multilingual datasets are available with multiple modalities and methods for handling such datasets.

2.4 Research Objectives

The proposed objectives are based on identified research needs:

- **RO1:** To propose a novel framework for predicting sentiment using images and captions.
- **RO2:** To propose a novel framework for aspect-based multimodal sentiment analysis.
- **RO3:** To propose an effective method for predicting "emoticons" in multimodal content.
- **RO4:** To propose a novel framework for predicting sentiment from videos having visual, audio and textual (subtitles) information.
- **RO5:** To propose a novel framework for the detection of sarcasm for multimodal data.

2.5 Research Contributions

The main objective of the thesis is to design and develop novel architectures capable of identifying sentiments in multimodal content. Hence, the following architectures and frameworks are proposed to accomplish this:

- **VABDC-Net** Depending on the context, the same phrase may generate different sentiments in several scenarios. Hence, it is essential to use both visual and textual content for more accurate prediction. Motivated by this, we developed a novel Visual Attention and Bi-Directional Caption Processing network (VABDC-Net) for visual-caption sentiment recognition tasks. The proposed methodology more effectively analyses the interaction between visual and captions modalities than earlier innovative methods. The model is organized into three components: an attentional tokenizer-based bi-directional caption branch for extracting features from the captions, an attentional visual branch for visual feature extraction, and cross-modal feature fusion.

- **VECT-Net** Proposed a novel approach called the Visual-to-Emotional-Caption Translation Network (VECT-Net) technique. The primary objective of this strategy is to effectively acquire visual sentiment clues by analysing facial expressions. Additionally, it effectively aligns and blends the obtained emotional clues with the target attribute of the caption mode. This study aims to employ Target-Dependent Multimodal Sentiment Recognition (TDMSR) to identify the level of sentiment associated with every target (aspect) stated within a multimodal post consisting of a visual-caption pair.

❖ Developed a multimodal architecture based on the principle of contrastive learning for the emoticon prediction task to effectively simulate the relationship and compatibility between image and text content. The proposed multimodal architecture comprises of three primary components: An Image encoder, a Text encoder, and a Contrastive learning component. The Image encoder is responsible for acquiring image embeddings, while the Text encoder acquires textual embeddings. The Contrastive learning element examines the pertinent attributes and similarities between the textual and image embeddings obtained in the preceding steps. The proposed model employs the joint training of dual-branch encoder along with the contrastive learning to accurately map text and images into a common latent space. Our key finding is that by integrating the principle of contrastive learning with that of the other two branches yields superior results.

❖ Introduced a deep-learning-based model for Multimodal Sarcasm Detection, that combines DistilBERT and RegNet to efficiently capture meaningful information across text, visual, and acoustic modalities in a single task, To check the robustness of our model so that it can effectively and reliably imitate the multi-modal depiction of textual, visual & acoustic modalities and yield ground-breaking results for Multimodal Sarcasm Detection, we perform extensive analysis and test our proposed approach using one of the benchmark datasets, MUStARD.

The following research works form the basis of this chapter:

❖ **A. Pandey** and D. K. Vishwakarma, "Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey," ***Applied Soft Computing***, vol. 152, November. 2024, doi: https://doi.org/10.1016/j.asoc.2023.111206.

Chapter 3: Visual-Caption Sentiment Recognition

3.1 Scope of this Chapter

This chapter focuses on the area of VCSR (Visual Caption Sentiment Recognition) by proposing a novel approach which seeks to interpret a picture and text combination by integrating textual (spoken words) and visual modalities (facial expressions). A novel architecture named VABDC-Net (Visual Attention and Bi-Directional Caption Processing Network) have been proposed. The model's framework, comprised of three modules: an attentional tokenizer-based bidirectional caption expert branch to retrieve useful textual features, an attention visual expert branch to retrieve appropriate visual features, and a cross-modal feature fusion module to merge features and predict sentiment. For convenience, image-text sentiment polarity analysis is referred to as visual-caption sentiment recognition. Based on our findings, it is evident that our method VABDC-Net outperforms the other models, indicating that it is better able to learn the variety of different feature modalities and perform more robustly. Due to the attentional tokenizer-based bi-directional caption and spatial-depth visual attention modules, VABDC-Net outperforms the other models. The attention modules essentially improve the interactions between two categories of modality information; particularly, it enables caption information to aid in the acquisition of visual features (and vice versa) to achieve balanced learning of both sorts of modalities.

3.2 VABDC-Net: A Framework for Visual-Caption Sentiment Recognition via Spatio-Depth Visual Attention and Bi-Directional Caption Processing

3.2.1 Abstract

People are becoming accustomed to posting images and captions on social media platforms to express their opinions. Hence, Visual-Caption Sentiment Recognition (VCSR) has been a subject of growing attention recently. Thus, the correlation between visual and caption modalities is crucial for VCSR. However, most recent VCSR strategies concatenate features from the visual and caption modalities with the help of pre-trained deep learning models containing millions of trainable parameters without adding a dedicated attention module, ultimately leading to less desirable results. Motivated by this observation, we have proposed a novel model VABDC-Net, that integrates an attention module with the convolutional neural network to focus on the most relevant information from the visual modality and attentional

tokenizer-based method to extract the most relevant contextual information from the caption modality. Demanding to this dire need, the following are the significant contributions of our experimentation: (1) an attentional tokenizer-based bi-directional caption branch to retrieve useful textual features from the captions, (2) an attentional visual branch to retrieve appropriate visual features, and (3) a cross-domain feature fusion to merge multi-modal features and predict sentiment. Thorough experimentation on two benchmark datasets, Twitter-15, with an accuracy of $\mathbf{83.80\%}$, and Twitter-17, with an accuracy of $\mathbf{72.42\%}$, indicates that our technique outperforms existing methods for VCSR.

3.2.2 Proposed Methodology

In this section, the proposed framework of VCSR is thoroughly discussed. The problem is defined in the first portion of this section. Then, we present the model's framework, comprised of three modules: an attentional tokenizer-based bidirectional caption expert branch to retrieve useful textual features, an attention visual expert branch to retrieve appropriate visual features, and a cross-modal feature fusion module to merge features and predict sentiment. For convenience, image-text sentiment polarity analysis is referred to as visual-caption sentiment recognition.

Problem definition

Visual-caption sentiment recognition problem is briefly summarized as follows. Assume that $\boldsymbol{S}$ and $\boldsymbol{V}$ illustrate a caption and a visual sample space, respectively, where one string of caption and its accompanying visual information constitute an example. Each example belongs to a class label, $\boldsymbol{C^j}$. In other words, each example is a triplet that includes a caption, a piece of visual information, and a class label. It can be expressed as follows:

$$\boldsymbol{E} = \{(\boldsymbol{S^0}, \boldsymbol{V^0}, \boldsymbol{C^0}), (\boldsymbol{S^1}, \boldsymbol{V^1}, \boldsymbol{C^1}), \dots \dots, (\boldsymbol{S^j}, \boldsymbol{V^j}, \boldsymbol{C^j}), \dots \dots, (\boldsymbol{S^{n-1}}, \boldsymbol{V^{n-1}}, \boldsymbol{C^{n-1}})\}, \qquad (\mathbf{1})$$

where $\boldsymbol{E}$ is a collection of example triplets, $\boldsymbol{S^j}$ denotes caption information, $\boldsymbol{V^j}$ denotes visual information, $\boldsymbol{C^j}$ is a class label of the caption-visual pair in the $\boldsymbol{j^{th}}$ example, and $\boldsymbol{n}$ is the count of the number of samples in a dataset. VCSR aims to learn a mapping function $\boldsymbol{F}: (\boldsymbol{S}, \boldsymbol{V}) \longrightarrow \boldsymbol{C}$ from the multi-modal training examples $\{(\boldsymbol{S^j}, \boldsymbol{V^j}, \boldsymbol{C^j}) \big| \mathbf{0} \leq \boldsymbol{i} \leq \boldsymbol{n} - \mathbf{1}\}$. For a sentiment polarity categorization task, $\boldsymbol{C^j} \in \{\boldsymbol{positive}, \boldsymbol{negative}, \boldsymbol{and\ neutral}\}$.

Visual Attention and Bi-Directional Caption Processing Network (VABDC-Net)

We proposed VABDC-Net for the visual-textual sentiment recognition task to simulate the interaction between caption and visual information and to examine the compatibility between caption and visual content. **Figure 3.1** depicts the VABDC-Net framework. The model is organized into three components: an attentional tokenizer-based bi-directional caption branch for extracting features from the captions, an attentional visual branch for visual feature extraction, and cross-modal feature fusion. The details of the proposed framework have been discussed below. The proposed framework is presented in algorithmic form as
Table 3.3.

Attentional Tokenizer-Based Bidirectional Caption Branch

This section thoroughly explains the proposed approach for extracting crucial information from caption modality.

Attentional Tokenization

Inputs for NLP models like [97] must be numeric vectors, which often require transforming attributes like vocabulary into numbers. The BERT model, released by Google [98] in 2019, is a member of the group of NLP-based language models known as transformers and includes tokenizers to convert phrases into a numeric representation. It is superior to previous word-embedding approaches, such as TF-IDF, Word2Vec, Glove, etc., because it has been pre-trained on vast text datasets. Hence, the BERT model's tokenizer produces superior word embeddings. For generating embeddings of words, the BERT model uses the attention mechanism. As a result, it captures associations for each word depending on the words on both sides of the phrase. Positionally encoded word embeddings keep track of the sequence and arrangement of each word in a phrase. As a result, it delivers high-quality context-aware or contextualized word embeddings by traversing each BERT encoder layer. Hence, the proposed framework BERT-base version (110 million parameters) with 12 blocks of the transformer, 768 hidden layer sizes, and 12 attention heads have been employed for pre-processing and converting the captions into a sequence of vectors. For a given caption, $\boldsymbol{S^j}$ the caption with m words can be denoted as $\boldsymbol{S^j} = \{\boldsymbol{w_{j1,}} \boldsymbol{w_{j2,}} \dots \dots \dots., \boldsymbol{w_j}\}$. Each word $\boldsymbol{w_{jk}}$ is embedded in a vector representation where $\boldsymbol{v_j} \in \ \boldsymbol{\Re^d}$ is the $\mathcal{d}$-dimension vector for $\boldsymbol{k^{th}}$ word. The final tokenization is denoted as $\mathbb{V}^{\boldsymbol{j}} = \{\boldsymbol{v_{j1,}} \boldsymbol{v_{j2,}} \dots \dots \dots., \boldsymbol{v_{jn}}\}$. Illustration of word to a vector representation using a BERT-base encoder is depicted in **Figure 3.2**.

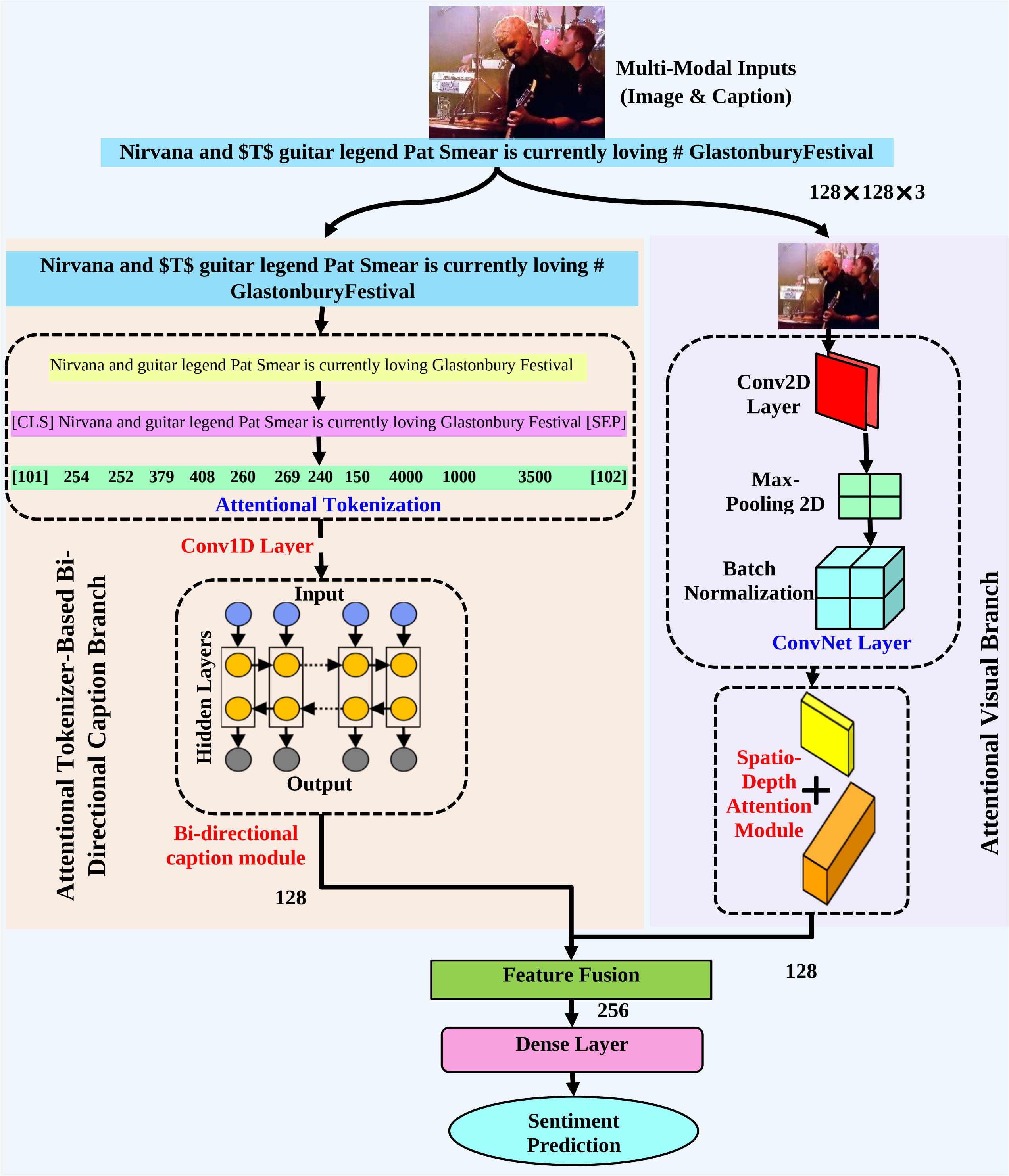

Multi-Modal Inputs
(Image & Caption)
Nirvana and T guitar legend Pat Smear is currently loving # GlastonburyFestival
128✕128✕3
Nirvana and T guitar legend Pat Smear is currently loving # GlastonburyFestival
Nirvana and guitar legend Pat Smear is currently loving Glastonbury Festival
[CLS] Nirvana and guitar legend Pat Smear is currently loving Glastonbury Festival [SEP]
[101] 254 252 379 408 260 269 240 150 4000 1000 3500 [102]
Attentional Tokenization
Conv1D Layer
Input
Hidden Layers
Output
Bi-directional caption module
Attentional Tokenizer-Based Bi-Directional Caption Branch
128
Conv2D Layer
Max-Pooling 2D
Batch Normalization
ConvNet Layer
Spatio-Depth Attention Module
+
Attentional Visual Branch
128
Feature Fusion
256
Dense Layer
Sentiment Prediction

Figure 3.1 Proposed visual attention and bi-directional caption processing network (VABDC-Net) for the visual-caption sentiment recognition.

Bi-Directional Caption Branch

Various sequential models can be used to extract features from script-based information. Traditionally, conventional RNNs (Recurrent neural networks) were used to tackle the sequential data. The main advantage of using RNN instead of a standard neural network is that the weights were not shared in conventional neural networks. There are other advantages, too, such as this type of network can recall their previous inputs. Apart from several benefits, standard RNNs, aren't very good at capturing long-term dependencies. This is fundamentally correlated to the issue of vanishing gradients, which means that while training an intense network, gradients drop off exponentially as they move down the layers.

To overcome this problem, [97] was introduced. This network can recall previous inputs for the longest period and provide a wide range of hyperparameters such as learning rate, weights, and biases for input and output. Hence, there is no need for fine adjustment. Apart from this, the complexity of updating the weights is reduced to O(1), which is also an advantage. Although these became popular because they could overcome the problem of diminishing gradients. However, they fail to eradicate it. The issue is that the data must still be sent from cell to cell for analysis. Furthermore, the cell has grown rather complicated with new features (such as forget gates). These models can't maintain the details from the past and future because, in such networks, data can flow either in a forward or backward direction.

To overcome these drawbacks, [99] were developed. In these networks, information flows in both directions. Hence it can model sequential dependencies between words in both directions of the sequence, and as a result, the past and future details are preserved. Due to this reason, such architectures produce a more meaningful output. Motivated by these advancements, [99] have been utilized in the proposed framework to extract features from the caption modality. For caption feature extraction, the embedding vector $\mathbb{V}^{j} = \{\boldsymbol{v_{j1,}} \boldsymbol{v_{j2,}} \dots \dots \dots., \boldsymbol{v_{jn}}\}$, obtained from an attentional tokenization module, is fed to the three consecutive layers of convolution 1D to reduce the number of features.

$$\boldsymbol{T^{j}} = \boldsymbol{Conv1D}(\mathbb{V}^{j}) \qquad (\mathbf{3.2})$$

The acquired feature vector $\boldsymbol{T^{j}}$ is then processed through three consecutive layers of [99], indicated as ∂, as well as one dense layer and batch normalization, to extract the final feature vector $\boldsymbol{F^{j}}$ from the caption modality. The caption feature of $\boldsymbol{S^{j}}$ was obtained as $\boldsymbol{F_{s}^{j}}$ in **Eqn.**$(\mathbf{3.4})$

$$\boldsymbol{F^{j}} = \boldsymbol{\partial}(\boldsymbol{T^{j}}), \boldsymbol{j} \in [\mathbf{1}, \boldsymbol{n}] \qquad (\mathbf{3.3})$$

$$\boldsymbol{F_{s}^{j}} = \{\boldsymbol{f_{s}^{1}}, \boldsymbol{f_{s}^{2}}, \boldsymbol{f_{s}^{3}}, \dots \dots \dots \dots \dots \dots \dots, \boldsymbol{f_{s}^{n}}\} \qquad (\mathbf{3.4})$$

Spatial dropout of 1-dimensional was also utilized to prevent overfitting. **Table 3.1** shows the architectural details of the proposed model for caption feature extraction.

Table 3.1 Architectural details of the proposed framework for caption feature extraction.

Components of the proposed framework for caption feature extraction	Hyperparameters
Embedding size	512
Three consecutive layers of convolution 1D	128, 64, and 32 units
Max-pooling 1D with pool size	2
Stride	1
Layer 1 of ∂	64 units
Layer 2 of ∂	128 units
Layer 3 of ∂	256 units
Activation function except for the dense layer	Leaky ReLu
Dense layer	128 units
Spatial dropout	0.4
Optimizer	Adam
Learning rate	0.001
Loss function	Categorical cross-entropy
Activation function for a dense layer	Softmax

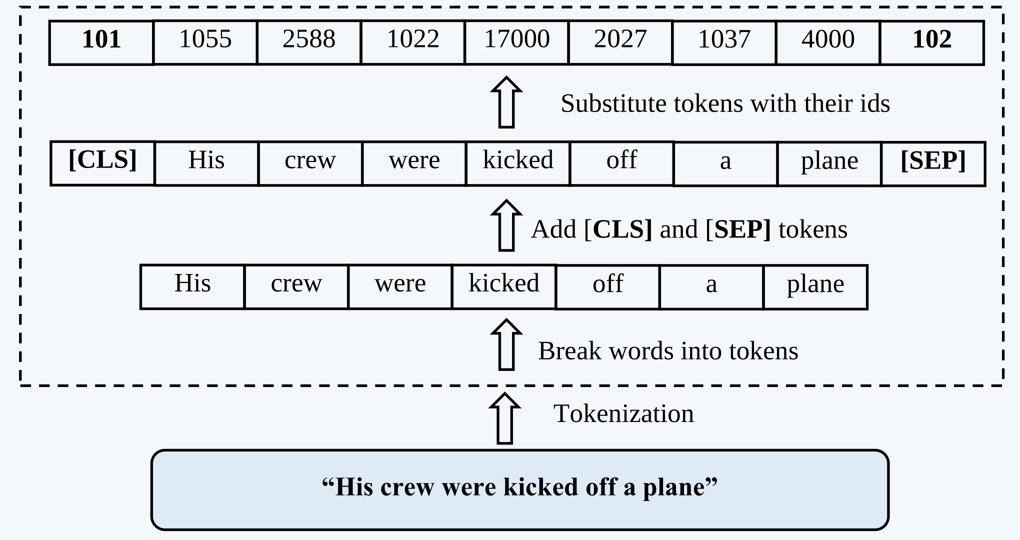

101 1055 2588 1022 17000 2027 1037 4000 102
Substitute tokens with their ids
[CLS] His crew were kicked off a plane [SEP]
Add [CLS] and [SEP] tokens
His crew were kicked off a plane
Break words into tokens
Tokenization
"His crew were kicked off a plane"

Figure 3.2 Illustration of phrase to a vector representation using BERT tokenizer.

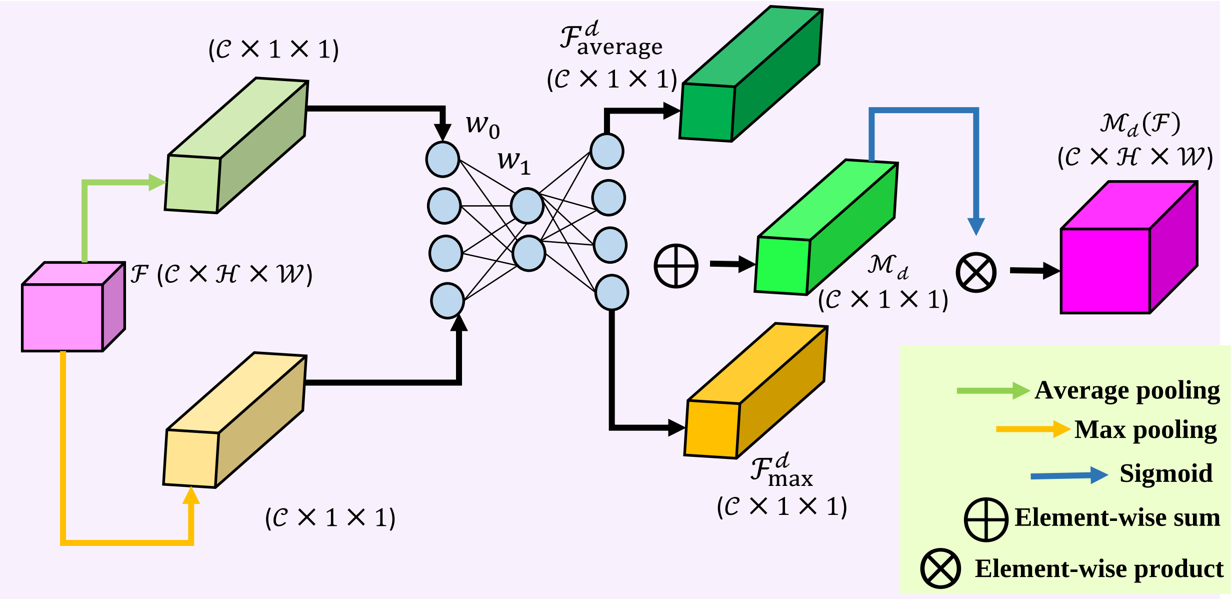
Attentional Visual Branch

This section will thoroughly explain the proposed method for extracting crucial information from visual modality for recognizing sentiment.

Spatio-Depth Attention

Excellent representational capabilities of ConvNets have substantially improved their performance on visual tasks. To enhance ConvNet performance, recent research has concentrated on three crucial network properties: depth, breadth, and cardinality. In addition to these aspects, we look into another factor of architectural design called attention. The importance of attention has been well investigated in earlier research [100], [101]. By leveraging attention mechanisms, such as focusing on crucial attributes and suppressing irrelevant ones, we hope to boost the strength of representation. In the proposed framework, the spatio-depth attention module [102] is integrated with the layers of ConvNet to emphasize meaningful information along the depth and spatial axes. To do this, we implemented depth and spatio attention modules, which enable our model to learn "what" and "where" to pay attention along the depth and spatial axes, respectively. This improves network information flow by determining which aspect to emphasize or ignore.

The term "depth" often refers to the number of channels, which are effectively the feature maps layered in a tensor, in which each cross-sectional slice is simply a feature map with depth ($\boldsymbol{\mathcal{H}} \times \boldsymbol{\mathcal{W}}$). The depth attention simply assigns a value to each channel; as a result, the channels that contribute the most to learning are prioritized for refinement, thereby improving the model's overall performance. And the term “spatio” refers to the overall domain space enclosed by each feature map. By enhancing the feature maps using the spatio attention module, we feed the refined input to the consecutive convolutional layers, boosting the model's effectiveness. Therefore, the attention module used to extract the most relevant features from the visual modality is collectively referred to as “spatio-depth attention.”



$(\mathcal{C} \times 1 \times 1)$
$\mathcal{F}^d_{\text{average}}$
$(\mathcal{C} \times 1 \times 1)$
W_0
W_1
$\mathcal{M}_d(\mathcal{F})$
$(\mathcal{C} \times \mathcal{H} \times \mathcal{W})$
$\mathcal{F}\ (\mathcal{C} \times \mathcal{H} \times \mathcal{W})$
$\mathcal{M}_d$
$(\mathcal{C} \times 1 \times 1)$
$(\mathcal{C} \times 1 \times 1)$
$\mathcal{F}^d_{\text{max}}$
$(\mathcal{C} \times 1 \times 1)$
Average pooling
Max pooling
Sigmoid
⊕ Element-wise sum
⊗ Element-wise product

Figure 3.3 Depth attention module.

In the depth attention module (**Figure 3.3**), the average and max-pooling operations are employed to aggregate the spatial information of a feature, yielding two distinct spatial context attributes: $\boldsymbol{\mathcal{F}^d_{average}}$and $\boldsymbol{\mathcal{F}^d_{max}}$, which stand for average and max-pooled features explicitly. Following that, the depth attention feature map, $\boldsymbol{\mathcal{M}_d \in \mathcal{R}^{\mathcal{C}\times 1\times 1}}$ is created using a multi-layer perceptron. After applying the multi-layer perceptron to each feature, the resultant feature vectors are element-wise summed and passed to a sigmoid activation function. In brief, the depth attention is calculated as follows:

$$\boldsymbol{\mathcal{M}_d(\mathcal{F}) = sigmoid\left(w_1\left(w_0(\mathcal{F}^d_{average})\right) + w_1\left(w_0(\mathcal{F}^d_{max})\right)\right)} \qquad \mathbf{(3.5)}$$

where, $\boldsymbol{w_1}$ and $\boldsymbol{w_0}$ are the shared input weights of the multi-layer perceptron.

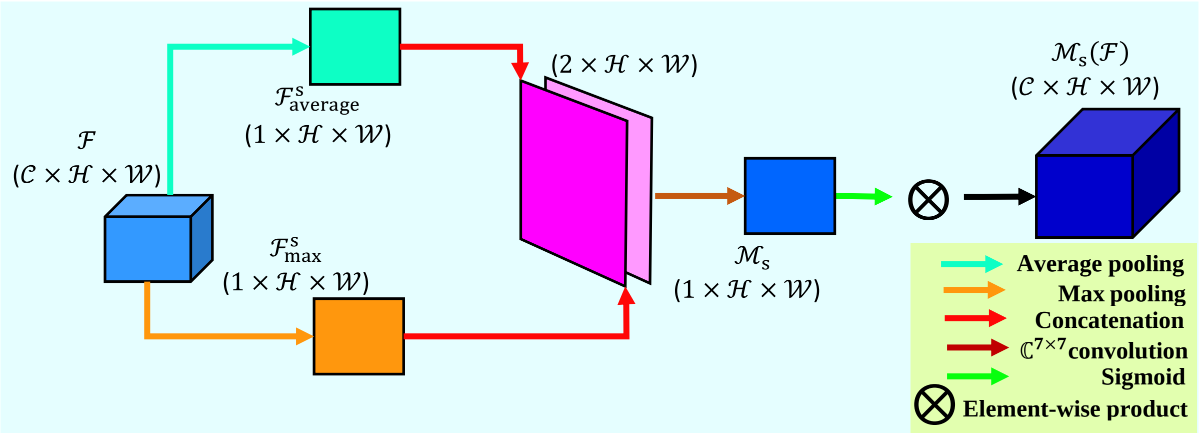

$\mathcal{F}$
$(\mathcal{C} \times \mathcal{H} \times \mathcal{W})$
$\mathcal{F}^s_{\text{average}}$
$(1 \times \mathcal{H} \times \mathcal{W})$
$\mathcal{F}^s_{\text{max}}$
$(1 \times \mathcal{H} \times \mathcal{W})$
$(2 \times \mathcal{H} \times \mathcal{W})$
$\mathcal{M}_s$
$(1 \times \mathcal{H} \times \mathcal{W})$
$\mathcal{M}_s(\mathcal{F})$
$(\mathcal{C} \times \mathcal{H} \times \mathcal{W})$
Average pooling
Max pooling
Concatenation
$\mathbb{C}^{7\times7}$convolution
Sigmoid
⊗ Element-wise product

Figure 3.4 Spatio attention module.

Meanwhile, the spatio attention module (**Figure 3.4**) comprises three successive operations. The first portion is called the depth Pool, and it divides the input matrix of dimensions $(\boldsymbol{\mathcal{C} \times \mathcal{H} \times \mathcal{W}})$ into $(\boldsymbol{2 \times \mathcal{H} \times \mathcal{W}})$, with each channel illustrating max and average pooling across the depth. The obtained matrix of size $(\boldsymbol{2 \times \mathcal{H} \times \mathcal{W}})$ is fed as an input to the ConvNet layer, which generates a 1-depth feature vector with a dimension of $(\boldsymbol{1 \times \mathcal{H} \times \mathcal{W}})$. Hence, this results in two distinct features $\boldsymbol{\mathcal{F}^s_{average} \in \mathcal{R}^{1\times\mathcal{H}\times\mathcal{W}}}$ and $\boldsymbol{\mathcal{F}^s_{max} \in \mathcal{R}^{1\times\mathcal{H}\times\mathcal{W}}}$. As a result, this ConvNet layer $\mathbb{C}^{7\times7}$ is a spatial dimension preserving convolution. The obtained result is then delivered to a sigmoid activation function, which converts all values to a range between 0 and 1. The obtained resultant spatio attention mask, $\boldsymbol{\mathcal{M}_s \in \mathcal{R}^{1\times\mathcal{H}\times\mathcal{W}}}$ is then applied to all the input tensor feature maps using element-wise multiplication. In summary, the spatio attention is determined as follows:

$$\boldsymbol{\mathcal{M}_s(\mathcal{F}) = sigmoid\left(\mathbb{C}^{7\times7}([average - pooling(\mathcal{F}); max - pooling(\mathcal{F})])\right)}$$

$$\boldsymbol{\mathcal{M}_s(\mathcal{F}) = sigmoid\left([\mathcal{F}^s_{average}\, ; \mathcal{F}^s_{max}\,]\right)} \qquad \mathbf{(3.6)}$$

Baseline Architecture for Visual Feature Extraction

Deep learning-based architectures outperform traditional handcrafted-based classification algorithms in visual, textual, and acoustic recognition. The key to this achievement is the utilization of massive datasets, the extensive usage of GPU cards, and the evolution of deep architectures. ConvNet-based deep learning approaches have surpassed other deep learning models in image categorization because, in these approaches, features from visual data have been extracted using the localization principle in the convolution layer, and the image is scaled down to fewer dimensions with more defining characteristics in the pooling layer. However, training massive deep models, on the other hand, takes time and requires expensive GPU resources. Considering these factors, we have designed a customized lightweight ConvNet of 5 layers incorporating a spatio-depth attention module to reduce computational cost. We have defined the visual-caption pair as $(\boldsymbol{S}, \boldsymbol{V})$, where $\boldsymbol{S}$ and $\boldsymbol{V}$ denote caption and visual modality explicitly. The $\boldsymbol{n}$ is the number of visual-caption pairs in the corpus. We utilized **Eqn.** $(\mathbf{3.7})$ to extract features from a given $\boldsymbol{V^j}$. The visual feature of $\boldsymbol{V^j}$ was obtained as $\boldsymbol{F_v^j}$ in **Eqn.** $(\mathbf{3.8})$**.** This architecture has a total of 246265 trainable parameters, which is relatively low, making it lightweight. **Table 3.2** shows the architectural details of the proposed model used for visual feature extraction.

$$\boldsymbol{F^j_{ConvNet+spatio-depth\,attention}} = \boldsymbol{ConvNet + spatio - depth\,attention}(\boldsymbol{V^j}), \boldsymbol{j} \in [\mathbf{1}, \boldsymbol{n}] \qquad (\mathbf{3.7})$$

$$\boldsymbol{F_v^j} = \{\boldsymbol{f_v^1}, \boldsymbol{f_v^2}, \boldsymbol{f_v^3}, \dots\dots\dots\dots\dots., \boldsymbol{f_v^n}\} \qquad (\mathbf{3.8})$$

Table 3.2 Architectural details of the proposed framework for visual feature extraction.

Components of the proposed framework for visual feature extraction	Hyperparameters
Image shape	128×128×3
5 Layers of Conv2D	32, 64, 128, 64, and 64 units
Kernel size	3×3
Max-pooling 2D pool size	2×2
Stride	1
Activation function except for the dense layer	ReLu
Dense layer	128 units
Spatial dropout	0.1
Optimizer	Adam
Learning rate	0.001
Loss function	Categorical cross-entropy
Activation function for a dense layer	Softmax

Feature Fusion

Multi-modal data in a social media post communicate a user's opinion. In visual-caption sentiment identification, the correct expression of the post's meaning is ensured by appropriately constructing the link between modalities. In several studies, remarkable results were reached utilizing early fusion, whereas, in others, high performance was obtained using late fusion. In this research study, we conducted feature-level fusion. The merging of information from distinct layers or branches, known as feature fusion, is an essential component of current network architectures. It is typically accomplished through fundamental operations such as summation or concatenation. The visual-caption pair characteristics were defined as $(\boldsymbol{S}, \boldsymbol{V})$. **Eqn.** $(\mathbf{3.4})$ and $(\mathbf{3.8})$ showed that $\boldsymbol{F_s^j}$ and $\boldsymbol{F_v^j}$ reflected the visual and caption attributes of the $\boldsymbol{j}$ visual and caption pairs, respectively. **Eqn.** $(\mathbf{3.9})$ integrates the characteristics extracted from the visual and the caption modality, where $\boldsymbol{n}$ represents the collection's sample count and $\oplus$ denotes the concatenation operation for implementing feature fusion. The obtained vector, after fusion, is then sent into a softmax layer for sentiment recognition.

$$\boldsymbol{F^j} = \boldsymbol{F_s^j} \oplus \boldsymbol{F_v^j} \in [\mathbf{1}, \boldsymbol{n}] \qquad (\mathbf{3.9})$$

Table 3.3 Algorithm for the proposed framework.

```
Algorithm 1: Visual-Caption Sentiment Recognition based on visual attention and bi-directional caption processing network (VABDC-Net)
Aim: To learn a mapping function $F: (S^j, V^j, C^j) \longrightarrow$ from the multi-modal training examples $\{(S^j, V^j, C^j) | 0 \le i \le n - 1\}$.
Input: Caption set $S^j = \{w_{j1,} w_{j2,} \dots\dots\dots., w_j\}$ and visual set $V^j$
Output: Sentiment polarity categorization task, $C^j \in \{positive, negative, and\ neutral\}$
1.  Tokenization of words for the entire caption set $\mathbb{V}^j$;
2.  Extract features from the caption content $F_s^j$;
3.  Extract features from the visual content $F_v^j$;
4.  for $E \leftarrow 1$ to Epochs do
        $\mathbb{V}^j = \{v_{j1,} v_{j2,} \dots\dots\dots., v_{jn}\} \leftarrow S^j = \{w_{j1,} w_{j2,} \dots\dots\dots., w_j\}$ words to vector representation;
        $F_s^j = \{f_s^1, f_s^2, f_s^3, \dots\dots\dots\dots\dots\dots\dots, f_s^n\} \leftarrow \mathbb{V}^j = \{v_{j1,} v_{j2,} \dots\dots\dots., v_{jn}\}$ caption feature extraction by Eqn. (3.2), Eqn. (3.3), and Eqn. (3.4);
        $\mathcal{M}_d(\mathcal{F})$ and $\mathcal{M}_s(\mathcal{F})$ refined feature map from the visual modality by Eqn. (3.5), Eq. (3.6);
        $F_v^j = \{f_v^1, f_v^2, f_v^3, \dots\dots\dots\dots\dots., f_v^n\} \leftarrow \mathcal{M}_d(\mathcal{F})$ and $\mathcal{M}_s(\mathcal{F})$ by Eqn. (3.7), Eqn. (3.8);
        $F^j = F_s^j \oplus F_v^j \in [1, n]$ concatenation of the features by Eqn. (3.9);
        $C^j \in \{positive, negative, and\ neutral\} \leftarrow F^j$ by fully connected and softmax;
        Calculate the loss and perform backpropagation;
5.  end
```

3.2.3 Experimental Results and Discussion

This section contains detailed information regarding the dataset utilized during the research, the experimental settings of the proposed framework, and performance assessments.

Dataset Description

Two separate datasets acquired from Twitter were utilized in this study to validate the performance of our proposed framework. Twitter-15 and Twitter-17 [71] are the datasets used to evaluate our model. In both datasets, each tweet was composed of visual-caption pair. The Twitter-15 dataset has 5347 tweets, and the Twitter-17 dataset contains 5972 tweets. Each visual-caption pair in these datasets was categorized as positive, negative, or neutral based on its caption and visual modality. In both datasets, “0” refers to the negative, “1” refers to the neutral, and “2” refers to the positive class samples. We utilize a random split of the data in the ratio of 8:1:1 and compare accuracy, macro-precision, macro-recall, and macro-F1scores as metrics to the baseline approaches.

Table 3.4 Statistics of the Twitter-15 and Twitter-17 dataset.

Dataset	Positive	Negative	Neutral	Total
Twitter-15	1548	630	3169	5347
Twitter-17	2516	728	2728	5972

Implementation Details

To compare with prior cutting-edge approaches, datasets were randomly divided into 80% train, 10% validation, and 10% test sets. The training and validation sets were used to train both dataset’s attentional tokenizer-based bidirectional caption and visual attentional branch models to extract deep features. The proposed framework has been run for 50 epochs with a batch size of 32 while categorizing the sentiments to evaluate the results. We examined multiple values for each hyperparameter randomly and then freeze those values that generated the best results for our proposed framework. **Table 3.1** and **Table 3.2** provide the hyperparameters of the proposed methodology used in the study. Accuracy($\boldsymbol{\mathcal{A}}$), macro-precision($\boldsymbol{M\mathcal{P}}$), macro-recall($\boldsymbol{M\mathcal{R}}$), and macro-F1($\boldsymbol{M\mathcal{F}1}$) are the performance metrics used for evaluating our proposed model. The proposed model used in the study was trained and evaluated using Google COLAB pro plus with NVIDIA V100 GPU, 40GB of graphics memory, driver version 460.32.03, CUDA version 11.2, 80GB of RAM, and 100GB of hard disc space. For our experiment, we used up 20.1 GB of RAM, 12.2 GB of graphics memory, and 30.3 GB of disc space. Models were generated using the TensorFlow and Keras frameworks to extract meaningful information from visual-caption pairings.

Experimental Results and Analysis

We evaluate our proposed framework on two benchmark datasets. Note that we have used Twitter-15 and Twitter-17 datasets for visual-caption polarity categorization. **Table 3.5** depicts the results of our VABDC-Net model, which is based on the visual attention and bi-directional caption processing network.

Table 3.5 Experimental results of our proposed framework VABDC-Net on Twitter-15 and Twitter-17.

Datasets	VABDC-Net (Ours)			
	$\mathcal{A}$	$\mathcal{MP}$	$\mathcal{MR}$	$\mathcal{MF}1$
Twitter-15 (visual + caption)	83.80	83.55	83.78	83.58
Twitter-17 (visual + caption)	72.42	72.38	73.26	72.50

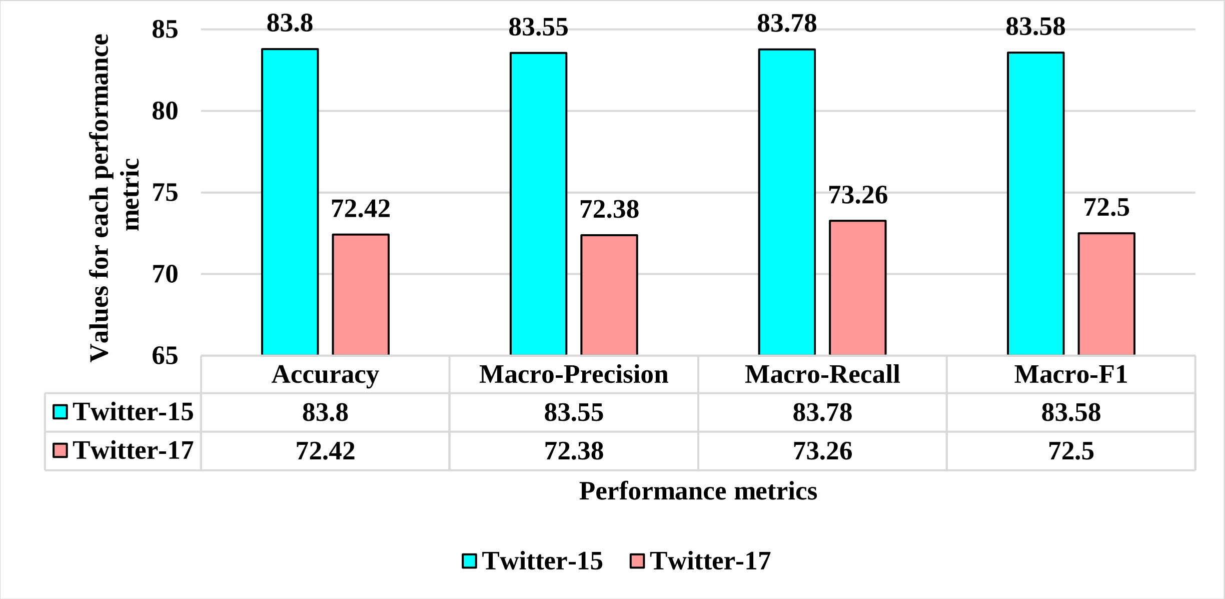
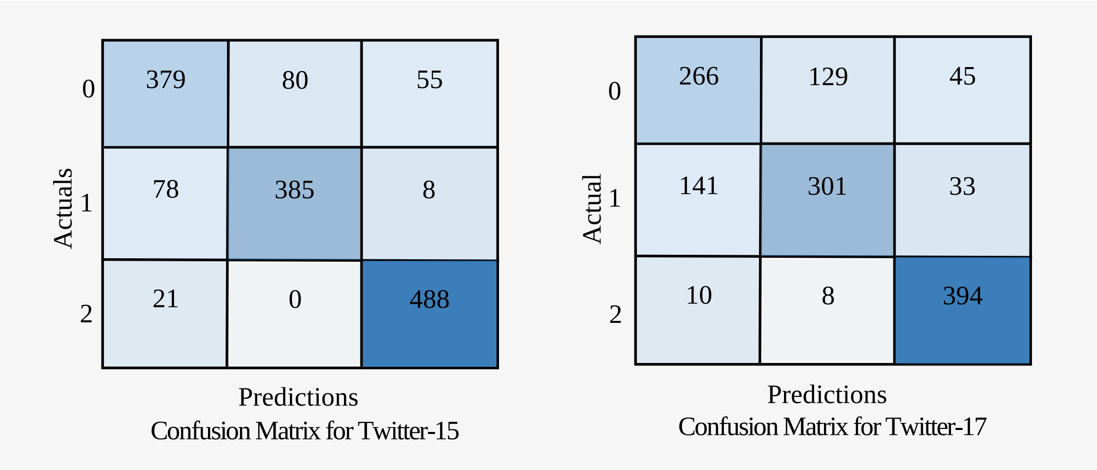


Figure 3.5 Graphical representation of the experimental results on Twitter-15 and Twitter-17 datasets using VABDC-Net.

From the results following observations can be made. First, regarding accuracy and macro-F1-score, VABDC-Net surpasses the other models, implying that it can better learn the variety of the multiple feature modalities and perform more robustly. We find that VABDC-Net outperforms the other models due to the attentional tokenizer-based bi-directional caption and spatio-depth visual attention modules. The functionality of these two modules has been thoroughly detailed in earlier sections. The attention modules primarily increase the interactions of two categories of modality information; mainly, it enables caption information to help in the acquisition of visual characteristics (and vice versa) to attain balanced learning of both types of modality details, which improves visual-caption sentiment recognition. Our model either outperforms or is competitive with the baseline approaches for various evaluation measures. As a result, we believe our method is effective.

To further analyze the effectiveness of our proposed VABDC-Net for each sentiment category, we use the confusion matrix to demonstrate our methods' prediction accuracy on the Twitter-15 and Twitter-17 datasets for three sentiment categories, as shown in **Figure 3.6**. According to **Figure 3.6**, our technique has the best prediction accuracy for the positive category, demonstrating the usefulness of the proposed model for VCSR. The fundamental cause for misclassified samples might be that the visual and caption aspect in both datasets is so similar that they are easily confused in multi-modal sentiment identification.



Actuals	0	1	2
0	379	80	55
1	78	385	8
2	21	0	488

Predictions
Confusion Matrix for Twitter-15

Actual	0	1	2
0	266	129	45
1	141	301	33
2	10	8	394

Predictions
Confusion Matrix for Twitter-17

Figure 3.6 Confusion Matrix for twitter-15 and Twitter-17 datasets where 0, 1 and 2 belongs to negative, neutral and positive class labels.

Table 3.6 Analysis of VABDC-Net prediction on multiple test samples of Twitter-15 dataset where ✓ and ✗ represents right and wrong predictions.

Visual modality	Caption	Ground truth	Prediction by VABDC-Net
	RT @ donnabrazile: Congratulations to @ Oprah and @ GloriaSteinem. T. Presidential Medal of Freedom	**Positive**	**Positive** ✓
	RT @ JordanStrack: T presented the check for winning the 2015 Marathon Classic. Chella Choi	**Positive**	**Positive** ✓
	RT @ NiallOfficial: another shot from T! @ CalAurand Santiago	**Positive**	**Neutral** ✗
	T moved back to Chicago to care for her mom: And it ' s been terrible # NextDayChi 17: Harriette	**Neutral**	**Negative** ✗

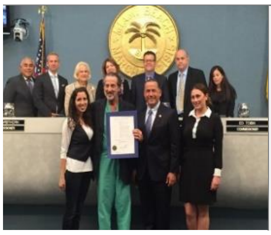
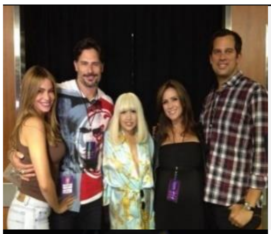
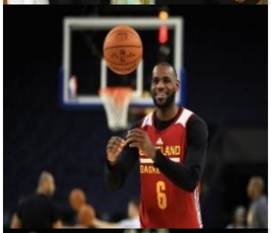
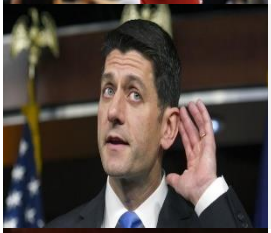
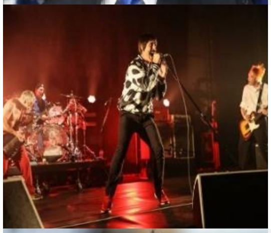
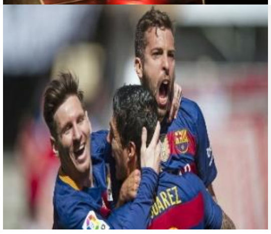
Visual modality	Caption	Ground truth	Prediction by VABDC-Net
	Speaking of the Galbuts, commission just honored memory of matriarch T, who passed away a few weeks ago. Bessie Galbut	Negative	Negative ✓

Table 3.7 Analysis of VABDC-Net prediction on multiple test samples of Twitter-17 dataset where ✓ and ✗ represents right and wrong predictions.

Visual modality	Caption	Ground truth	Prediction by VABDC-Net
	#ThrowBackThursdayBackstage with T and the family. # TBT Lady Gaga	Positive	Negative ✗
	T Finals: LeBron James's Record Has Improved with Age NBA	Positive	Positive ✓
	Just go ahead: 64 % of likely T voters say Paul Ryan should endorse Donald Trump Republican	Neutral	Neutral ✓
	T ' illness could affect Rock on the Range festival Anthony Kiedis	Negative	Negative ✓
	T Crowned 2016 La Liga Champions, Suarez Sink . . . Barcelona	Positive	Neutral ✗

Baselines for Comparison

Our model is compared to the baseline approaches discussed below. **TomBERT** [103]adopts ResNet-152 and BERT to obtain visual and caption features. Then elements of both modalities are combined and passed to the BERT encoder layers to modal the interaction across both modes. Finally, the final hidden state of the "[CLS]" token is used to categorize the sentiment. **ESAFN** [88] utilized LSTM and ResNet architectures to learn the features from text and image-based information. Then a multi-modal fusion layer is applied, followed by a softmax layer for sentiment prediction from multi-modal tweets. **EF-Net** [94] employs Bi-GRU and ResNet-152 to extract visual and textual features and then passes the fused vector of the two modalities to the dense layer for final sentiment categorization. **ModalNet** [104] suggested an attentional fusion network by extracting features with bi-LSTM and ResNet-50 for sentiment prediction. **HIMT** [105] developed a transformer-based model consisting of a BERT module for textual and an F-RCNN module for visual feature extraction. **CapBERT** [106] suggested a framework that first converts existing images in the dataset to captions. The generated and previously available captions are then given to the BERT encoder layers for sentiment prediction. Before integrating the features for prediction, [79] use F-RCNN and BART for feature extraction. **VAuLT** [107] utilizes a pre-trained vision-language transformer for multi-modal sentiment recognition. The performance comparison of VABDC-Net with all the baseline methods is demonstrated in **Table 3.8**. **Figure 3.7** and **Figure 3.8** show a graphical illustration of the year-wise comparison of accuracy and macro-F1 scores for all baseline approaches with our proposed VABDC-Net.

Table 3.8 Comparison of different baseline methods on Twitter-15 and Twitter-17 datasets.

Methods for (visual + caption)	**Twitter-15**				**Twitter-17**			
	$\mathcal{A}$	$\mathcal{MP}$	$\mathcal{MR}$	$\mathcal{MF}1$	$\mathcal{A}$	$\mathcal{MP}$	$\mathcal{MR}$	$\mathcal{MF}1$
TomBERT [103]	77.15	-	-	71.75	70.50	-	-	68.04
ESAFN [88]	73.38	-	-	67.37	67.83	-	-	64.22
EF-Net [94]	73.65	-	-	-	67.77	-	-	-
ModalNet [104]	79.03	-	-	72.50	72.36	-	-	69.19
HIMT [105]	78.14	-	-	73.68	71.14	-	-	69.16
CapBERT [106]	77.92	-	-	73.90	72.30	-	-	70.20
VAuLT [107]	75.60	-	-	70.0	70.20	-	-	67.80
VABDC-Net (Ours)	**83.80**	**83.55**	**83.78**	**83.58**	**72.42**	**72.38**	**73.26**	**72.50**

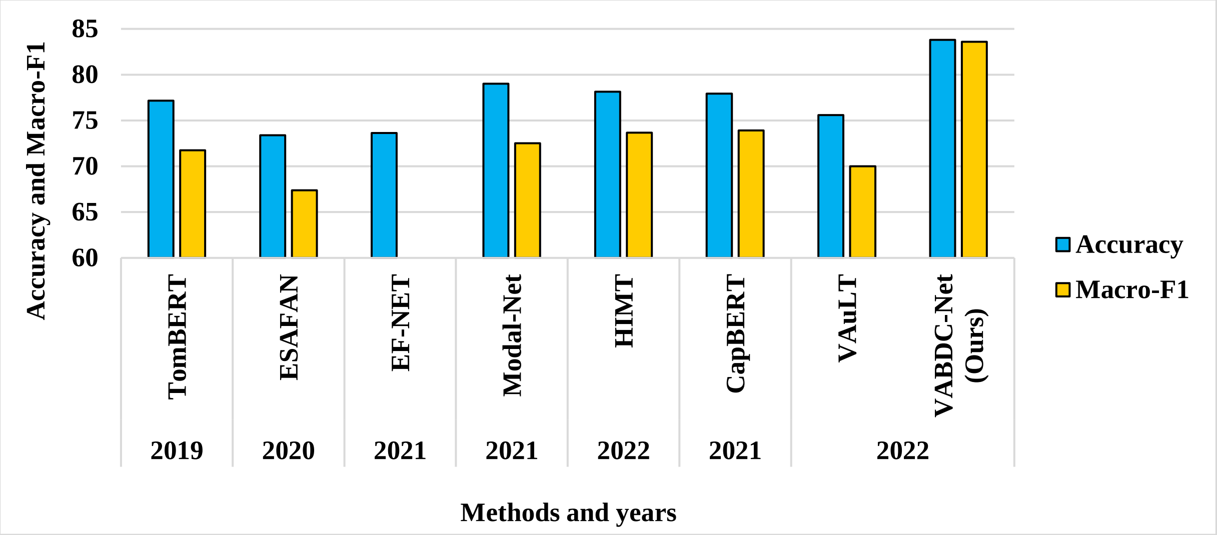


Figure 3.7 Comparison of accuracy and macro-F1 of all the baseline methods and our proposed framework for the Twitter-15 dataset.

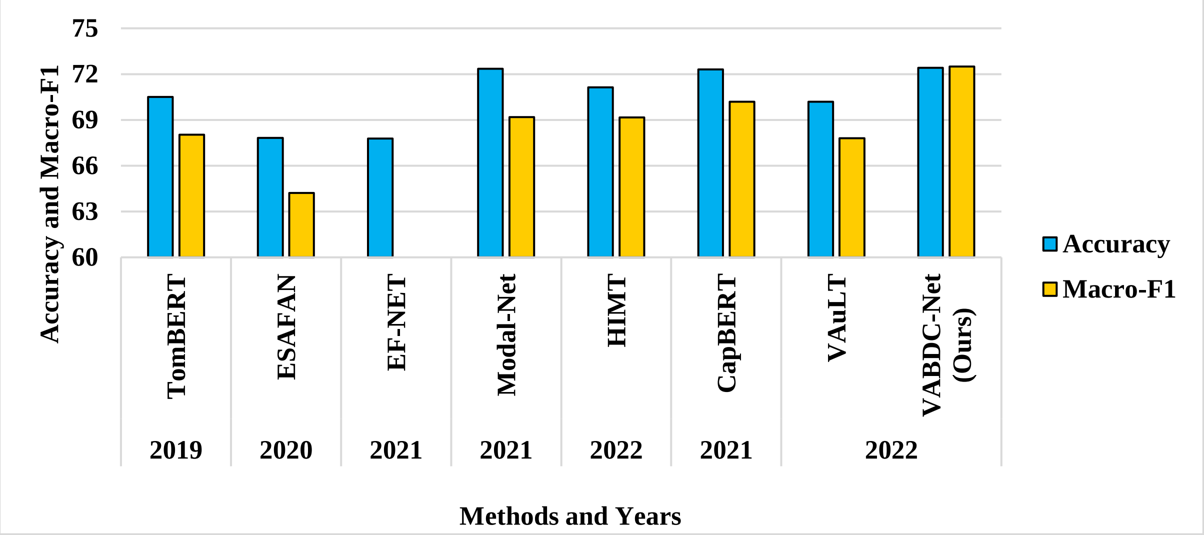


Figure 3.8 Comparison of accuracy and macro-F1 of all the baseline methods and our proposed framework for the Twitter-17 dataset.

On Twitter-15, our approach beats the best baseline model **ModalNet** [104] by 4.77% and 11.08%, respectively, in terms of $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{M\mathcal{F}1}$ scores. When applied to the Twitter-17 dataset, our model significantly achieves the competition with **ModalNet** [104] by 0.06% and 3.31% in terms of $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{M\mathcal{F}1}$ along with various other methods such as **TomBERT** [103], **ESAFN** [88], **EF-Net** [94], **HIMT** [105], and **VAuLT** [107]. Altogether, these findings show the benefit of the proposed VABDC-Net for VCSR. The proposed attentional tokenizer-based bi-direction caption module and visual-attention module capture the interactions between visual and caption modalities at a deeper level. Apart from these benefits, the proposed model is lightly weighted since the total number of trainable parameters in the caption branch is 1658275, and the total number of trainable parameters in the visual branch is 246265.

Ablation Study

In this section, we perform numerous ablation experiments on two Twitter datasets to further test the performance of each suggested module. Based on the VABDC-Net model, we remove the attentional tokenizer from the bi-directional caption branch module and the spatio-depth attention module from the attentional visual branch, denoted as "VABDC-Net w/o attentional tokenizer" and "VABDC-Net w/o spatio-depth attention" in **Table 3.9**. The results of these ablation trials are summarised in **Table 3.9**.

Table 3.9 Ablation study result on Twitter-15 and Twitter-17 datasets.

Methods	**Twitter-15**				**Twitter-17**			
	$\mathcal{A}$	$\mathcal{MP}$	$\mathcal{MR}$	$\mathcal{MF}1$	$\mathcal{A}$	$\mathcal{MP}$	$\mathcal{MR}$	$\mathcal{MF}1$
VABDC-Net w/o attentional tokenizer	79.42	76.23	75.36	78.36	63.28	62.43	62.21	64.13
VABDC-Net w/o spatio-depth attention	80.05	79.93	80.18	80.00	65.10	64.75	65.68	65.11
VABDC-Net	**83.80**	**83.55**	**83.78**	**83.58**	**72.42**	**72.38**	**73.26**	**72.50**

These observations lead us to the following conclusions: (a) the proposed VABDC-Net, which includes all modules, achieves the best performance on both Twitter datasets. (b) The elimination of any one module would result in inferior prediction results. From the above observations, we may conclude that each suggested module is essential and contributes to overall performance.

Generalization Study

Despite exceptional performance on VCSR datasets, most emerging recent VCSR methods primarily rely on thorough in-dataset architectural engineering while ignoring generalization ability, which is essential when algorithms are required to analyze examples from other datasets or domains. As a result, rather than assessing the quality of VABDC-Net exclusively on one dataset, we present a cross-dataset assessment in which a model trained on one dataset is tested on another dataset.

Hence, to evaluate the robustness of VABDC-Net, we perform a generalization study by conducting a cross-dataset analysis in addition to the previously described experiment. For this experimental analysis, we train our proposed approach VABDC-Net on the Twitter-17 dataset, then evaluate it using test samples of the Twitter-15 dataset.

Table 3.10 Cross-dataset analysis to test the robustness of VABDC-Net where Twitter-17 is used for training and Twitter-15 is used for testing.

Datasets	**VABDC-Net (Ours)**			
	$\mathcal{A}$	$\mathcal{MP}$	$\mathcal{MR}$	$\mathcal{MF}1$
Twitter-17 (for training) & Twitter-15 (for testing)	72.22	67.62	78.08	70.33

During the training phase of the cross-dataset analysis, out of a total of 5972 samples from the Twitter-17 dataset, 80% of the samples were randomly utilized for training, and 10% were used for validation. And during the testing phase, out of 5347 image-text pairings from the Twitter-15 dataset, 10% of the samples were picked at random to evaluate the effectiveness of the proposed model based on several parameters such as accuracy ($\mathcal{A}$), macro-precision ($\mathcal{MP}$), macro-recall ($\mathcal{MR}$), and macro-F1($\mathcal{MF}\mathbf{1}$). **Table 3.10** depicts the results of the cross-dataset analysis to test the robustness of VABDC-Net, which is based on visual attention and bi-directional caption processing modules. The results show that our proposed model, VABDC-Net, is accurate and generalizable, implying that our method potentially recognizes instances from datasets other than the ones it was trained on. As a result, we conclude that our technique is more reliable and robust than previous cutting-edge alternatives.

Discussion

The prevalence of social media and digital platforms has encouraged users to express themselves through the use of multimodal content, such as photos and text. Utilizing machine learning algorithms to interpret the psychological orientation inherent in this multimedia information, we can efficiently capture people's views about specific occurrences, necessitating a focus on multimodal sentiment analysis in our research. Our objective differs from the conventional sentiment analysis of determining if a text reflects a positive or negative sentiment; rather, we seek to deduce the user's latent sentiment. We demonstrate that our multimodal model, which combines caption and visual data, outperforms previous cutting-edge models which are solely based on either text or images. We have contributed to the area of VCSR (Visual Caption Sentiment Recognition) by proposing a novel approach which seeks to interpret a picture and text combination by integrating textual (spoken words) and visual modalities (facial expressions). Based on our findings, it is evident that our method VABDC-Net outperforms the other models, indicating that it is better able to learn the variety of different feature modalities and perform more robustly. Due to the attentional tokenizer-based bi-directional caption and spatial-depth visual attention modules, VABDC-Net outperforms the other models. The attention modules essentially improve the interactions between two categories of modality information; particularly, it enables caption information to aid in the acquisition of visual features (and vice versa) to achieve balanced learning of both sorts of modalities.

3.2.4 Conclusion

This paper proposes a novel Visual Attention and Bi-Directional Caption Processing network (VABDC-Net) for visual-caption sentiment recognition. Our proposed methodology more effectively analyses the interaction between visual and captions modalities than earlier innovative methods. Depending on the context, the same phrase may generate different sentiment in several scenarios. Hence, it is essential to use both visual and textual content for more accurate prediction. Motivated by this, we developed a novel visual attention branch for extracting relevant information from the images and a bi-directional caption processing network for extracting crucial features from the caption modality. The extensive experiments on two publicly accessible datasets revealed that our proposed model beats highly competitive baseline models. Although promising results have been obtained, there are still many avenues open for further study. These include incorporating adversarial learning into the multi-modal feature fusion module, improving feature extraction techniques, expanding our model to incorporate aspect-based multi-modal sentiment recognition, and exploring a wider variety of multimedia data, including video and audio. Furthermore, considering that most existing multi-modal approaches concentrate on sentiment analysis, we intend to examine multi-modal continuous emotion intensity in future research, which might offer a deeper semantic relationship.

3.3 Significant Outcomes of this Chapter

The significant outcomes of this chapter are as follows:

- To predict sentiment by employing visual-caption pairs using a framework named as “VABDC-Net (Visual Attention and Bi-Directional Caption Processing Network)”. The proposed model comprised of three modules: an attentional tokenizer-based bidirectional caption expert branch to retrieve useful textual features, an attention visual expert branch to retrieve appropriate visual features, and a cross-modal feature fusion module to merge features and predict sentiment.

- Performed an ablation and generalization study to evaluate the resilience of the proposed model.

The following research studies serve as the foundation for this chapter:

- **A. Pandey** and D. Kumar Vishwakarma, "VABDC-Net: A framework for Visual-Caption Sentiment Recognition via spatio-depth visual attention and bi-directional caption processing," ***Knowledge-Based Systems***, vol. 269, June. 2023, doi: https://doi.org/10.1016/j.knosys.2023.110515.
- **A. Pandey** and D. K. Vishwakarma, "Attention-based Model for Multi-modal sentiment recognition using Text-Image Pairs," in *2023* ***4th International Conference on Innovative Trends in Information Technology (ICITIIT)***, Institute of Electrical and Electronics Engineers (IEEE), March. 2023, pp. 1–5. doi: 10.1109/ICITIIT57246.2023.10068626.

Chapter 4: Target-Dependent Multimodal Sentiment Recognition

4.1 Scope of this Chapter

Target-Dependent Sentiment Recognition (TDSR) determines the sentiment polarity towards specific attributes that are explicitly mentioned within a given input text. For example, “*The river gives a pleasant view, however, the quality of the roll was disappointing*”. The user conveys a positive opinion regarding the *river′s* location while expressing dissatisfaction or negative sentiment for the target aspect *′roll′*. With the growing popularity of multimodal information across social media, it has become inadequate to rely just on written text for the purpose of aspect-based sentiment categorisation. Multimodal postings often include visual elements, such as photos and emoji, which may frequently provide a significant understanding of people's emotions. The opinion towards a target aspect, expressed by a user, is often influenced by the accompanying image. This is because the textual information in such posts is typically unstructured, informal and brief. Hence, motivated by the recent advancements [108], [109], [110], and [111] made in understanding facial emotion within the domain of computer vision, we provide an effective and simple approach, i.e., visual-to-emotional-caption translation network (VECT-Net) for target-dependent multimodal sentiment recognition. This technique aims to convert the sentiment conveyed in the image into a textual representation by generating descriptions of facial expressions.

4.2 Target-Dependent Multimodal Sentiment Analysis Via Employing Visual-to-Emotional-Caption Translation Network using Visual-Caption Pairs

4.2.1 Abstract

Target-dependent sentiment recognition is a highly intriguing and significant domain within affective computing. Substantial advancements have been achieved in this field, with notable contributions such as those presented in [112]. The natural language processing and multimedia field has seen a notable surge in interest in multimodal sentiment recognition. Hence, this study aims to employ Target-Dependent Multimodal Sentiment Recognition (TDMSR) to identify the level of sentiment associated with every target (aspect) stated within a multimodal post consisting of a visual-caption pair. Despite the recent advancements in multimodal sentiment recognition, there has been a lack of explicit incorporation of emotional clues from the visual

modality, specifically those pertaining to facial expressions. The challenge at hand is to proficiently obtain visual and emotional clues and subsequently synchronise them with the textual content. In light of this fact, this study presents a novel approach called the Visual-to-Emotional-Caption Translation Network (VECT-Net) technique. The primary objective of this strategy is to effectively acquire visual sentiment clues by analysing facial expressions. Additionally, it effectively aligns and blends the obtained emotional clues with the target attribute of the caption mode. The experimental findings demonstrate that our methodology is capable of producing ground-breaking outcomes when applied to two publicly accessible multimodal Twitter datasets, namely, Twitter-2015 and Twitter-2017. The experimental results show that the suggested model achieves an accuracy of **81.23%** and a macro-F1 of **80.61%** on the Twitter-15 dataset, while **77.42%** and **75.19%** on the Twitter-17 dataset, respectively. The observed improvement in performance reveals that our model is better than others when it comes to collecting target-level sentiment in multimodal data using the expressions of the face.

4.2.2 Proposed Methodology

In this section, the proposed framework of VECT-Net is thoroughly discussed. The problem is defined in the first portion of this section. Then, we present the model's framework, comprised of four distinct components: Facial emotion description module, Target alignment and refinement of the face descriptions module, Image captioning module, and Fusion module.

Problem Formulation

The TDMSR can be precisely described as outlined below: Consider a collection of visual-caption pair examples denoted as $M = \{E_1,\ E_2,\ E_3, \dots\dots\dots, E_M\}$, where |M| represents the total number of instances. For each given example, an image $I \in\ R^{3\times H\times W}$ is provided where 3, $\mathcal{H}$ and $\mathcal{W}$ indicate the number of channels, height and width. Every visual sample in this study is associated with textual content represented by a set of K- words provided by the captions $C = \{w_1, w_2, w_3, \dots\dots\dots, w_K\}$, which comprises a subsequence of N-word that represents the target entity, defined as $T = \{w_1, w_2, w_3, \dots\dots\dots, w_N\}$. Our study aims to develop a sentiment classifier to predict a sentiment label Y from multimodal examples accurately. A combination of variables $E = \{I, C, T\}$ represents each sample in E. The sentiment labels are categorised into three classes: $Y \in \{positive, negative, neutral\}$. Consider **Figure 4.1** as an example. When omitting the accompanying image, the anticipated sentiments towards the targets "Justin," "America," and "Lydia" seem to be neutral, positive, and neutral, respectively. However, this prediction is inaccurate. In the instances mentioned above (**Figure 4.1**), users convey their sentiments towards "Justin," "America," and "Lydia" by utilising distinct visual representations. Specifically, a crying face is employed to express negative sentiment towards Justin, a neutral image is used to convey a neutral feeling towards America, and a happy look is employed to denote a positive view towards Lydia.

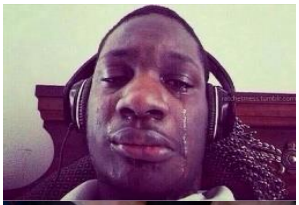
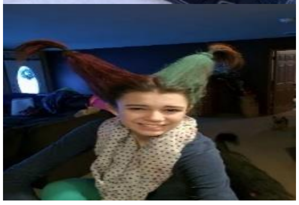
Visual	Textual	Target	Sentiment
	RT @ irauhlcarlyrae : Justin tweeted # mybeliebers so Beliebers trended # ourJustin	Justin	**Negative**
	Randy from America knows how to wear a good sock. Hide it under your boot for extra warmth. # tweetusyoursocks	America	**Neutral**
	Crazy hair day! Lydia is a contender. :)	Lydia	**Positive**

Figure 4.1 A few instances of Target-Dependent multimodal sentiment recognition are provided, including the identified targets and their corresponding sentiments.

Visual-to-Emotional-Caption Translation Network (VECT-Net)

The proposed framework Visual-to-Emotional-Caption-Translation Network (VECT-Net) illustrated in **Figure 4.2** has three distinct components: Facial emotion description module, Target alignment and refinement of the face descriptions, and Fusion module. For a given tweet consisting of a visual-caption pair, denoted as $E = \{I, C, T\}$, we first take the input image $'I'$ and feed it into a facial emotion description unit to generate face description $D = \{D_1, D_2, D_3, \dots \dots \dots, D_{\mathcal{F}}\}$ comprises of different features such as age, gender, emotion, etc., where $\mathcal{F}$ is the number of faces present in an input image and $D_i = \{D_1, D_2, D_3, \dots \dots \dots, D_L\}$ represents a phrase consisting of L-word. The extraction and textualisation of facial expressions within an image, which represents an immense amount of information on the individual's sentiments, is the primary emphasis of this module. Since the input image $'I'$ may include several facial expressions, it is necessary to match or align the facial description D_T with the target entity T. The target alignment and refinement of the face descriptions module estimates cosine similarity between visual input I and face descriptions with target D_T. The facial description D_T is selected and rewritten based on similarity scores. Since the visual scenes may provide extra semantic details, we employ the image-to-text transformer (*Wang et al.* [113], 2022) to produce image captions for the scene $I_C = \{I_{C1}, I_{C2}, I_{C1}, \dots\dots\dots, I_{CG}\}$, where G represents caption length. At last, in the fusion component, we employ two robustly optimised pre-trained language models based on BERT to simulate image captions and face descriptions, followed by a gating mechanism for feature fusion and noise reduction. For target-dependent sentiment recognition, the gated unit output flows via a linear layer. The pseudocode for the proposed algorithm is presented in **Table 4.2**. The following subsections will provide detailed insights for every module. **Table 4.1** presents a comprehensive collection of significant symbols along with their respective meanings, aiming for a better understanding of the proposed strategy.

Table 4.1 Listed below are some important symbols and their respective meanings.

Symbols	***Meaning***
E	Multimodal example
I	Visual modality available in the dataset
C	Caption modality available in the dataset
T	Target entity available in the dataset
Y	Output sentiment labels
$\mathcal{F}$	Total number of faces extracted from the visual modality
D	The set represents the face description obtained from $\mathcal{F}$.
D_T	Refined face descriptions concatenated with the target entity
I_C	Set of captions generated by using visual modality I of the dataset
$O_{D\&T}$ and O_I	Embedding of visual and textual content
V and b	Trainable parameters (Weights & Bias)
$\oplus$	Denotes concatenation
$\odot$	Denotes gated fusion

Table 4.2 Pseudocode for the proposed algorithm.

Algorithm 1: Target-Dependent Sentiment Recognition based on visual-to-emotional-caption translation network (VECT-Net)

Aim: To learn a mapping function $\boldsymbol{F}: (\boldsymbol{I}, \boldsymbol{C}, \boldsymbol{T}, \boldsymbol{Y}) \longrightarrow$ from the multi-modal training examples $\boldsymbol{E}$.

Input: visual set $\boldsymbol{I} \in \boldsymbol{\mathcal{R}}^{3\times\mathcal{H}\times\mathcal{W}}$;

Caption set $\boldsymbol{C} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \dots\dots\dots, \boldsymbol{w}_K\}$ where $\boldsymbol{K} \in$ set of words provided by a caption corresponding to a visual sample;

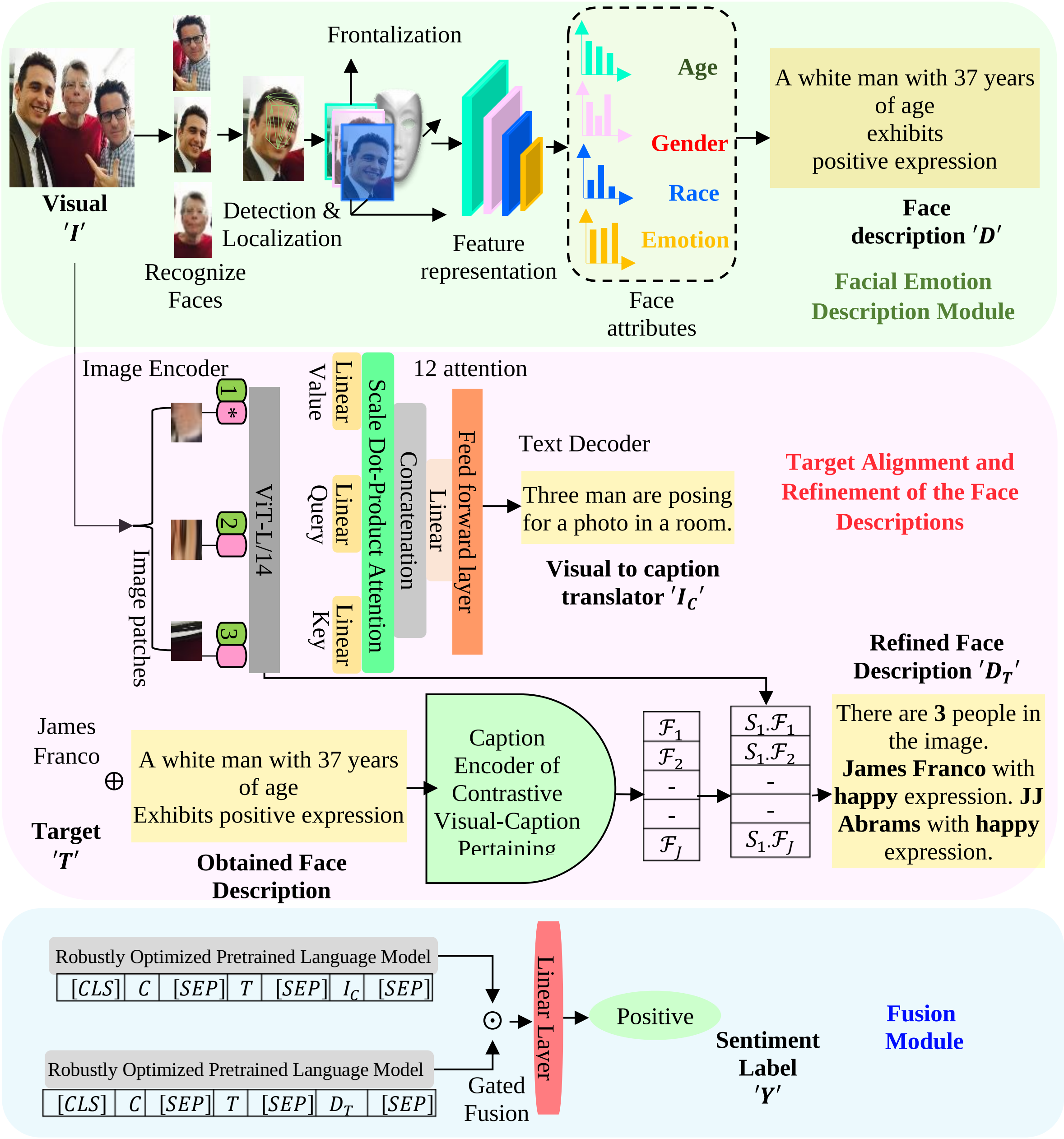
Target entity set $\boldsymbol{T} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \dots\dots\dots, \boldsymbol{w}_N\}$ where caption with a subsequence of $\boldsymbol{N}$-word represents the target entity and;

Output: Categorization of sentiment label $\boldsymbol{Y} \in \{\boldsymbol{positive}, \boldsymbol{negative}, \boldsymbol{neutral}\}$ based on target $\boldsymbol{T}$

```
1.  for $\boldsymbol{E} \leftarrow \boldsymbol{1}$ to Epochs do
      $\boldsymbol{\mathcal{F}} = \{\boldsymbol{\mathcal{F}}_1, \boldsymbol{\mathcal{F}}_2, \boldsymbol{\mathcal{F}}_3, \dots\dots\dots, \boldsymbol{\mathcal{F}}_J\} \leftarrow \boldsymbol{I}$ extraction of faces from the visual set using Eqn. (4.1);
      $\boldsymbol{D} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3, \dots\dots\dots, \boldsymbol{D}_{\boldsymbol{\mathcal{F}}}\} \leftarrow \boldsymbol{\mathcal{F}}$ generation of fluent linguistic emotional face descriptions using extracted faces obtained from the previous step;
      $\boldsymbol{D}_{\boldsymbol{T}} \leftarrow \boldsymbol{D}$ refined face descriptions are obtained concatenated with the target entity using Eqn. (4.2) to Eqn. (4.6);
```

Algorithm 1: Target-Dependent Sentiment Recognition based on visual-to-emotional-caption translation network (VECT-Net)

```
$I_C = \{I_{C1}, I_{C2}, I_{C1}, \dots \dots \dots, I_{CG}\} \leftarrow I$ generate captions from the visual modality using Eqn. (4.7);
$[CLS] w_1^C, \dots, w_K^C [SEP] w_1^T, \dots, w_N^T [SEP] w_1^{D_T}, \dots, w_L^{D_T} [SEP]$ fine-tune the combination of available text, target and refined facial descriptions using robustly optimised language model using Eqn. (4.8);
$[CLS] w_1^C, \dots, w_K^C [SEP] w_1^T, \dots, w_N^T [SEP] w_1^{I_C}, \dots, w_G^{I_C} [SEP]$ fine-tune the combination of available text, target and generated captions using robustly optimised language model using Eqn. (4.9);
Eqn. (4.8) $\odot$ Eqn. (4.9) gated fusion of the result obtained from the previous two steps;
$Y \in \{positive, negative, neutral\}$ by passing Eqn. (4.8) $\odot$ Eqn. (4.9) by a fully connected layer and softmax layer;
Calculate the loss $\mathbb{L} = -\frac{1}{|D|} \sum_{\ell=0}^{|D|} \log \mathcal{P}\{Y^\ell | O^\ell\}$ and perform backpropagation; Eqn. (4.10)
end
```



A white man with 37 years of age exhibits positive expression

Three man are posing for a photo in a room.

A white man with 37 years of age Exhibits positive expression

There are **3** people in the image. **James Franco** with **happy** expression. **JJ Abrams** with **happy** expression.

Figure 4.2 Proposed Methodology ('⊕' denotes concatenation and '⊙' denotes gated fusion).

Facial Emotion Description Module

The proposed module addresses two fundamental difficulties in TDMSR. First, the complex images in multimodal tweets might make it challenging to extract object-level emotional indicators. Another issue is translating emotional signals obtained from visual modality into a sequence of words.

To address the first challenge, as previously stated, leveraging the wide range of facial expressions in images proves to be an efficient method for extracting emotional cues from visual mode. The first step involves using a tool represented as $\oint$ developed by Serengil *et al.*[114] to recognise multiple faces within an image of the dataset as stated in **Eqn.** $(\mathbf{4.1})$. Let $\mathcal{F}$ represent the set of faces, denoted as $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \dots \dots \dots, \mathcal{F}_J\}$, where J represents the total number of faces and $\mathcal{F}_J \in \mathcal{R}^{3 \times \mathcal{H}_\mathcal{F} \times \mathcal{W}_\mathcal{F}}$ represents a face area with $three$ channels, $\mathcal{H}_\mathcal{F}$ height, and $\mathcal{W}_\mathcal{F}$ width. The obtained faces $\mathcal{F}$ are then fed into a pre-trained classification model (Serengil *et al.* [115]) for facial attribute analysis, which involves gender, age, race (Indian, Black, Asian, White, Latino, and Middle Eastern,) and facial expression to predict sentiments.

$$\boldsymbol{\mathcal{F}} = \oint(\boldsymbol{I}) \tag{4.1}$$

For the second challenge, facial attributes obtained from the previous step are transformed into the textual representation without training an additional visual-caption model. The prediction confidence score α is used to pick out facial attributes. The face attributes that have a score below the threshold $\alpha = 0.5$ are filtered out. We manually develop a visual feature pattern to create fluent linguistic emotional face descriptions $D = \{D_1, D_2, D_3, \dots \dots \dots, D_\mathcal{F}\}$. An example of facial description generation is shown in **Table 4.3**.

Table 4.3 A fluent face description, for example, 1 of Figure 21, may be generated via a template.

Template	**Instances**	**Visual features**				**Fluent linguistic emotional face descriptions**
A [Race] [Gender] with [Age]-year-of-age exhibits [Sentiment] expression		Age	Gender	Race	Sentiment	A Black man with 43 years of age exhibits a negative expression
	Example 1	43	man	Black	negative	
	Prediction confidence score	1.00	1.00	0.879	0.9467	

Target Alignment and Refinement of the Face Descriptions

Sometimes, a multi-face image sample with varied facial expressions fails to estimate the correct emotion of the target entity. On the other hand, the inclusion of redundant facial emotions produces noise and diminishes the overall effectiveness. Therefore, it is essential to effectively synchronise the facial emotions shown in the visual sample with the desired target entity. In our proposed framework, VECT-Net, this component focuses on aligning facial expressions with the target object, resulting in more detailed facial descriptions. The TDMSR challenge needs external visual-caption alignment information for more fine-grained alignments due to restricted dataset size and the absence of direct visual-caption alignment supervision. Hence, to achieve fine-grained alignment, we use caption and visual encoders of a recently developed contrastive visual-caption pre-training architecture trained on a variety of visual-caption pairs [116] denoted as '$\boldsymbol{\tau}$' to encode the face descriptors D associated with target T and the visual I. The resulting embeddings for images and descriptions of faces are shown in **Eqn. (4.2)** and **Eqn. (4.3),** where '$\bigoplus$' denotes concatenation.

$$\boldsymbol{O_{D_T} = Caption\_Encoder\ (\tau)(D \oplus T)} \tag{4.2}$$

$$\boldsymbol{O_I = Visual\_Encoder(\tau)(I)} \tag{4.3}$$

Subsequently, the obtained feature embeddings are projected into the same feature space. Then, we compute the Levenshtein distance L for these feature embeddings using L2-normalization $'\gamma'$. Next, we choose and regenerate the face description that best fits the current image as the visual, emotional clue for the current target based on the similarity score using L. The refined face description only contains the target object and expressions based on predicted facial traits.

$$\boldsymbol{O'_{D_T} = \gamma(O_{D\&T} \cdot V_{D\&T})} \tag{4.4}$$

$$\boldsymbol{O'_I = \gamma(O_I \cdot V_I)} \tag{4.5}$$

$$\boldsymbol{L = \ O'_I \cdot ({O'_{D_T}}^{\mathbb{T}}) * e^t} \tag{4.6}$$

$V_{D\&T}$ and V_I are trainable weights, and t is the scaling factor of the generative visual-to-caption transformer model [113]. This module is only used for multi-face visual samples. The target is concatenated directly with the acquired face description in this module. Subsequently, the newly concatenated phrases are used as input for the textual encoder of '$\boldsymbol{\tau}$', while the picture is employed as input for the visual encoder of '$\boldsymbol{\tau}$'. The cosine similarities between the visual and textual features are then computed using **Eqn. (4.5)**, **(4.6)** and **(4.7)**. Then, we choose the facial description with the highest score and modify it to get a more refined description.

Furthermore, considering the effect of image scene details on multimodal semantics, we use a recently developed, more effective generative transformer [113] for visual-to-caption translation '$\boldsymbol{\delta}$' to provide an overall comprehensive description of all the image samples of the dataset using **Eqn. (4.8)**.

$$\boldsymbol{I_C = \delta(I)} \qquad (\mathbf{4.8})$$

Ultimately, we achieve the accurate alignment of face descriptions and visual captions, which are then used as input for the succeeding module.

Fusion Module

This module aims to combine already available caption(C), target entity (T), refined facial description (D_T), and the generated caption (I_C). To leverage the pre-trained language model's robust textual context analysis, we concatenate the refined face descriptions and image caption with available text and target to create two new phrases as shown in **Eqn. (4.9)** and **(4.10)** below:

$$\boldsymbol{[CLS]w_1^C, \ldots, w_K^C[SEP]w_1^T, \ldots, w_N^T[SEP]w_1^{D_T}, \ldots, w_L^{D_T}[SEP]} \qquad (\mathbf{4.9})$$

$$\boldsymbol{[CLS]w_1^C, \ldots, w_K^C[SEP]w_1^T, \ldots, w_N^T[SEP]w_1^{I_C}, \ldots, w_G^{I_C}[SEP]} \qquad (\mathbf{4.10})$$

Fine-tuning two robustly optimised per-trained language models [117] with these new phrases yields $[CLS]$ token $O_{D_T}^{[CLS]} \in R^{768}$ and $O_{I_C}^{[CLS]} \in R^{768}$ pooler outputs. The gate mechanism is used to reduce noise in feature representations of $O_{D_T}^{[CLS]}$ and $O_{I_C}^{[CLS]}$. At last, to predict sentiment, fused feature representations (**Eqn. (4.11)** and **(4.12)**) are sent via a linear classifier using **Eqn. (4.13)**, where V_{D_T}, V_{I_C} and V are trainable weights of dimensions $R^{768\times768}$, $R^{768\times768}$, and $R^{768\times3}$. In contrast, b_j and b are learnable biases with dimensions R^{768} and R^3.

$$\boldsymbol{jt = tanh\left(V_{D_T}O_{D_T}^{[CLS]} + V_{I_C}O_{I_C}^{[CLS]} + b_j\right)} \qquad (\mathbf{4.11})$$

$$\boldsymbol{O = jt * O_{D_T}^{[CLS]} + jt * O_{I_C}^{[CLS]}} \qquad (\mathbf{4.12})$$

$$\boldsymbol{\mathcal{P}(Y|O) = Softmax\big((V * O) + b\big)} \qquad (\mathbf{4.13})$$

All module parameters are optimised using conventional cross-entropy loss defined in **Eqn. (4.14)**.

$$\mathbb{L} = -\frac{\mathbf{1}}{|\boldsymbol{D}|}\sum_{\boldsymbol{\ell=0}}^{|\boldsymbol{D}|} \mathbf{log}\,\boldsymbol{\mathcal{P}\{Y^{\ell}|O^{\ell}\}} \qquad (\mathbf{4.14})$$

4.2.3 Experimental Setup and Results

This section includes comprehensive details about the dataset used in the study, the experimental configuration of the proposed framework, and the evaluations of its performance.

Experimental Details and Dataset Description

Our model was trained and evaluated using two publically accessible benchmark datasets, Twitter-2015 and Twitter-2017. Both datasets include tweets consisting of visual-caption pairs. Each caption has been tagged with some target entity and its associated sentiment polarity. Our approach primarily emphasises cases that include facial images. Therefore, we extract samples with facial images from the aforementioned two datasets to create the Tweet1517-Face dataset. Subsequently, we evaluate the effectiveness of our proposed model on these samples to prove its superiority over the others. The comprehensive statistical information for the three datasets can be seen in **Table 4.4**.

Additionally, the model's learning rate has been configured to be 2e-5. The batch size has been set to 32, and a dropout rate of 0.4 has been employed. The proposed approach has undergone fine-tuning for a total of 15 epochs. This work was implemented using the PyTorch framework and is executed on a high-end NVIDIA TITAN RTX (48GB) GPU system with an Intel Xeon Silver 4116 CPU, 10TB storage, and 128 GB RAM. The final result has been determined by calculating the mean of five independent training iterations.

Table 4.4 Statistical information that describes all the datasets utilised in the evaluation of our proposed model.

Name of the Dataset	Positive Samples			Negative Samples			Neutral Samples			Average Number of Targets		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
Twitter-15	928	303	317	368	149	113	1883	679	607	1.34	1.33	1.35
Twitter-17	1508	515	493	1638	517	573	416	144	168	1.41	1.43	1.45
Tweet1517-Face	1285	449	442	408	137	156	1531	514	494	1.37	1.37	1.39

Baselines for Comparison

This section compares our proposed model with several cutting-edge baseline approaches for the task of TDMSR for both unimodal and multimodal networks. **Table 4.5** demonstrates that the experimental results of our proposed model are more accurate than those of the other baseline methodologies in terms of $\mathrm{Accuracy}\ (\mathrm{A})$ and $\mathrm{macro} - \mathrm{F1}$ score, thus proving our model's superiority.

Table 4.5 Experimental results of our proposed model compared with the multimodal baseline approaches.

Methods (visual + captions)	Twitter-2015		Twitter-2017	
	A	$\mathbf{macro - F1}$	A	$\mathbf{macro - F1}$
TomBERT [118]	76.18	71.27	70.50	68.04
ESAFN [119]	73.38	67.37	67.83	64.22
EF-Net [120]	73.65	67.90	67.77	65.32
ModelNet [121]	79.03	72.50	72.36	69.19
R-GCN [122]	-	75.00	-	87.11
HIMT [123]	78.14	73.68	71.14	69.16

Methods	Twitter-2015		Twitter-2017	
(visual + captions)	A	macro − F1	A	macro − F1
FGSN [124]	74.61	65.84	-	-
[6]	-	72.97	-	71.76
EF-CaTrBERT [125]	77.92	73.90	72.30	70.20
Our Proposed (VECT-Net)	**81.23**	**80.61**	**77.42**	**75.19**

Figure 4.3 gives a year-by-year visual representation to compare our model with the techniques that are considered to be state-of-the-art in terms of Accuracy (A) and macro − F1 score for the Twitter-2015 and Twitter-2017 datasets, respectively.

Analysis of Our Experimental Results

This section presents experimental results and analysis of our proposed framework for all three datasets. **Table 4.6**, **Table 4.7** and **Table 4.8** highlight the best scores on each performance measure for all three datasets. Our method demonstrates superior performance when compared to all other multimodal baselines. This serves as evidence of the efficacy of the proposed VECT-Net framework. In the fusion module, we conducted fine-tuning using RoBERTa-base and RoBERTa-Large language models [117] and found that RoBERTa-Large yielded superior results. The model we developed indicates higher accuracy when using a more robust language model. This observation highlights the significant impact of the language model's contextual modelling capacity during the fusion step. However, this observation is not evident in the comparison of the base version model. We believe the inadequate text context modelling by the pre-trained language model in the base version is the cause of this issue.

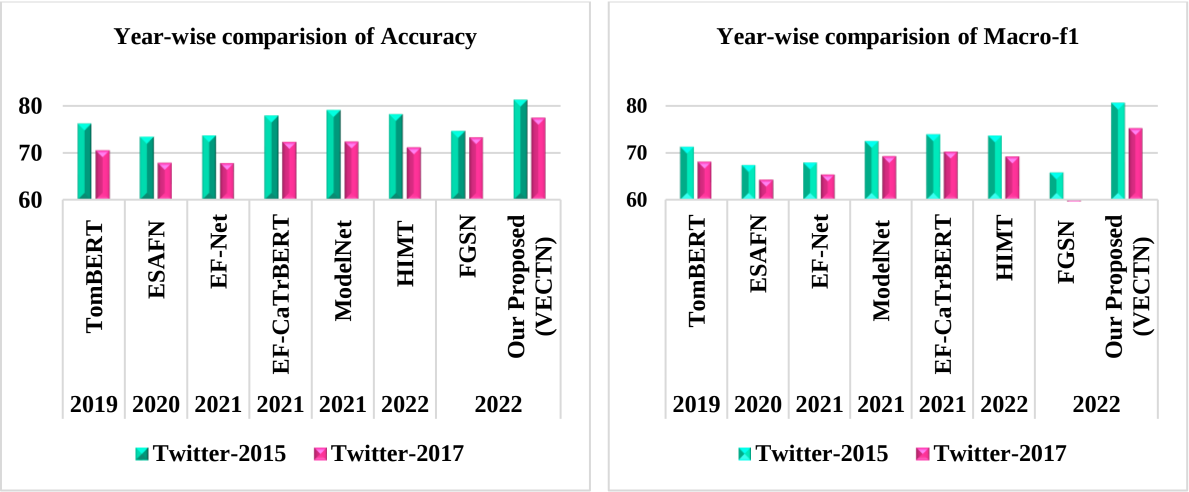


Figure 4.3 Year-wise comparison of our proposed model with already existing cutting-edge multimodal networks in terms of accuracy and macro-f1 scores.

Table 4.6 Experimental results of our proposed approach for Twitter-2015.

Methods	Twitter-2015					
	Unimodal				Multimodal	
	Caption		visual		Caption + Visual	
	A	macro − F1	A	macro − F1	A	macro − F1
VECT-Net-Roberta-base	73.19	71.74	75.20	72.51	79.63	77.21
VECT-Net-Roberta-Large	75.83	75.20	77.59	76.86	**81.23**	**80.61**

Table 4.7 Experimental results of our proposed approach for Twitter-2017.

Methods	Twitter-2017					
	Unimodal				Multimodal	
	Caption		visual		Caption + Visual	
	A	macro − F1	A	macro − F1	A	macro − F1
VECT-Net-Roberta-base	67.42	64.35	68.54	67.40	72.37	70.84
VECT-Net-Roberta-Large	72.11	70.65	74.48	71.79	**77.42**	**75.19**

Figure 4.4 shows that combining visual and textual modes enhances performance. Therefore, this demonstrates that our proposed framework has the capability to accurately represent facial expressions shown in images. Additionally, it emphasises the need to explicitly integrate emotional cues in visual analysis.

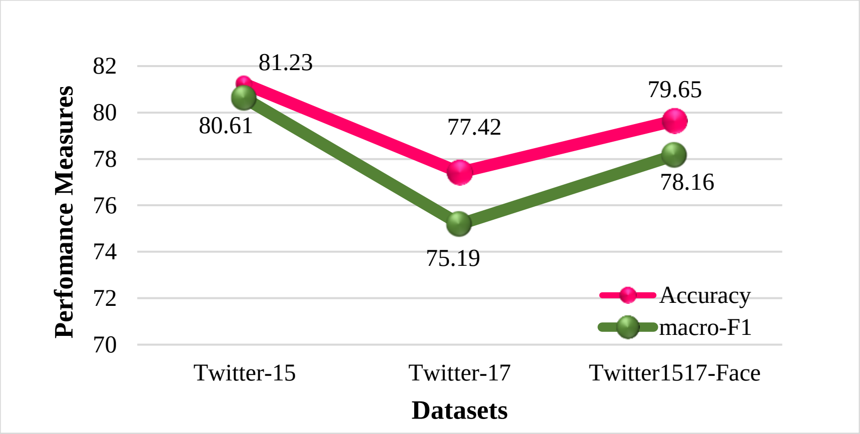


Figure 4.4 Graphical representation to show the performance of the proposed model on all three datasets for multimodal (Image + Caption) configuration.

Table 4.8 Experimental results of our proposed approach for Tweet1517-Face.

Methods	Tweet1517-Face Dataset					
	Unimodal				Multimodal	
	Caption		visual		Caption + Visual	
	A	macro − F1	A	macro − F1	A	macro − F1
VECT-Net-Roberta-base	64.32	62.67	65.22	63.80	74.02	71.28
VECT-Net-Roberta-Large	70.15	67.46	74.57	72.30	**79.65**	**78.16**

Ablation Study

To understand the influence of each component of our technique, we performed an extensive ablation study utilising the VECT-Net-RoBERTa-Large version. The results for the same are depicted in **Table 4.9**. Initially, the sentiment label of the target is predicted by combining the results of linguistic models. These outcomes are then used as input to a linear classification layer without the inclusion of a gating mechanism. The overall performance of the architecture experiences a significant decrease as a result of the presence of noise during the facial emotion description module. On Twitter 2015, A and $\mathrm{macro}-\mathrm{F1}$ scores dropped 1.83% and 1.96%, respectively. On the Twitter-2017 dataset, A drops 2.15%, and the $\mathrm{macro}-\mathrm{F1}$ score drops 2.88%. This suggests that the gating technique is responsible for reducing noise and extracting more useful features. Additionally, it is seen from **Table 4.9** that the exclusion of the target alignment module also results in a decrease in performance. This result suggests that the alignment between the visual and emotional cues and the target entity is crucial. At last, we analyse the impact of excluding the visual caption from the scene, which results in a significant decrease in the model's performance. This finding provides evidence that the utilisation of visual-to-caption translation contributes to the advancement of visual-caption fusion.

Table 4.9 Ablation study conducted on Twitter-2015, Twitter-2017 and Tweet1517-Face.

Methods	**Twitter-2015**		**Twitter-2017**		**Tweet1517-Face**	
	Multimodal (Caption + visual)		Multimodal (Caption + visual)		Multimodal (Caption + visual)	
	$\mathbf{A}$	$\mathbf{macro-F1}$	$\mathbf{A}$	$\mathbf{macro-F1}$	$\mathbf{A}$	$\mathbf{macro-F1}$
VECT-Net-Roberta-Large	**81.23**	**80.61**	**77.42**	**75.19**	**79.65**	**78.16**
VECT-Net without gating mechanism	79.40	78.65	75.27	72.31	77.29	76.08
VECT-Net without Target Alignment	79.11	79.20	76.44	73.60	78.31	76.82
VECT-Net without visual caption of the scene	78.89	78.50	75.23	72.56	77.48	75.27

Predictive Analysis of Few Samples

To provide a more comprehensive demonstration to highlight the benefits of the proposed method, this section of the manuscript will include the actual predictions made by our model on a few samples gathered from Twitter-15 and Twitter-17. As shown in **Table 4.10**, the VECT-Net model accurately forecasts positive sentiments for the target terms $[LeBron\ James's]$, $[JordanStrack]$, negative sentiments for aspect words $[Harriette]$, $[Anthony\ Kiedis]$, while the neutral sense of emotion for $[Donald\ Trump\ Republcian]$. As a result, this study demonstrates that the proposed approach efficiently focuses on multimodal sentimental regions to leverage the interaction between the image and the target phrase more extensively than existing methods. Hence, the VECT-Net model can deeply examine the local semantic relationship between image and text in contrast with baseline models. In simple terms, the proposed model exhibits a higher degree of advantage.

Table 4.10 Sentiment prediction by employing the proposed model "VECT-Net" on a few multimodal samples of Twitter-15 and Twitter-17.

Modalities		**VECT-Net Prediction**
Text	**Image**	
Finals: [***LeBron James's***]$_{Positive}$ ***Record Has Improved with Age NBA***		*LeBron Jame's* – *Positive*
[***JordanStrack***]$_{Positive}$: ***presented the che for winning the* 2015 *Marathon Classic. Chella Choi***		*JordanStrack* – *Positive*
Just go ahead*: 64 % *of likely* [*Donald Trump Republican*]$_{Neutral}$ *voter say Paul Ryan should endorse		*Donald Trump Republican* – *Neutral*
[**Harriette**$_{Negative}$] ***moved back to Chicago to care for her mom*: *And it ' s been terrible***		Harriette – *Negative*
[***Anthony Kiedis***$_{Negative}$] ***illness could affect Rock on the Range festival***		Anthony Kiedis – *Negative*

4.2.4 Conclusion and Future Direction

This paper presents a target-dependent multimodal sentiment recognition strategy called a visual-to-emotional-caption translation network. The idea put forward utilises facial emotions depicted in images as visual indicators of emotions. In this study, we propose a novel and efficient approach to establish a correlation between the target entity in textual information and the facial expressions depicted in visual media. Our approach has successfully achieved ground-breaking results on the Twitter2015 and Twitter-2017 datasets. The results indicate that our proposed solution surpasses a set of baseline models. This showcases the strength of our method in gathering emotional clues from the visual modality and achieving cross-modal alignment on visual-caption sentimental information. In the future, we would like to extend our proposed method for other multimodal tasks, such as the identification of hate speech, sarcasm, fake news, etc. The analysis of emotions conveyed by video is another exciting field of study with promising future prospects.

4.3 Significant Outcomes of this Chapter

The significant outcomes of this chapter are as follows:

- Developed a novel framework called the Visual-to-Emotional-Caption Translation Network (VECTN) to perform Target-Dependent Multimodal Sentiment Analysis (TDMSA). This network is composed of three main modules: the facial emotion description module, the target alignment and refinement module for face description, and the fusion module.
- The facial emotion description unit is responsible for generating a face description that includes various features such as age, gender, and emotion. The target alignment and refinement module estimates the cosine similarity between the visual input and the face descriptions with the target. In the fusion component, two robustly optimized pre-trained language models are utilized to simulate images, captions and face descriptions by a gating mechanism for feature fusion and noise reduction.
- We perform extensive studies and experiments utilizing standard datasets, namely Twitter-2015 and Twitter-2017, to demonstrate the effectiveness and reliability of our approach in simulating multimodal representations. Through our research, we aim to achieve remarkable cutting-edge results.

The following research works form the basis of this chapter:

- ❖ **A. Pandey** and D. Kumar Vishwakarma, "Target-Dependent Multimodal Sentiment Recognition Via a Visual-to-Emotional-Caption Translation Network using Visual-Caption Pairs." Under Minor Revision in ***Signal, Image and Video Processing*** (Pub: Springer). https://doi.org/10.48550/arXiv.2408.10248.

Chapter 5: Emoticon Prediction using Multimodal Content

5.1 Scope of this Chapter

In the digital age of social media platforms and the internet, an exciting new way of human interaction has emerged. It involves the combination of concise, readable text messages and imagery ideograms known as emoticons. Emoticons are tiny symbols that represent individuals, settings, and objects. These symbolic expressions have gained widespread acceptance as a standard for communication on the web. It is commonly used not just on Twitter but also on other well-known platforms like YouTube, WhatsApp, Telegram, Facebook, Instagram, and LinkedIn. According to Google Trends, the popularity of emoticons has been on the rise over the last decade. Emoticon prediction based solely on text has garnered attention and has been studied extensively from the perspective of Natural Language Processing. However, there is a dire need for further research on predicting emoticons based on multiple modalities. Hence, this research demonstrates the importance of integrating visual information with texts in the realm of multimodal communication. Specifically, we highlight how combining texts and images in online communities can lead to more precise emoticon prediction models. We examine the utilization of emoticons within the popular social media platform Twitter. We propose a multimodal strategy based on contrastive learning to forecast the emoticons associated with a Twitter post, considering both its textual content and accompanying image. We rely on visual modality to enhance the process of selecting the most suitable emoticons for a post. Our research demonstrates that incorporating both text and images in posts enhances the precision of emoticon prediction in comparison to relying solely on textual information. It may be inferred that textual and visual content incorporate distinct yet complementary aspects of using emoticons.

5.2 Contrastive Learning-based Multi-Modal Architecture for Emoticon Prediction by Employing Image-Text Pairs

5.2.1 Abstract

The emoticons are symbolic representations that generally accompany the textual content to visually enhance or summarize the true intention of a written message. Although widely utilized in the realm of social media, the core semantics of these emoticons have not been extensively explored based on multiple modalities. Incorporating textual and visual

information within a single message develops an advanced way of conveying information. Hence, this research aims to analyze the relationship among sentences, visuals, and emoticons. For an orderly exposition, this paper initially provides a detailed examination of the various techniques for extracting multimodal features, emphasizing the pros and cons of each method. Through conducting a comprehensive examination of several multimodal algorithms, with specific emphasis on the fusion approaches, we have proposed a novel contrastive learning-based multimodal architecture. The proposed model employs the joint training of dual-branch encoder along with the contrastive learning to accurately map text and images into a common latent space. Our key finding is that by integrating the principle of contrastive learning with that of the other two branches yields superior results. The experimental results demonstrate that our suggested methodology surpasses existing multimodal approaches in terms of accuracy and robustness. The proposed model attained an accuracy of **91%** and an MCC-score of **90%** while assessing emoticons using the Multimodal-Twitter Emoticon dataset acquired from Twitter. We provide evidence that deep features acquired by contrastive learning are more efficient, suggesting that the proposed fusion technique also possesses strong generalisation capabilities for recognising emoticons across several modes.

5.2.2 Proposed Methodology

This section contains an in-depth analysis of the proposed paradigm for multimodal emoticon predictions. The problem is outlined in the first subsection. Then, a Contrastive Learning based Multimodal Architecture depicted in **Figure 5.1** consists mainly of 3 components is introduced. The proposed model employs a dual-branch encoder design to accurately map text and images into a common latent space. This functionality is achieved through the joint training of both the encoders. Transformer-based visual encoder, Transformer-based textual encoder and an additional component involves the use of contrastive learning to uncover the hidden relationships within the text and pictures. It has been demonstrated that by integrating the principle of contrastive learning with that of the other two branches yields superior results. For simplicity, emoticon prediction based on text-image analysis is referred to multimodal emoticon prediction.

Task Definition

The task of multimodal emoticon prediction is defined as follows: Let I and T denote the sample spaces for an image and text, respectively. An example is comprised of a singular text string accompanied by supplementary visual data. Hence, each example consists of three elements: an image, a piece of text, and an emoticon as a class label. The expression for it is shown below:

$$\mathbb{E} = \{(\boldsymbol{I^0}, \boldsymbol{T^0}, \boldsymbol{L^0}), (\boldsymbol{I^1}, \boldsymbol{T^1}, \boldsymbol{L^1}), \dots \dots, (\boldsymbol{I^i}, \boldsymbol{T^i}, \boldsymbol{L^i}), \dots \dots, (\boldsymbol{I^{m-1}}, \boldsymbol{T^{m-1}}, \boldsymbol{L^{m-1}})\} \qquad (\boldsymbol{5.1})$$

where, $\mathbb{E}$ denotes the entire set of instance triplets, I^j symbolizes the images information, T^j represents the text-based data, L^j denotes emoticons numbered from $0-9$ as a class label for the i^{th} sample, and m is the count of the total number of examples in the entire dataset. Multimodal emoticon prediction aims to learn a mapping function $\mathbb{F}$: $(I^i, T^i) \rightarrow L^i$ predict most suitable emoticon for a multimodal tweet $\{(I^i, T^i, L^i | 0 \leq i \leq m-1)\}$. For multimodal emoticon prediction task, $L^i \in \{0,1,2,3,4,5,6,7,8,9\}$, where 0 represents 😭 while 9 denotes 👍. **Table 5.1** displays few samples of the dataset to represent multimodal Twitter posts for emoticon prediction.

Table 5.1 Few samples from the dataset to show emoticon prediction based on text-image pairs.

Number	**I**	**II**	**III**	**IV**
Textual modality	RT @lovsickgirls : lisa's reaction after she was told about her solo achievements	She is sooooo beautiful #MehndiHaiRachneWaali	RT @offl_Lawrence : A small request for my physically abled dancer boys	RT @AldrineEsther : My mother is no morehow am I going to survive oh God
Visual modality				
Predicted Emoticon	😍	❤️	🥺	😭

Contrastive Learning-based Multimodal Architecture

We have proposed a multimodal architecture based on the principle of contrastive learning for the emoticon prediction task to effectively simulate the relationship and compatibility between image and text content. **Figure 5.1** illustrates the proposed architecture. The proposed multimodal architecture comprises of three primary components: An Image encoder, a Text encoder, and a Contrastive learning component. The Image encoder is responsible for acquiring image embeddings, while the Text encoder acquires textual embeddings. The Contrastive learning element examines the pertinent attributes and similarities between the textual and image embeddings obtained in the preceding steps. The proposed model is demonstrated in the form of pseudocode in **Table 5.2**.

Table 5.2 Pseudocode for the proposed Contrastive Learning-Based Multimodal Architecture.

Pseudocode: Emoticon prediction using Contrastive Learning-Based Multimodal Architecture

Objective: To acquire knowledge of a mapping function $\mathbb{F}: (\boldsymbol{I^i}, \boldsymbol{T^i}) \rightarrow \boldsymbol{L^i}$ from a set of multimodal tweets $\left\{(\boldsymbol{I^i}, \boldsymbol{T^i}, \boldsymbol{L^i} \middle| \mathbf{0} \leq \boldsymbol{i} \leq \boldsymbol{m} - \mathbf{1})\right\}$.

Input: Image set $\boldsymbol{I} = \{\boldsymbol{I^0}, \boldsymbol{I^1}, \dots, \boldsymbol{I^i}, \dots, \boldsymbol{I^{m-1}}\}$ and Text set $\boldsymbol{T} = \{\boldsymbol{T^0}, \boldsymbol{T^1}, \dots, \boldsymbol{T^i}, \dots, \boldsymbol{T^{m-1}}\}$, $\boldsymbol{m}$ is the count of the total number of examples in the entire dataset.

Output: Multimodal emoticon prediction task, $\boldsymbol{L^i} \in \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}, \mathbf{7}, \mathbf{8}, \mathbf{9}\}$, where **0** represents 😭 while **9** denotes 👍.

```
1. Extract image features "$\mathbf{e_h}$" of size "$\mathbb{D} - \mathbf{768}$" from the visual samples using Transformer-based visual
   encoder by Eqn. (5.2), Eqn. (5.3), and Eqn. (5.4) ;
2. Extract text features "$\mathbb{T}$" of size "$\mathbb{D} - \mathbf{768}$" from the caption samples using Transformer-based textual
   encoder.
3. for $\mathbf{E_{epochs}}$←1 to Epochs do;

    #identify and extract modality-specific feature representations
    $\boldsymbol{e_h}$ = visual _encoder($\boldsymbol{I}$) $\mathbf{Eqn.\,(5.2), Eqn.\,(5.3), and\ Eqn.\,(5.4)}$;
    $\mathbb{T}$ = textual _encoder($\boldsymbol{T}$);

    # joint multimodal embedding where $\mathbf{W_I}$ and $\mathbf{W_T}$ are weights which are learned projection of image and text
    to embedding
    $\mathbb{Y} = \mathbb{L}_2\_\boldsymbol{normalization}(\boldsymbol{np.dot}(\boldsymbol{e_h}, \boldsymbol{W_I}), \boldsymbol{axis} = \mathbf{1})$
    $\mathbb{Z} = \mathbb{L}_2\_\boldsymbol{normalization}(\boldsymbol{np.dot}(\mathbb{T}, \boldsymbol{W_T}), \boldsymbol{axis} = \mathbf{1})$

    # cosine similarity between pairs of vectors
    $\boldsymbol{logits} = \boldsymbol{np.dot}(\mathbb{Y}, \mathbb{Z}.\boldsymbol{T}) * \boldsymbol{np.exp}(\boldsymbol{\tau})$

    # calculating contrastive loss function to optimize the cosine similarity between the text and image
    embedding for $\boldsymbol{N}$ real pairs in the batch
    $\boldsymbol{labels} = \boldsymbol{np.arange}(\boldsymbol{m})$
    $\boldsymbol{loss\_I} = \boldsymbol{cross\_entropy\_loss}(\boldsymbol{logits}, \boldsymbol{labels}, \boldsymbol{axis} = \mathbf{0})$
    $\boldsymbol{loss\_T} = \boldsymbol{cross\_entropy\_loss}(\boldsymbol{logits}, \boldsymbol{labels}, \boldsymbol{axis} = \mathbf{1})$
    $\boldsymbol{loss} = (\boldsymbol{loss\_I} + \boldsymbol{loss\_T})/\mathbf{2}$ using $\boldsymbol{Eqn.}\,(\mathbf{5.6}), \boldsymbol{Eqn.}\,(\mathbf{5.7}), \boldsymbol{and\ Eqn.}\,(\mathbf{5.8})$

    # At last The pair with the maximum cosine similarity scores are then fed into the simple artificial neural
    network to provide probabilities for each label.
    $\boldsymbol{L^i} = \boldsymbol{Artificial\,neural\,network}\big((\mathbb{Y}, \mathbb{Z})_{\boldsymbol{Maximum\_cosine}}\big)$ find the loss and execute the backpropagation;

    end
```

Transformer-based Visual Encoder

This sub-section will provide a comprehensive explanation of the proposed methodology for extracting pertinent details from visual cues. In the realms of image, audio and text analysis, the performance of deep neural network-based architectural designs succeeds over the conventional hand-crafted approaches for classification [71], [126], [127], [128]. The primary reason for their effectiveness lies in their capacity to enhance end-to-end relationships, facilitate autonomous feature learning, ensure efficient scalability, establish semantic representations, and offer flexibility. Furthermore, several researchers have been working to enhance the performance of pre-trained and custom-built ConvNets over the last few years by including attention [5], [129], [130] an additional architectural design component. The utilization of ConvNets is not deemed essential in modern times. This is because a standalone transformer model [131] can effectively handle visual classification tasks by directly processing sequences of patches of images. Each patch is then converted into a vector via a Large Language Model referred as LLMs and processed using a transformer architecture. Hence, Dosovitskiy et al. [131] is capable of gathering and analysing global information in images, unlike ConvNets, which can only extract local aspects.

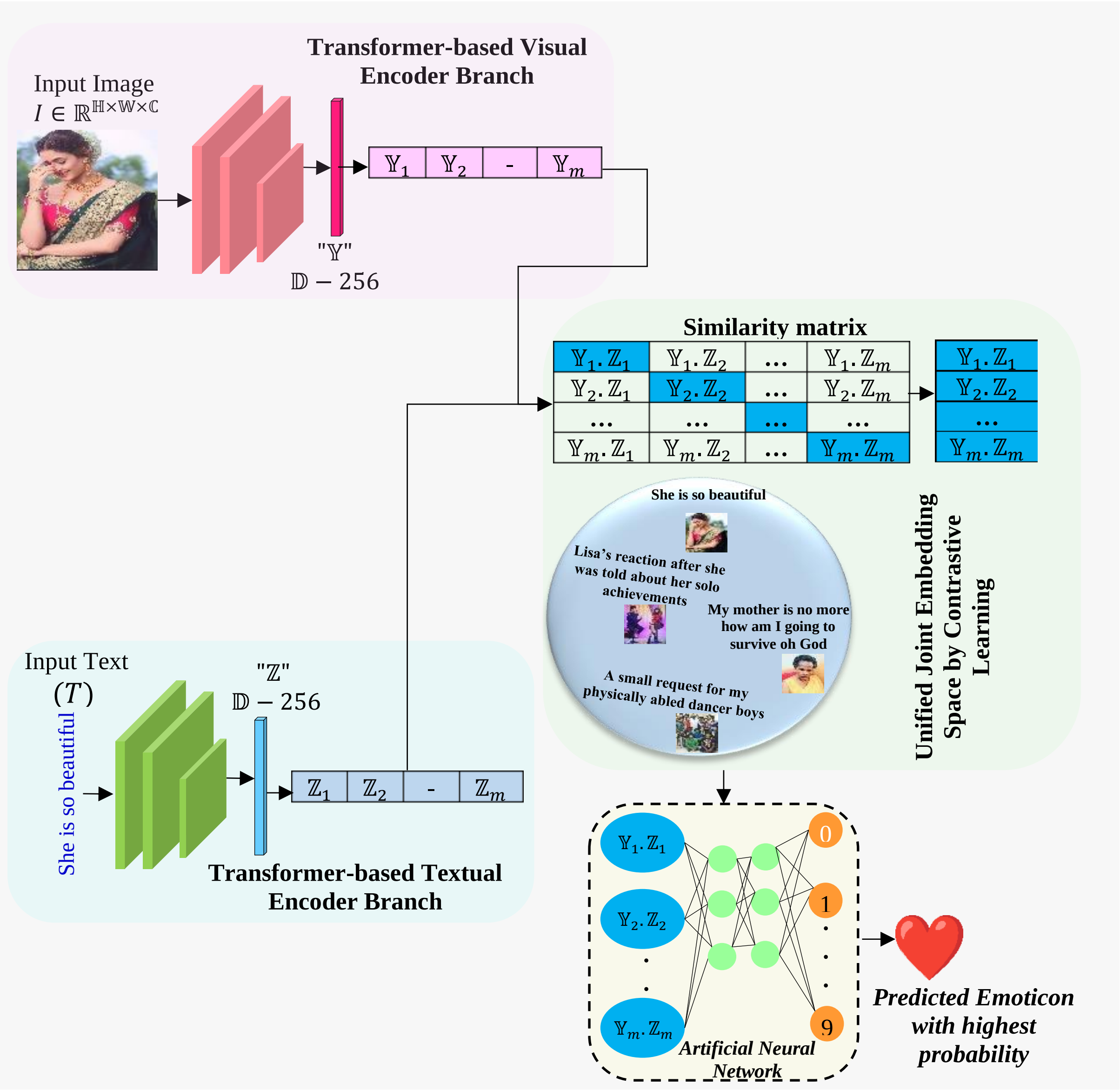

Transformer-based Visual Encoder Branch
Input Image
$I \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}\times\mathbb{C}}$
$\mathbb{Y}_1$ $\mathbb{Y}_2$ - $\mathbb{Y}_m$
"$\mathbb{Y}$"
$\mathbb{D}-256$
Similarity matrix
$\mathbb{Y}_1.\mathbb{Z}_1$ $\mathbb{Y}_1.\mathbb{Z}_2$ … $\mathbb{Y}_1.\mathbb{Z}_m$
$\mathbb{Y}_2.\mathbb{Z}_1$ $\mathbb{Y}_2.\mathbb{Z}_2$ … $\mathbb{Y}_2.\mathbb{Z}_m$
… … … …
$\mathbb{Y}_m.\mathbb{Z}_1$ $\mathbb{Y}_m.\mathbb{Z}_2$ … $\mathbb{Y}_m.\mathbb{Z}_m$
$\mathbb{Y}_1.\mathbb{Z}_1$ $\mathbb{Y}_2.\mathbb{Z}_2$ … $\mathbb{Y}_m.\mathbb{Z}_m$
She is so beautiful
Lisa's reaction after she was told about her solo achievements
My mother is no more how am I going to survive oh God
A small request for my physically abled dancer boys
Unified Joint Embedding Space by Contrastive Learning
Input Text
(T)
She is so beautiful
"$\mathbb{Z}$"
$\mathbb{D}-256$
$\mathbb{Z}_1$ $\mathbb{Z}_2$ - $\mathbb{Z}_m$
Transformer-based Textual Encoder Branch
$\mathbb{Y}_1.\mathbb{Z}_1$
$\mathbb{Y}_2.\mathbb{Z}_2$
$\mathbb{Y}_m.\mathbb{Z}_m$
0
1
9
Artificial Neural Network
Predicted Emoticon with highest probability

Figure 5.1 Proposed Contrastive learning based multi-modal architecture for emoticon prediction.

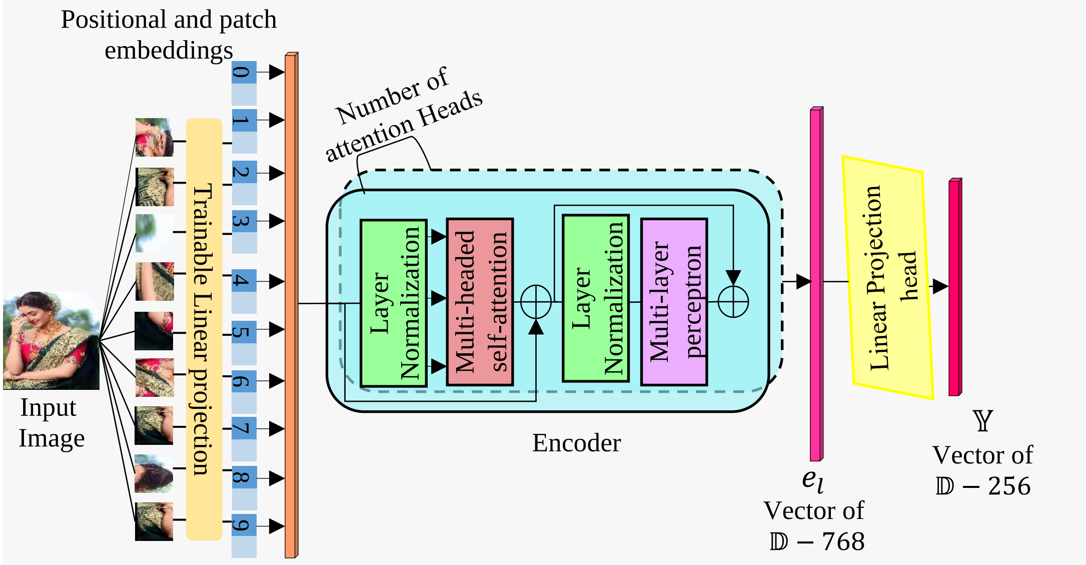

Positional and patch embeddings
0
1
2
3
4
5
6
7
8
9
Trainable Linear projection
Input Image
Number of attention Heads
Layer Normalization
Multi-headed self-attention
Layer Normalization
Multi-layer perceptron
Encoder
e_l
Vector of $\mathbb{D}-768$
Linear Projection head
$\mathbb{Y}$
Vector of $\mathbb{D}-256$

Figure 5.2 Image encoder to obtain the embeddings for all the image samples present in the dataset.

Taking these considerations into account, a transformer-based model Vit-base-patch32 depicted in **Figure 5.2** is utilized to retrieve the most prominent features from the image samples of Multimodal-TwitterEmoticon dataset. Firstly, the initial step involves the scaling of image samples denoted as $I \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}\times\mathbb{C}}$ from the entire collection of the dataset to a resolution of 224×224. This scaling is done to prepare the samples for subsequent processing. Typically, transformers take a 1-dimensional series of token embeddings as input. We address this by transforming the two-dimensional input image into a sequence of flattened patches $I_p \in \mathbb{R}^{(\mathbb{P}^2\cdot\mathbb{C})\times\mathbb{N}}$of size $(\mathbb{P} \times \mathbb{P}) \in 32 \times 32$. A trainable linear projection is used to transform the acquired patches into $\mathbb{D}$ −dimensional space using **Eqn.** $(\mathbf{5.2})$. This allows the uniform latent vector size $\mathbb{D}$ of the transformer to be employed throughout all the layers. The projection's end product is called patch embedding. Additionally, to preserve positional information, position embeddings $E_{Position}$ are appended to the patch embeddings. Analogous to BERT's $[class]$ token [57], we augment the patches with a learnable embedding $(e_0^0 = I_{Class})$. The final obtained patches are then pass through a visual encoder layer which is nothing but the h number of self-attention operations, called "heads", which will run in parallel. The final outcome of the Transformer-based visual encoder is used as the representation for the images denoted as e_h of $\mathbb{D}-768$ in **Eqn.** $(\mathbf{5.3})$ and **Eqn.** $(\mathbf{5.4})$. The GELU non-linearity is used on two consecutive layers to build the multilayer perceptron.

Finally, we need to translate the image and text embeddings into the same vector space so that we may process this acquired vector together with the textual embedding. This is why the obtained vector of $\mathbb{D}-768$ goes through a linear projection head, basically just an artificial neural network with a linear activation function to get the desired result. Thus, the final vector $\mathbb{Y}$ will be $\mathbb{D} - 256$, as stated in **Eqn. (5.5)**, and will be used for further operations.

$$\boldsymbol{e_0 = \left[I_{Class};\ I_p^1E; I_p^1E; \dots\dots; I_p^1E\right] + E_{Position}} \tag{5.2}$$

$$\boldsymbol{e'_h = Multi - headed\ self\ attention\big(Layer\ Normaliztion(e_{h-1})\big) + e_{h-1}} \tag{5.3}$$

$$\boldsymbol{e_h = Multi - layer\ perceptron\big((e'_h)\big) + e'_h} \tag{5.4}$$

$$\mathbb{Y} = \boldsymbol{Linear\ projection\ head(e_h^0)} \tag{5.5}$$

where, $\mathbb{H} \times \mathbb{W}$ denotes the height and width of the original image sample I, $\mathbb{C}$ symbolizes the depth of an image, $\mathbb{P} \times \mathbb{P}$ denotes the resolution for each transformed patch I_p of an image, $h = \{1,2,\dots,H\}$ denotes number of attention heads, $E \in \mathbb{R}^{\mathbb{D}\times(\mathbb{P}^2\cdot\mathbb{C})}$ and $E_{Position} \in \mathbb{R}^{\mathbb{D}\times(\mathbb{N}+1)}$.

Transformer-based Textual Encoder

The latest advancements in transformer-based models incorporate a self-attention mechanism to prioritise relevant information while disregarding irrelevant information. Recent studies primarily utilised these architectures [57] exclusively for text processing. Although there are different variants of BERT that aim to extract text features, but they differ from the [132] architecture design due to their reliance on absolute position encoding rather than relative position codification [133]. Raffel et al. [132] utilises relative positional encoding to incorporate information between pairwise positions, whereas absolute positional encoding does not take this into consideration. In absolute positional encoding, the embeddings for each location are initialised at random. As a result, the relationship between different positions is unknown. Instead of random embedding initialization, relative positional encoding generates a pairwise vector of size $(V, 2 * V - 1)$, with the row index representing the desired word and the column index representing its position distance from previous and subsequent words. This information regarding relative positioning is dynamically integrated into the keys and values as part of the computation process in attention modules. Therefore, it is advantageous to use a transformer-based model that supports relative positional encoding. This feature provides greater flexibility to the model and leads to more accurate result.

Considering these factors, we have employed the encoder of [132] to transform the text into their respective embedding's. During the first step, the sentence T provided as input to the Transformer-based text encoder is partitioned into tokens $T = \{t_1, t_2, \dots\dots, t_n\}$, and each token is subsequently transformed into a vector representation $T^{vectorization} = \{t_1^v, t_2^v, \dots\dots, t_n^v\}$. Next, the acquired vectors are sent for relative position encoding. The results achieved through relative position encoding are then fed into the transformer-based text encoder component, which produces a text embedding vector $\mathbb{T}$ of size $\mathbb{D} - 768$. In order to ensure that the image and text embedding's are aligned in the identical vector space, the$\mathbb{D} - 768$ is sent to a linear projection head to compress the text embedding to a vector $\mathbb{Z}$ of size $\mathbb{D} - 256$.

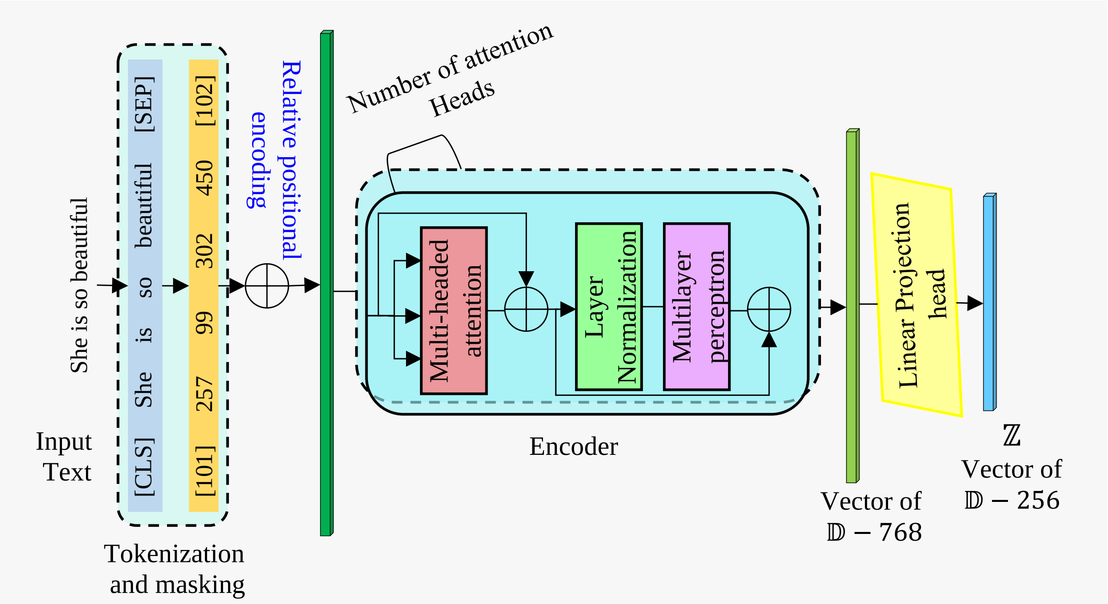


Figure 5.3 Text encoder to obtain the embeddings for all the text samples present in the dataset.

Contrastive Learning

The primary goal of multi-modal learning is to understand the relationships between images and text in a given batch $B = \{I^i, T^i\}_{i=1}^{m}$ which consists of m examples represented as (T^i, T^i). Motivated by the architectural design of Contrastive visual-textual pre-training, we applied the concept of a similarity matrix to examine the relevant features and similarities between the embeddings of image-text pairs acquired by the transformer-based encoder in previous stages. In order to do this, firstly contrastive visual-textual pre-training model [134] trains an image encoder $f(I^i, L^i)$ and a text encoder $g(T^i, L^i)$, such that the embeddings of image-text pairings $\{I^i, T^i\}_{i=1}^{m} = \{f\left(I^i, L^i\right), g(T^i, L^i)\ \}_{i=1}^{M}$ become more similar to each other. It is important to mention that $\mathbb{Y}$ and $\mathbb{Z}$ are unit vectors of size $\mathbb{D} - 256$ that have been normalised using the $\mathbb{L}_2$ norm in the encoding phase. These vectors are then located on the similar hypersphere. Contrastive visual-textual pre-training model employs the loss function $C\ (\cdot,\cdot)$ as specified in the cited approach [13] with the goal of ensuring that pairs $(\mathbb{Y}, \mathbb{Z})$ possess both similarity and distance. The formulation is represented by the **Eqn. (5.6) and Eqn. (5.7)**:

$$\boldsymbol{C}\ (\boldsymbol{I},\boldsymbol{T}) = \frac{\mathbf{1}}{\boldsymbol{m}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m}} -\boldsymbol{log}\,\frac{\mathbf{exp}\ (\boldsymbol{sim}(\mathbb{Y}_{\boldsymbol{i}},\mathbb{Z}_{\boldsymbol{i}}\ )/\boldsymbol{\tau}}{\boldsymbol{y}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m}}\mathbf{exp}\ (\boldsymbol{sim}(\mathbb{Y}_{\boldsymbol{i}},\mathbb{Z}_{\boldsymbol{i}})/\boldsymbol{\tau}} \qquad (\mathbf{5.6})$$

$$\boldsymbol{\mathcal{L}}_{\textbf{Contrastive visual}-\textbf{textual}} = \frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{C}\ (\mathbb{Y}_{\boldsymbol{i}}, \mathbb{Z}_{\boldsymbol{i}}) + \ \boldsymbol{C}\ \mathbb{Y}_{\boldsymbol{i}}, \mathbb{Z}_{\boldsymbol{i}}) \qquad (\mathbf{5.7})$$

Similar to other approaches in computational linguistics, [134] uses a dot product to calculate similarity $(sim(.,.))$ between two vectors. It also employs a learnable temperature parameter $(\boldsymbol{\tau})$ to adjust the magnitude of the observed similarity and $\mathcal{L}_{\text{Contrastive visual}-\text{textual}}$ is the mean loss of combined image and text pairs. In general, the model learns a multi-modal embedding space by training its encoders to maximize the cosine similarity between the picture and text embeddings of the N correct pairs in the batch, while minimizing the cosine similarity between the embeddings of the $N^2 - N$ incorrect pairings. Hence, by representing both images and texts using the coherent embedding space, our proposed contrastive learning-based model is capable of doing two essential tasks: (a) Optimize the cosine similarity between the image and text embeddings for N real pairs in the batch, aiming to maximize it. (b) Additionally, minimize the cosine similarity between the embeddings of $N(N-1)$ erroneous pairings. During the pre-training phase of [134], the model is trained using a contrastive loss function, as depicted in **Eqn.** $(\mathbf{5.7})$. The main objective of this loss function is to promote the aggregation of embeddings for related or *positive pairings* (text and picture that match) while simultaneously driving apart the embeddings for unrelated or *negative pairs* (text and image that do not match). The purpose of the model is to minimise the loss for positive pairs and maximise the loss for negative pairs. The pair with the maximum cosine similarity scores, as indicated by **Eqn.** $(\mathbf{5.8})$, is between two $\mathbb{D}$-$dimensional$ vectors, referred to as $\mathbb{Y}$ and $\mathbb{Z}$. These obtained vectors are then fed into the simple artificial neural network and processed using softmax as an activation function to provide probabilities for each label.

$$\boldsymbol{Sim}(\boldsymbol{I},\boldsymbol{T}) = \boldsymbol{\delta} = \frac{\mathbb{Y}\cdot\mathbb{Z}}{||\mathbb{Y}||\cdot||\mathbb{Z}\,||} = \frac{\boldsymbol{\Sigma}_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m}}\mathbb{Y}_{\boldsymbol{i}}\times\mathbb{Z}_{\ \boldsymbol{i}}}{\sqrt{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m}}(\mathbb{Y}_{\boldsymbol{i}})^{\mathbf{2}}}\times\sqrt{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m}}(\mathbb{Z}_{\ \boldsymbol{i}})^{\mathbf{2}}}} \qquad (\mathbf{5.8})$$

5.2.3 Experimental Setup and Results

The following section of the research article will present a detailed description of the optimal experimental configurations, the dataset utilised, and the experimental results obtained using our proposed approach.

Experimental Configuration

It is crucial to determine an appropriate range for hyper-parameters in a learning algorithm, as they play a significant role in controlling the learning process. The hyper-parameters underwent a random testing process using various values. As a result, the values that yielded the most optimal outcomes for our proposed framework were subsequently set as fixed. Since, the joint embedding from both modalities was obtained using the concept of contrastive learning. Therefore, the hyper-parameter values used were almost similar to those in [134]. The encoders utilised for both images and text in our proposed model referred as Contrastive Learning based Multimodal Architecture are equipped with 12 number of attention heads. The image samples denoted as $I \in \mathbb{R}^{\mathbb{H} \times \mathbb{W} \times \mathbb{C}}$ from the dataset are scaled to a resolution of 224×224 before being passed to a visual encoder. Adam is used as an optimizer with the default learning rate of 0.001. The values of β_1 and β_2 for using Adam are set to 0.9 and 0.99. A dropout rate of 0.2 has been implemented. The contrastive loss function is also used (shown in **Eqn.** $(\mathbf{5.7})$) with a temperature scaling factor $\tau = 0.1$ and margin $M = 0.7$ to combine the embeddings of images and text with the highest similarity score. The proposed architecture is fine-tuned for 20 epochs with a batch size of 32. By the 20th epoch, it becomes evident that there has been no noticeable rise in accuracy, as the results reach a point of saturation. From a total of 21k samples, 16k were allocated for training purposes, while the remaining 5k samples were dedicated to testing our proposed model.

Hardware Configuration

An Azure virtual machine with premium specifications was used for training and evaluation of the proposed model. This machine provides 100 GB of hard drive space, NVIDIA-A100 GPU with 80 GB of graphics memory, CUDA version 12.4, and 384 GB of RAM. We utilised an $8 \times A100$ GPU cluster for 41 days, which is roughly comparable to 1000 GPU hours, and a pool memory of 640 GB to train the entire model from end to end. To derive useful information from text-image pairings, models were constructed using the Pytorch frameworks.

Dataset Description

We have employed the Multimodal-TwitterEmoticon dataset, provided by Ebrahimian et al. [135], to evaluate our suggested approach for emoticon prediction. Multimodal-TwitterEmoticon dataset consists of 21k English tweets, each containing an image, the accompanying text, and a single emoticon. Emoticons used in tweets varies based on their popularity. **Figure 5.4 (A)** illustrates the frequency in percentage for most widely used emoticons. These ten most prevalent and often used emoticons have been taken into account for analyses. **Figure 5.4 (B)** also depicts the statistical information regarding the number of samples associated with each 10 emoticons across the entire dataset where 0 represents 😭 while 9 denotes 👍 . The experiment focuses on predicting emoticons based on image-text pairs. From a total of 21k samples, 16k were allocated for training purposes, while the remaining 5k samples were dedicated to testing our proposed model.

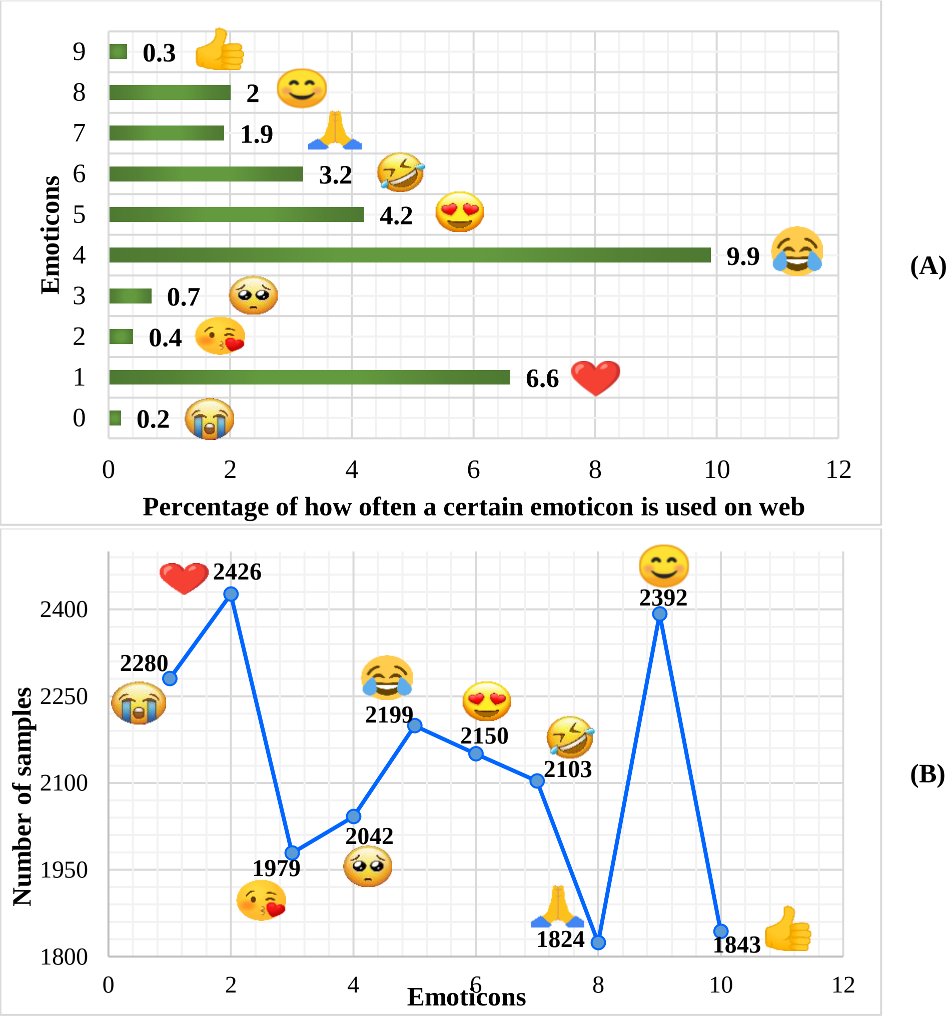


Figure 5.4 (A) provides information about how often multiple emoticons are utilized on web, while **(B)** indicates the number of samples corresponds to each emoticon in Multimodal-TwitterEmoticon dataset.

Experimental Results, Baseline Comparison and Analysis

We evaluate our proposed Contrastive Learning-Based Multimodal Architecture on Multimodal-TwitterEmoticon dataset provided by provided by Ebrahimian et al. [135]. The experimental results of our model are presented in **Table 5.3**. In addition to calculating the overall accuracy, macro-average and weighted-average, we have computed the $Precision$, $Recall$, and $F1 - score$ for each class separately to better analyse the results. By taking into consideration every aspect of the confusion matrix represented in **Figure 5.7**, Matthews Correlation Coefficient ($MCC - Score$) provides a more fair evaluation. Also, when the repercussions of false positives and false negatives differ, MCC becomes more relevant than F1. Hence, we have also calculated MCC-Score. **Figure 5.5** shows the training and validation loss curves, which can be used to better rely on the results. Receiver operating characteristic (ROC) curve is considered to be more significant when evaluating model's performance. Unlike $Accuracy$, which solely evaluates the number of right predictions, this statistic takes into account the trade-offs between $Recall$ and $Precision$. It reveals the degree to which the model can differentiate across categories. A higher AUC indicates that the model is effective at making accurate predictions. **Figure 5.6** displays the ROC curve, which has an area under the curve (AUC) of **0.82**.

Table 5.3 Experimental Results on Multimodal-TwitterEmoticon dataset for emoticon prediction.

$\boldsymbol{Class\ Labels}$	$\boldsymbol{Precision}$	$\boldsymbol{Recall}$	$\boldsymbol{F1 - score}$
0	0.85	0.83	0.84
1	0.92	0.94	0.93
2	0.89	0.91	0.90
3	0.93	0.92	0.92
4	0.90	0.93	0.91
5	0.88	0.90	0.89
6	0.91	0.89	0.90
7	0.94	0.95	0.94
8	0.95	0.93	0.94
9	0.93	0.91	0.92
$\boldsymbol{Macro - average}$	**0.910**	**0.911**	**0.909**
$\boldsymbol{Weighted - average}$	**0.91**	**0.91**	**0.91**
$\boldsymbol{Overall\ Accuracy}$	**0.91**		
$\boldsymbol{MCC - Score}$	**0.90**		

After examining several recent cutting-edge research articles in the field of emoticon prediction, it becomes clear that there is a scarcity of studies that have effectively combined visual and textual data to predict emoticons. As a result, to assess the strength of our proposed model, we have identified only two baseline methods that incorporate both text as well as images to predict emoticons. In baseline 1 Francesco et al. [136] scraped Instagram posts to collect multimodal content for emoticon prediction. This dataset is not publicly accessible. In this work, a Bi-LSTM model [99] as well as FastText [137] was utilised to extract the most significant features from the text samples present in the dataset. Additionally, a ResNet-101 model was employed to obtain features from the picture samples of the dataset.

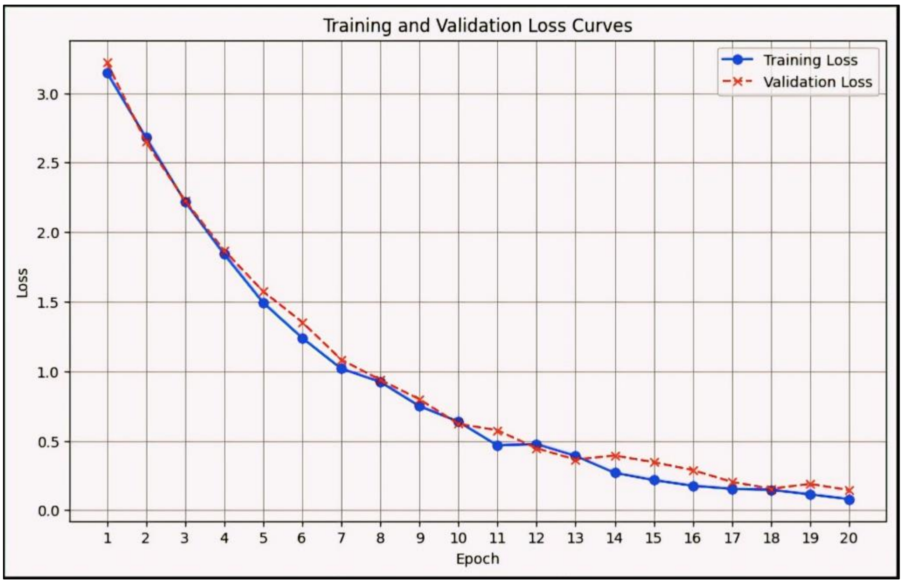


Figure 5.5 Training and validation loss curves for the proposed architecture.

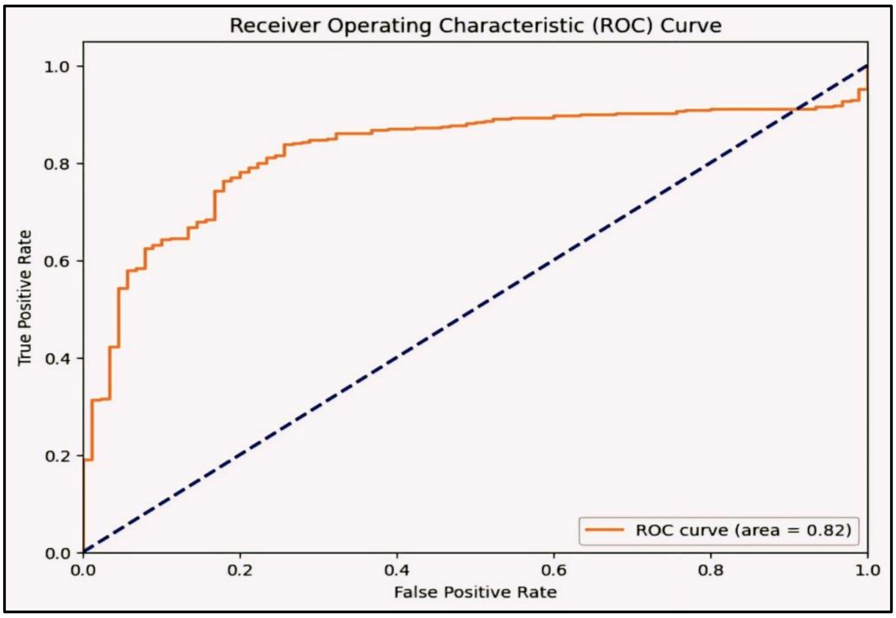
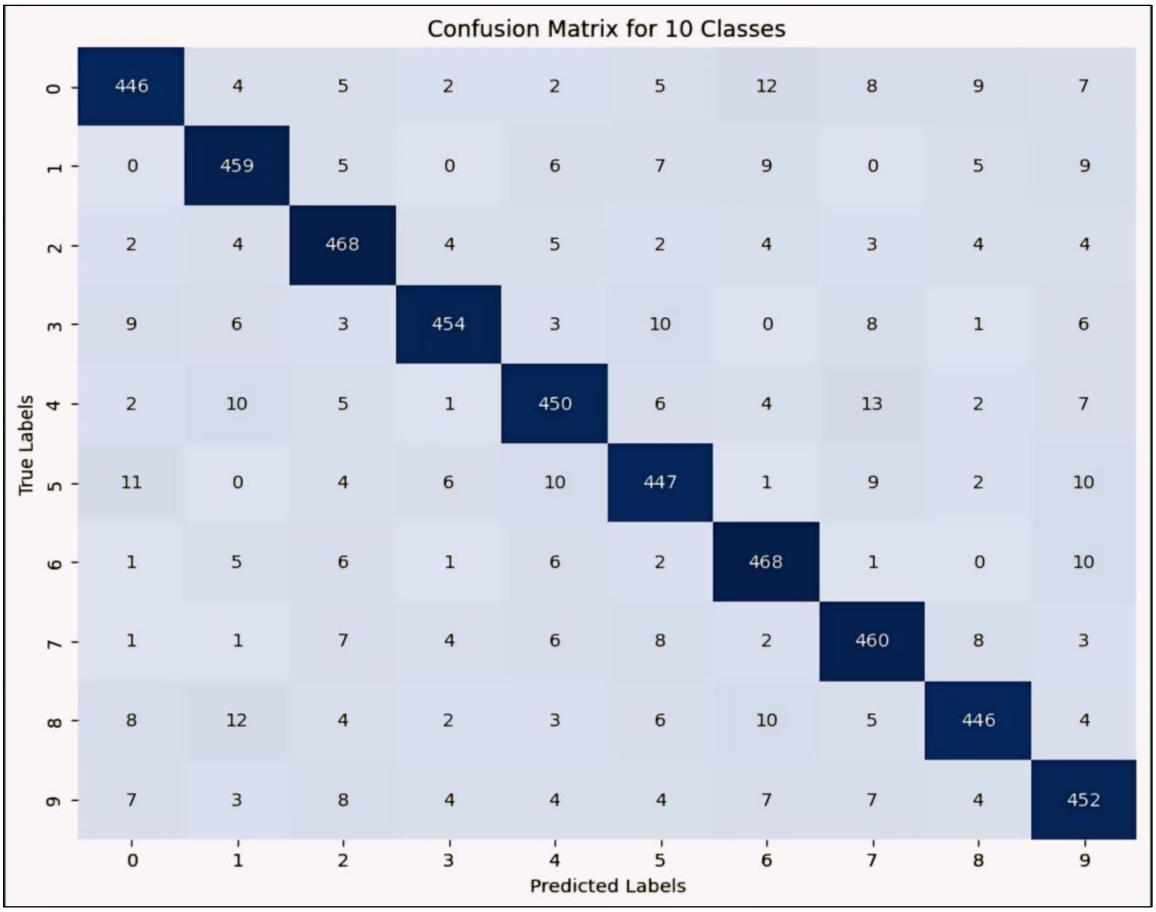


Figure 5.6 Receiver operating characteristic curve for the proposed approach.

Baseline 1 also includes a comparison of the FastText model with the character and word based B-LSTMs proposed by Barbieri et al. [138]. The FastText model proves to be highly effective, even outperforming the character-based B-LSTM in emoticon forecasting task. The findings of this research demonstrate that FastText works well for representing brief text from social networking platforms like Instagram or Twitter. Due to the unavailability of the dataset for this study, we have utilised the Multimodal-TwitterEmoticon dataset provided by Ebrahimian et al. [135] to evaluate our approach and compare it with the method proposed by Francesco et al. [136]. The results for the evidence are presented in **Table 5.4**.



	0	1	2	3	4	5	6	7	8	9
0	446	4	5	2	2	5	12	8	9	7
1	0	459	5	0	6	7	9	0	5	9
2	2	4	468	4	5	2	4	3	4	4
3	9	6	3	454	3	10	0	8	1	6
4	2	10	5	1	450	6	4	13	2	7
5	11	0	4	6	10	447	1	9	2	10
6	1	5	6	1	6	2	468	1	0	10
7	1	1	7	4	6	8	2	460	8	3
8	8	12	4	2	3	6	10	5	446	4
9	7	3	8	4	4	4	7	7	4	452

Figure 5.7 Confusion matrix obtained for the test samples for the ten class labels by using our proposed method.

Table 5.4 Experimental Results on Multimodal-TwitterEmoticon dataset by using the method of baseline 1 (Francesco et al. [32]).

$\boldsymbol{Class\ Labels}$	$\boldsymbol{Precision}$	$\boldsymbol{Recall}$	$\boldsymbol{F1 - score}$
0	0.75	0.72	0.73
1	0.72	0.73	0.72
2	0.69	0.70	0.69
3	0.73	0.72	0.72
4	0.70	0.73	0.71
5	0.78	0.69	0.73
6	0.71	0.70	0.70
7	0.73	0.74	0.73
8	0.74	0.73	0.73
9	0.74	0.71	0.72
$\boldsymbol{Macro - average}$	**0.73**	**0.72**	**0.72**
$\boldsymbol{Weighted - average}$	**0.73**	**0.72**	**0.72**
$\boldsymbol{Overall\ Accuracy}$	**73.00%**		
$\boldsymbol{MCC - Score}$	**0.71**		

In baseline 2, Ebrahimian et al. [135] also scrapped multimodal post from the Twitter platform. In this research EfficientNet-B7 has been utilized to obtain visual features. In addition, a topic modelling technique named Latent Dirichlet Allocation is also employed to discover embedded topics within the text. The extracted topics are then combined with the transformer-based network to enhance the efficacy of the proposed model. As compared to Ebrahimian et al. [135] our proposed Contrastive Learning-Based Multimodal Architecture performs far better in terms of multiple performance metrics. **Table 5.5** highlights the comparison of our results with both the baseline approaches discussed above. A comparative evaluation of the baseline methods in terms of $Accuracy, Precision,$ $Recall,$ and $F1 - score$ with our suggested framework on Multimodal-TwitterEmoticon dataset is illustrated in graphical form in **Figure 5.8.**

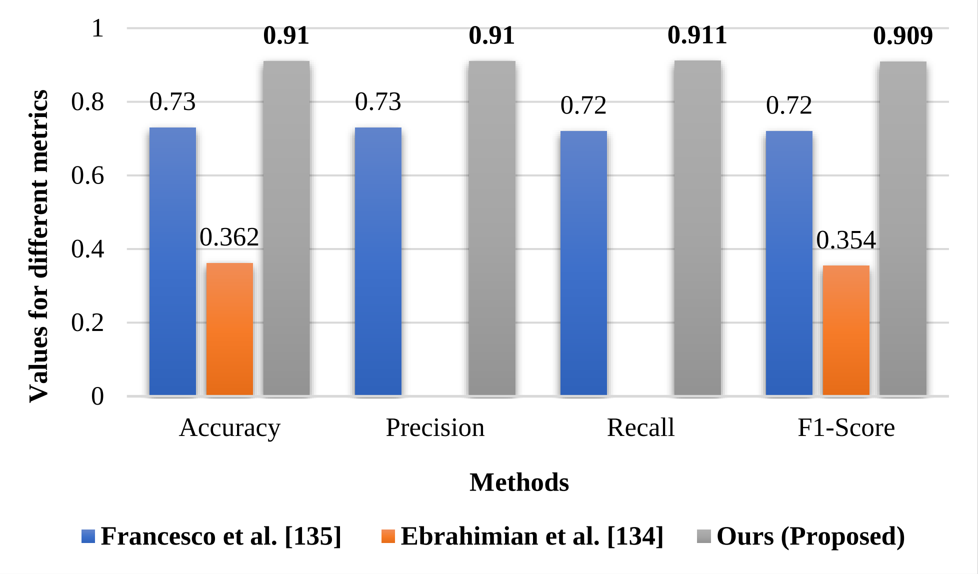


Figure 5.8 Graphical representation to represent that our proposed model surpasses existing cutting-edge baselines.

Table 5.5 Comparison of our proposed method with baseline methods on Multimodal-TwitterEmoticon dataset.

$Methods$	$Accuracy$	$Precision$	$Recall$	$F1 - score$
Francesco et al. [136]	0.73	0.73	0.72	0.72
Ebrahimian et al. [135]	0.362	-	-	0.354
Ours (Proposed)	**0.91**	**0.91**	**0.91**	**0.91**

Additionally, we have evaluated the predictions in the absence of contrastive learning to further test the robustness of our suggested model. In brief, we have combined the information retrieved from the image and text encoders and then passed them through a basic artificial neural network without depending on contrastive learning approach. However, it significantly diminishes the outcomes. Furthermore, we have evaluated the efficacy of our suggested model by means of t-Distributed Stochastic Neighbour Embedding (t-SNE). **Figure 5.9** shows that the model can create distinct clusters for different classes. Accordingly, it is able to glean valuable insights from the data.

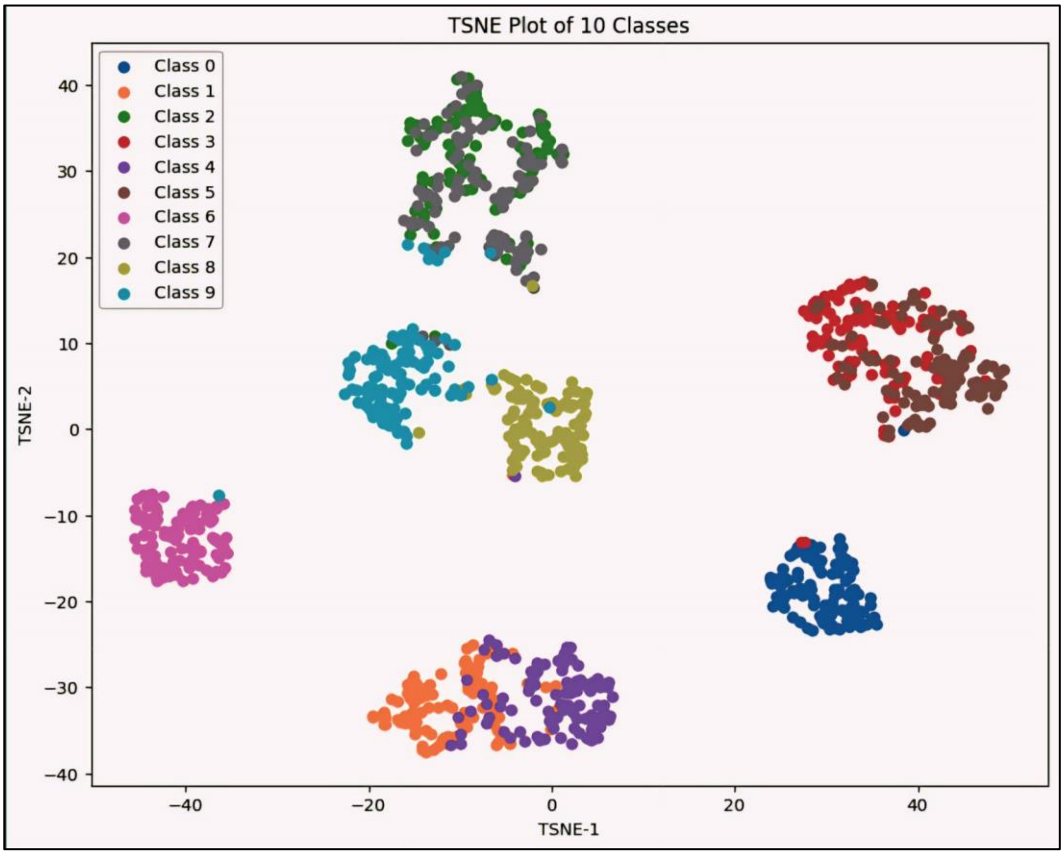


Figure 5.9 t-Distributed Stochastic Neighbour Embedding representation to test the strength of the proposed architecture.

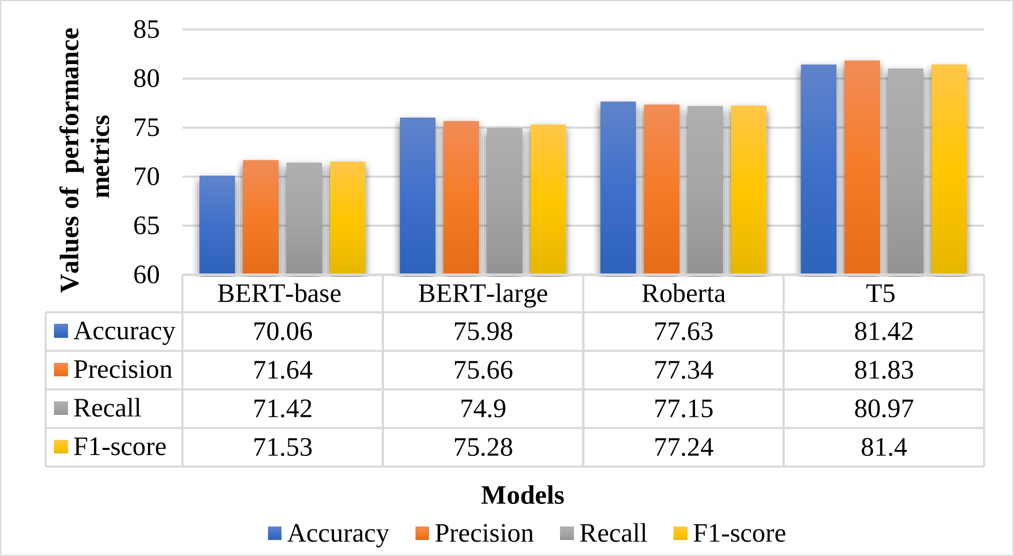
Hence, based on the obtained results, the following conclusion can be deduced. The proposed architecture demonstrates superior results compared to the baseline approaches, as it outperforms them in terms of $Accuracy$, P$recision$, $Recall$, and $F1 - score$. Thus it is reasonable to conclude that our contrastive-based multimodal architecture works effectively by uncovering the hidden relationships within the text and images. In general, the model learns a multi-modal embedding space by the joint training of two encoders to maximize the cosine similarity between the picture and text embedding of the correct pairs in the batch, while minimizing the cosine similarity between the embedding of the incorrect pairings.

Ablation Study

To evaluate the efficacy of each individual element in the proposed architecture, we have performed an ablation study. In the ablation experiment, the text encoder proposed in the study is substituted with BERT-base, BERT-large[98] , Roberta[117] and T5 models, while the image encoder is replaced with the ResNet-101[139], EfficientNet-B7[140], ResNext[141], RegNet[142] and Vit-base-patch32 models. The results have undergone analysis using both unimodal and multimodal configurations.

Table 5.6 Ablation study to assess the robustness of the text branch for unimodal configuration.

Text Encoder	$Accuracy$	$Precision$	$Recall$	$F1 - score$
BERT-base [98]	70.06	71.64	71.42	71.53
BERT-large [98]	75.98	75.66	74.90	75.28
Roberta[117]	77.63	77.34	77.15	77.24
T5 [132]	**81.42**	**81.83**	**80.97**	**81.40**



	BERT-base	BERT-large	Roberta	T5
■ Accuracy	70.06	75.98	77.63	81.42
■ Precision	71.64	75.66	77.34	81.83
■ Recall	71.42	74.9	77.15	80.97
■ F1-score	71.53	75.28	77.24	81.4

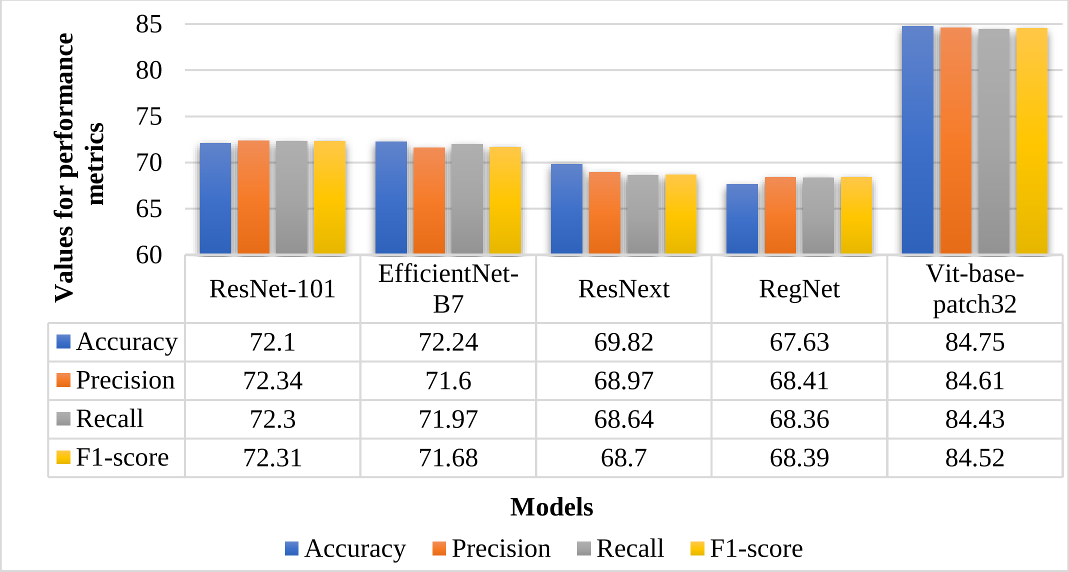
Figure 5.10 A graphical representation for an ablation study that aims to evaluate the resilience of the text branch in a unimodal configuration.

Based on the results of the ablation study presented in **Table 5.6** and **Figure 5.10** it is clear that the architecture used for the text encoder in the proposed framework is more efficient. The reason behind this is that the model such as BERT-base [98], BERT-large[98], and Roberta[117] supports absolute positional encoding instead of relative positional encoding. As a result, the relationship between different positions is unknown. Since, T5 [132] utilises relative positional encoding to incorporate information between pairwise positions. This information regarding relative positioning is dynamically integrated into the keys and values as part of the computation process in attention modules. Therefore, it is advantageous to use a transformer-based model that supports relative positional encoding. This feature provides greater flexibility to the model and leads to more accurate result.

Table 5.7 Ablation study to assess the robustness of the image branch for unimodal configuration.

Image Encoder	$Accuracy$	$Precision$	$Recall$	$F1 - score$
ResNet-101[139]	72.10	72.34	72.30	72.31
EfficientNet-B7[140]	72.24	71.60	71.97	71.68
ResNext[141]	69.82	68.97	68.64	68.70
RegNet[142]	67.63	68.41	68.36	68.39
Vit-base-patch32 [131]	**84.75**	**84.61**	**84.43**	**84.52**

For the visual encoder branch of the proposed architecture, similar to the textual branch, various types of image encoder have been tested. Furthermore, from the experimentation of the ablation study depicted in **Table 5.7** and **Figure 5.11** it has been discovered that [41] outperforms at extracting the most salient elements from visual data. This is because a standalone transformer model [131] can effectively handle visual classification tasks by directly processing sequences of patches of images. Each patch is then converted into a vector via a Large Language Model referred as LLMs and processed using a transformer architecture. Hence, Dosovitskiy et al. [131] is capable of gathering and analysing global information in images, unlike ConvNets, which can only extract local aspects. Also, Dosovitskiy et al. [131] can be regarded as superior to different variants of ConvNet architectures, particularly in terms of scalability and the attention mechanism.



	ResNet-101	EfficientNet-B7	ResNext	RegNet	Vit-base-patch32
Accuracy	72.1	72.24	69.82	67.63	84.75
Precision	72.34	71.6	68.97	68.41	84.61
Recall	72.3	71.97	68.64	68.36	84.43
F1-score	72.31	71.68	68.7	68.39	84.52

Figure 5.11 A graphical representation for an ablation study that aims to evaluate the resilience of the image branch in a unimodal configuration.

Based on the experimental results of the ablation study, it has been determined that the configurations [132] and [131] yield the most optimal outcomes for the unimodal setup. Therefore, we have employed the fusion of these encoders in the proposed contrastive-based architecture for the ultimate prediction.

5.2.4 Conclusion and Future Directions

The main aim of this research is to introduce an innovative and novel framework called Contrastive Learning-Based Multimodal Architecture for the prediction of emoticons. This architecture is designed to work effectively with image-text pairs. Our suggested approach surpasses existing cutting-edge techniques in analysing the interplay between caption and visual modalities. The sentiment elicited by a phrase can exhibit variability in various scenarios, depending upon the context. Thus, it is essential to employ a blend of textual and visual information to attain more accurate prediction. Motivated by this, we developed a novel dual-branch architecture comprises of three primary components: Transformer-based visual encoder, Transformer-based textual encoder and an additional component involves the use of contrastive learning to uncover the hidden relationships within the text and images. In general, the model learns a multi-modal embedding space by the joint training of two encoders to maximize the cosine similarity between the picture and text embeddings of the N correct pairs in the batch, while minimizing the cosine similarity between the embeddings of the $N^2 - N$ incorrect pairings. A thorough analysis on one of the standard datasets referred as Multimodal-TwitterEmoticon shown that our suggested strategy outperforms strong baseline models.

Despite the excellent outcomes that have been achieved, there remain numerous opportunities for research. These includes creation of publically accessible dataset with an objective of mutli-label emoticon prediction, enhancing feature extraction methods, integrating adversarial learning capabilities to the fusion module, investigating a broader range of multimedia data, encompassing both video and acoustic formats, in order to potentially uncover more detailed semantic relations.

5.3 Significant Outcomes of this Chapter

The significant outcomes of this chapter are as follows:

- Introduced an innovative dual-branch Contrastive Learning-Based Multimodal Architecture to deal with the challenge of multimodal emoticon prediction.
- The proposed multimodal architecture comprises of three primary components: Transformer-based visual encoder, Transformer-based textual encoder and an additional component involves the use of contrastive learning to uncover the hidden relationships within the text and images. It has been demonstrated that by integrating the principle of contrastive learning with that of the other two branches yields superior results.
- We conduct thorough analyses and experimentation with Multimodal-TwitterEmoticon standard dataset to illustrate our model's ability to accurately and consistently model the multimodal representations of images and descriptive text. Moreover, we highlight how our model delivers ground-breaking outcomes in the realm of multimodal emoticon prediction.

The following research works form the basis of this chapter:

- ❖ **A. Pandey** and D. K. Vishwakarma, “Contrastive Learning-based Multi-Modal Architecture for Emoticon Prediction by Employing Image-Text Pairs.” Under Review in Cognitive Computation (Pub: Springer). https://doi.org/10.48550/arXiv.2408.02571.

Chapter 6: Sarcasm Detection by Employing Videos

6.1 Scope of this Chapter

This chapter specifically addresses the issue of identifying sarcasm, criticism, and symbolic information that is concealed inside regular conversations. Prior to this, the primary focus of sarcasm recognition was largely on written material. However, it is crucial to take into account all written material, audio stream, facial expression, and body position in order to accurately identify sarcasm. Therefore, we present a new method that integrates a low-complexity depth attention module with a self-regulated ConvNet to focus on the most important characteristics of visual data. Additionally, we employ an attentional tokenizer-based technique to extract the most significant context-specific information from the textual data obtained from subtitles. Thorough testing conducted on the MUStARD benchmark video datasets showed that the proposed approach for Multi-modal Sarcasm Recognition achieved the highest accuracy for both speaker-dependent and speaker-independent configurations. This indicates that the proposed approach is superior to existing methods. As part of a generalization study, we have performed a cross-dataset investigation to evaluate the adaptability of the proposed model using unseen samples from another dataset called MUStARD++.

6.2 VyAnG-Net: A Novel Multi-Modal Sarcasm Recognition Model by Uncovering Visual, Acoustic and Glossary Features

6.2.1 Abstract

Various linguistic and non-linguistic clues, such as excessive emphasis on a word, a shift in the tone of voice, or an awkward expression, frequently convey sarcasm. The computer vision problem of sarcasm recognition in conversation aims to identify hidden sarcastic, criticizing, and metaphorical information embedded in everyday dialogue. Prior, sarcasm recognition has focused mainly on text. Still, it is critical to consider all textual information, audio stream, facial expression, and body position for reliable sarcasm identification. Hence, we propose a novel approach that combines a lightweight depth attention module with a self-regulated ConvNet to concentrate on the most crucial features of visual data and an attentional tokenizer-based strategy to extract the most critical context-specific information from the textual data provided by subtitles. The following is a list of the key contributions that our experimentation has made in response to performing the task of Multi-modal Sarcasm Recognition: (1) an attentional tokenizer branch to get beneficial features from the glossary content provided by

the subtitles of the utterances; (2) a visual branch for acquiring the most prominent features from the video frames; (3) an utterance-level feature extraction from acoustic content and (4) a multi-headed attention based feature fusion branch to blend features obtained from multiple modalities and predict class label as sarcasm or non-sarcasm. Extensive testing on one of the benchmark video datasets, MUStARD, yielded an accuracy of **79.86%** for speaker dependent and **76.94%** for speaker independent configuration demonstrating that our approach is superior to the existing methods for Multi-modal Sarcasm Recognition. We have also conducted a cross-dataset analysis as part of a generalization study to test the adaptability of VyAnG-Net with unseen samples of another dataset MUStARD++.

6.2.2 Proposed Approach

This section provides a comprehensive discussion of the proposed framework VyAnG-Net. The main objective is outlined in the first part of this section. Then, the model's framework is introduced, which consists of three modules: a glossary branch that uses the attention-based tokenization approach to acquire the most significant contextual features from the textual content provided by the subtitles of the video utterances, a visual branch with dedicated attention module to acquire the most prominent features from the video frames and lastly multi-headed attention based feature fusion to blend features acquired from each of the separate modalities. The term “multi-modal sarcasm recognition” is typically used in our study to denote the process of analysing sarcasm in video utterances for the sake of convenience. The proposed model has been named "VyAnG," which draws inspiration from the use of sarcasm in the Hindi language. The acronyms V, A, and G correspond to visual, acoustic, and glossary (textual) content, respectively.

Problem Definition

The problem of recognising sarcasm in video utterances can be thoroughly summed up as follows:

Let "$\mathcal{V}$" denote the sample space comprising video utterances. A sample of the dataset consists of the textual content conveyed through video subtitles "$\mathbb{G}$", visual frames "$\mathbb{V}$", and accompanying acoustic content "$\mathbb{A}$". Each of the samples is assigned to a class label denoted as "$\mathbb{C}$". In more technical terms, each sample can be defined as a quartet consisting of a subtitle, visual frames, acoustic information, and a class label. The following expression can be formulated as:

$$\mathbb{I} = \left\{(\mathbb{G}^{\mathbf{0}}, \mathbb{V}^{\mathbf{0}}, \mathbb{A}^{\mathbf{0}}, \mathbb{C}^{\mathbf{0}}), .., \left(\mathbb{G}^{\boldsymbol{i}}, \mathbb{V}^{\boldsymbol{i}}, \mathbb{A}^{\boldsymbol{i}}, \mathbb{C}^{\boldsymbol{i}}\right), .., (\mathbb{G}^{\boldsymbol{m-1}}, \mathbb{V}^{\boldsymbol{m-1}}, \mathbb{A}^{\boldsymbol{m-1}}, \mathbb{C}^{\boldsymbol{m-1}})\right\} \qquad (\mathbf{6.1})$$

where, $\mathbb{I}$ is the collection of sample quartets, $\mathbb{G}^i$ denotes subtitle information, $\mathbb{V}^i$ denotes visual frame information, $\mathbb{A}^i$ defines the acoustic content, $\mathbb{C}^i$ is the class label that corresponds to a specific utterance for the i^{th} sample and the variable m represents the cardinality of the sample space, which denotes the total number of samples in a given dataset. VyAnG-Net aims to learn a mapping function $\mathbb{F}: (\mathbb{G}, \mathbb{V}, \mathbb{A}) \longrightarrow \mathbb{C}$ from the multi-modal training examples $\left\{\left(\mathbb{G}^{\boldsymbol{i}}, \mathbb{V}^{\boldsymbol{i}}, \mathbb{A}^{\boldsymbol{i}}\right) \middle| \mathbf{0} \leq \boldsymbol{i} \leq \boldsymbol{m} - \mathbf{1}\right\}$. For a sarcasm recognition task, $\mathbb{C}^{\boldsymbol{i}} \in \{\boldsymbol{sarcasm\ and\ not\ sarcasm}\}$. Consider the cheering statement in **Figure 6.1**: *"if you're compiling a mix CD for a double suicide. Oh, I hope that scratching post is for you."* becomes sarcastic when spoken with an awkward face and a saucy tone, and in general, has a negative meaning. Naturally, humans can process this massive amount of simultaneous data. However, developing an approach that can possibly accomplish the same task requires a suitable representation of all of these different sources of information. It thus results in a significant increase in research interest.

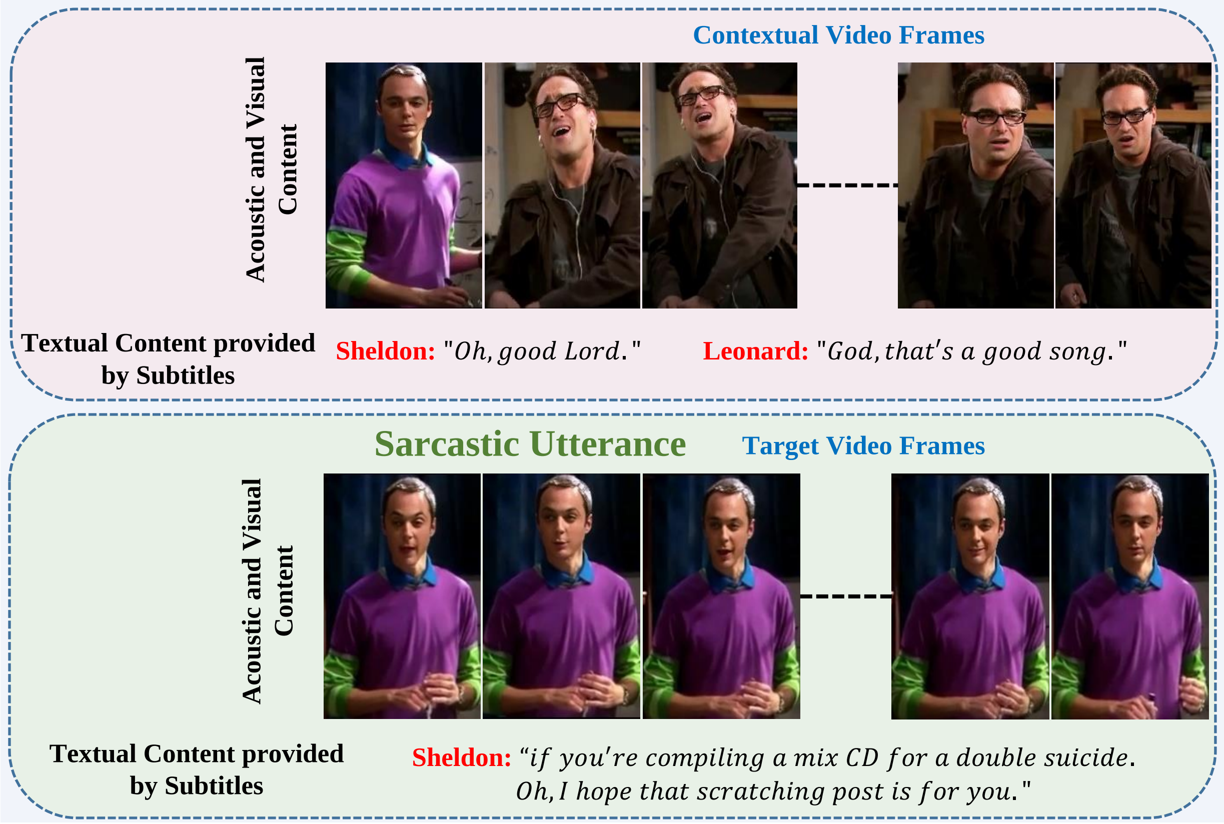

Contextual Video Frames
Acoustic and Visual Content
Textual Content provided by Subtitles
Sheldon: "Oh, good Lord."
Leonard: "God, that's a good song."
Sarcastic Utterance
Target Video Frames
Acoustic and Visual Content
Textual Content provided by Subtitles
Sheldon: "if you're compiling a mix CD for a double suicide. Oh, I hope that scratching post is for you."

Figure 6.1 A perfect illustration of a sarcastic statement from the dataset, accompanied by its context and a transcript of the utterance.

VyAnG-Net: A Novel Multi-Modal Sarcasm Recognition Model

The VyAnG-Net was proposed for the purpose of performing multi-modal sarcasm recognition to generate the relationship among visual, acoustic, and glossary (textual) information and to explore the compatibility between these three modalities. **Figure 6.2** illustrates the VyAnG-Net framework.

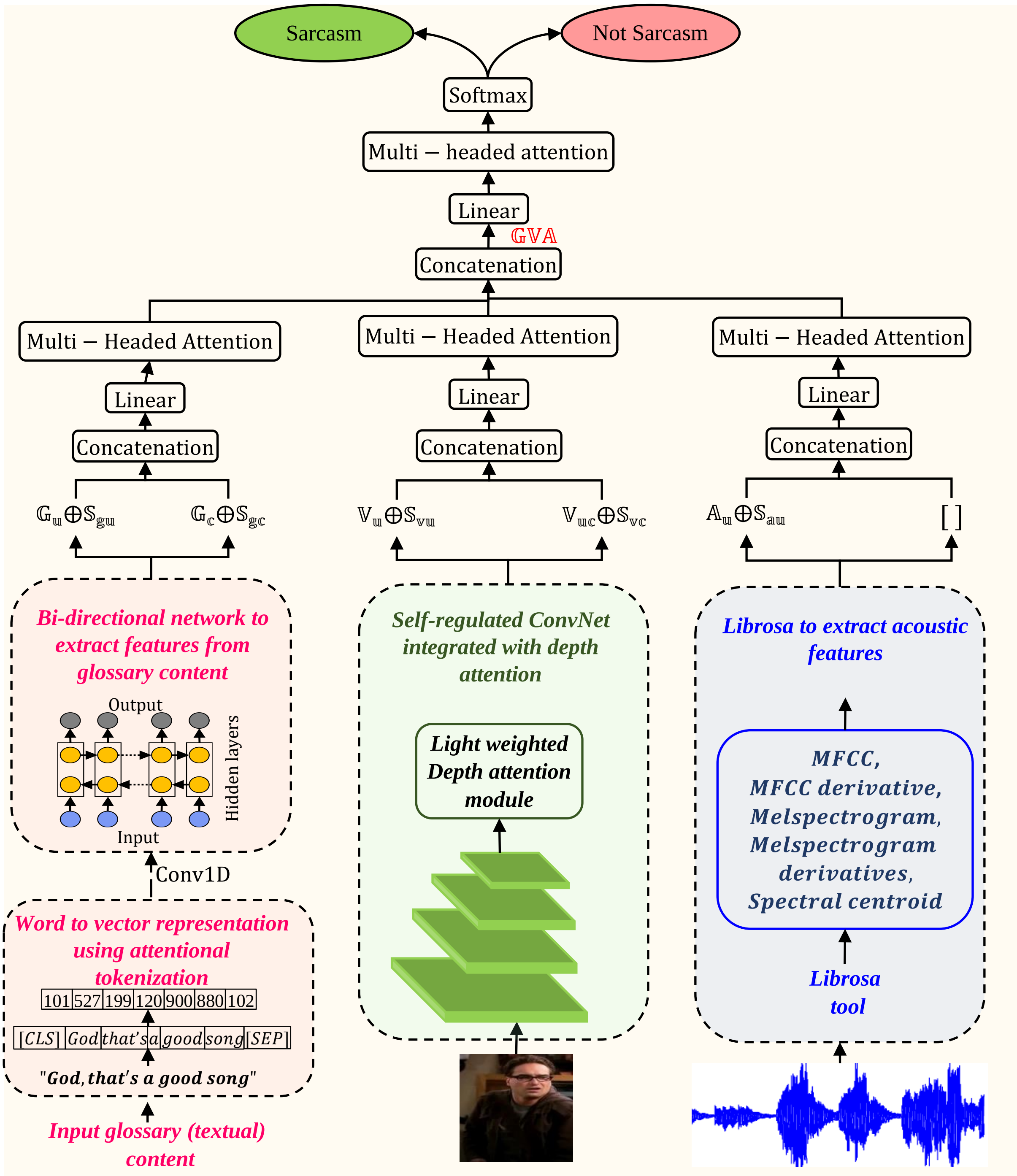

Sarcasm
Not Sarcasm
Softmax
Multi – headed attention
Linear
$\mathbb{GVA}$
Concatenation
Multi – Headed Attention
Linear
Concatenation
$\mathbb{G}_u \oplus \mathbb{S}_{gu}$
$\mathbb{G}_c \oplus \mathbb{S}_{gc}$
Multi – Headed Attention
Linear
Concatenation
$\mathbb{V}_u \oplus \mathbb{S}_{vu}$
$\mathbb{V}_{uc} \oplus \mathbb{S}_{vc}$
Multi – Headed Attention
Linear
Concatenation
$\mathbb{A}_u \oplus \mathbb{S}_{au}$
[ ]
Bi-directional network to extract features from glossary content
Output
Hidden layers
Input
Conv1D
Word to vector representation using attentional tokenization
101 527 199 120 900 880 102
[CLS] God that's a good song [SEP]
"God, that's a good song"
Input glossary (textual) content
Self-regulated ConvNet integrated with depth attention
Light weighted Depth attention module
Librosa to extract acoustic features
MFCC, MFCC derivative, Melspectrogram, Melspectrogram derivatives, Spectral centroid
Librosa tool

Figure 6.2 Proposed VyAnG-Net Framework where, $\mathbb{G}, \mathbb{V}, \mathbb{A}$ corresponds to glossary (textual), visual, and acoustic content. The notations $\mathbb{G}_u \oplus \mathbb{S}_{gu}$, $\mathbb{V}_u \oplus \mathbb{S}_{vu}$, and $\mathbb{A}_u \oplus \mathbb{S}_{au}$ refers to the utterance level features for all the three modalities. Additionally, $\mathbb{G}_c \oplus \mathbb{S}_{gc}$, $\mathbb{V}_{uc} \oplus \mathbb{S}_{vc}$ represents the context level feature for glossary and visual content. Lastly, [ ] denotes the empty list.

The model comprises of three distinct components. Firstly, a textual branch that employs an attention-based tokenization approach to extract the most salient contextual features from the glossary content presented in the video utterances' subtitles. Secondly, a visual branch that incorporates a dedicated attention module to capture the most prominent features from the video frames. Lastly, a multi-headed attention-based feature fusion mechanism is utilised to integrate the features obtained from each of the individual modalities. **Table 6.1** presents the proposed framework in algorithmic format.

Table 6.1 Pseudocode for the proposed VyAnG-Net.

Algorithm 1: VyAnG-Net: A Novel Multi-Modal Sarcasm Recognition Model by Uncovering Visual, Acoustic and Glossary Features.

Aim: To learn a mapping function $\mathbb{F}: (\mathbb{G}, \mathbb{V}, \mathbb{A}) \longrightarrow \mathbb{C}$ from the multi-modal training examples $\{(\mathbb{G}^i, \mathbb{V}^i, \mathbb{A}^i) | 0 \leq i \leq m - 1\}$.
Input: Glossary (textual) set $\mathbb{G} = \{\mathbb{G}_1, \mathbb{G}_2, \dots \dots, \mathbb{G}_i\}$, visual set $\mathbb{V} = \{\mathbb{V}_1, \mathbb{V}_2, \dots \dots, \mathbb{V}_i\}$, and acoustic set $\mathbb{A} = \{\mathbb{A}_1, \mathbb{A}_2, \dots \dots, \mathbb{A}_i\}$.
Output: sarcasm recognition task, $\mathbb{C}^{\boldsymbol{i}} \in \{sarcasm\ and\ not\ sarcasm\}$.

```
1.  Word-to-vector representation from the entire Glossary content set $\mathbb{R}$;
2.  Extract features at the level of utterance and context from vector representation of the glossary content
    $\mathbb{G}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{gu}}$, and $\mathbb{G}_{\mathbb{c}} \oplus \mathbb{S}_{\mathbb{gc}}$
3.  Extract features at the level of utterance and context from the visual frame $\mathbb{V}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{vu}}$, and $\mathbb{V}_{\mathbb{uc}} \oplus \mathbb{S}_{\mathbb{vc}}$;
4.  Extract features at the level of utterance and context from acoustic content $\mathbb{A}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{au}}$, and [ ]
5.  for $\mathbb{E} \leftarrow 1$ to $\mathbb{Epochs}$ do
            $\mathbb{R}_{\mathbb{u}} \leftarrow \mathbb{W}_{1:\mathbb{g}} = \{\mathbb{W}_1, \mathbb{W}_2, \dots \dots, \mathbb{W}_{\mathbb{g}}\}$ word to vector representation by Eqn. $(\mathbf{6.2})$;
            $\mathbb{R}_{\mathbb{c}} \leftarrow \mathbb{W}_{1:\mathbb{g}_{\mathbb{c}}^{\mathbb{i}}} = \{\mathbb{W}_1^{\mathbb{i}}, \mathbb{W}_2^{\mathbb{i}}, \dots \dots, \mathbb{W}_{\mathbb{g}_{\mathbb{c}}}^{\mathbb{i}}\}$ word to vector representation by Eqn.$(\mathbf{6.2})$ ;
            $\mathbb{G}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{gu}} \leftarrow (\mathbb{G}_{\mathbb{u}} \leftarrow \boldsymbol{\mu}(\mathbb{R}_{\mathbb{u}}) \oplus \mathbb{S}_{\mathbb{gu}})$ obtain utterance-level text-based features using
            Eqn.$(\mathbf{6.4})$;
            $\mathbb{G}_{\mathbb{c}} \oplus \mathbb{S}_{\mathbb{gc}} \leftarrow (\mathbb{G}_{\mathbb{c}}^{\mathbb{i}} \oplus \mathbb{S}_{\mathbb{gc}}^{\mathbb{i}})$ obtain context-level text-based features using Eqn.  and $(\mathbf{6.6})$;
            $\mathbb{V}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{vu}} \leftarrow \mathbb{V}\mathbb{S}_{\mathbb{u}}$ obtain utterance-level visual features using Eqn. $(\mathbf{6.10})$ and $(\mathbf{6.11})$;
            $\mathbb{V}_{\mathbb{uc}} \oplus \mathbb{S}_{\mathbb{vc}} \leftarrow \mathbb{V}\mathbb{S}_{\mathbb{c}}$ obtain context-level visual features using Eqn. $(\mathbf{6.12})$ and $(\mathbf{6.13})$;
            $\mathbb{A}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{au}} \leftarrow$ Librosa tool$(\boldsymbol{Concatenation}(\mathbb{A}_{\mathbb{u}}, \mathbb{S}_{\mathbb{au}}))$ obtain utterance-level acoustic
            features using Eqn. $(\mathbf{6.14})$;
            $\mathbb{G}_{cat} \leftarrow \mathbb{G}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{gu}} \oplus \mathbb{G}_{\mathbb{c}} \oplus \mathbb{S}_{\mathbb{gc}}$ final text-based features obtained by concatenating utterance
            and context-based features;
            $\mathbb{V}_{cat} \leftarrow \mathbb{V}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{vu}} \oplus \mathbb{V}_{\mathbb{uc}} \oplus \mathbb{S}_{\mathbb{vc}}$ final vision-based features obtained by concatenating
            utterance and context-based features;
            $\mathbb{A}_{cat} \leftarrow \mathbb{A}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{au}} \oplus$ [ ] final audio-based features obtained by concatenating utterance and
            context-based features;
            $\mathbb{L}_{\mathbb{G}} \leftarrow \mathbb{G}_{cat}$ obtain the most prominent textual features by applying multi-headed attention
            using Eqn. $(\mathbf{6.16})$;
            $\mathbb{L}_{\mathbb{V}} \leftarrow \mathbb{V}_{cat}$ obtain the most prominent visual features by applying multi-headed attention
            using Eqn. $(\mathbf{6.16})$;
            $\mathbb{L}_{\mathbb{A}} \leftarrow \mathbb{A}_{cat}$ obtain the most prominent acoustic features by applying multi-headed attention
            using Eqn.$(\mathbf{6.16})$
            $\mathbb{G}\mathbb{V}\mathbb{A} \leftarrow \mathbb{L}_{\mathbb{G}} \oplus \mathbb{L}_{\mathbb{V}} \oplus \mathbb{L}_{\mathbb{A}}$ concatenate all the features obtained from multiple modalities to get
            multi-modal feature representation using Eqn.$(\mathbf{6.17})$
            $\mathbb{C} \leftarrow \boldsymbol{Softmax}(\mathbb{G}\mathbb{V}\mathbb{A})$ pass the multi-modal features to the softmax layer to get the final
            prediction using Eqn. $(\mathbf{6.18})$;
            calculate loss and perform backpropagation;
6.  $End$
```

'⊕' denotes concatenation & '[ ]'denotes the empty list

Input Features

The dataset comprises of individual samples that include an utterance, its corresponding context, and associated labels. The utterance's context encompasses a series of prior utterances, typically $\mathbb{N}$ in number, that lead up to the given utterance within the dialogue. Each utterance is linked to its respective context and speaker, with the speaker of the utterance and the speaker of the context being distinct entities. Our study provides an extensive explanation of the utterance and its contextual factors across all modalities in the following subsections.

Textual Feature Extraction using Glossary Content

Assuming a given utterance consisting of $\mathbb{g}$ words, denoted as $\mathbb{W}_{1:\mathbb{g}} = \{\mathbb{W}_1, \mathbb{W}_2, \ldots \ldots, \mathbb{W}_{\mathbb{g}}\}$, where each word $\mathbb{W}_i$ belongs to the set of real numbers $\mathfrak{R}^{300}$. Each term is denoted as $\mathbb{W}_i$, corresponds to a vector that is generated through the utilization of [143] attention-based tokenization $(\boldsymbol{\tau})$ represented in **Eqn.**$(\mathbf{6.2})$. The acquisition of the contextual relationship among words is accomplished by means of employing a [144] model denoted as $\boldsymbol{\mu}$ using **Eqn.** $(\mathbf{6.3})$. Subsequently, utterance level features are obtained through the utilization of the final word embedding, represented as $\mathbb{G}_{\mathbb{u}}$.

$$\mathbb{R}_{\mathbb{u}} = \boldsymbol{\tau}\left(\{\mathbb{W}_1, \mathbb{W}_2, \ldots \ldots, \mathbb{W}_{\mathbb{g}}\}\right) \qquad (\mathbf{6.2})$$

$$\mathbb{G}_{\mathbb{u}} = \boldsymbol{\mu}(\mathbb{R}_{\mathbb{u}}) \qquad (\mathbf{6.3})$$

In cases where speaker information $\mathbb{S}_{\mathbb{g}\mathbb{u}}$is accessible, it is possible to combine it using **Eqn.**$(\mathbf{6.4})$ with $\mathbb{G}_{\mathbb{u}}$ to form a speaker-aware textual utterance, which is represented $\mathbb{G}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{g}\mathbb{u}}$.

$$\mathbb{G}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{g}\mathbb{u}} = \boldsymbol{Concatenation}\left(\mathbb{G}_{\mathbb{u}}, \mathbb{S}_{\mathbb{g}\mathbb{u}}\right) \qquad (\mathbf{6.4})$$

Assuming there is a set of utterances in the given context, each comprising $\mathbb{g}_{\mathbb{c}}$ words, the utterance-level representations for such a set of contextual videos are obtained by subjecting the words of each utterance to [144], using **Eqn.** $(\mathbf{6.5})$ following which the embedding of the last word of the glossary provided by the subtitle is utilized. The $\mathbb{i}^{\mathbb{th}}$ utterance in the context is denoted by $\mathbb{G}_{\mathbb{c}}^{\mathbb{i}}$.

$$\mathbb{G}_{\mathbb{c}}^{\mathbb{i}} = \boldsymbol{\mu}\left(\boldsymbol{\tau}\left(\left\{\mathbb{W}_{\mathbf{1}:\mathbb{g}_{\mathbb{c}}^{\mathbb{i}}} = \{\mathbb{W}_{\mathbf{1}}^{\mathbb{i}}, \mathbb{W}_{\mathbf{2}}^{\mathbb{i}}, \ldots \ldots, \mathbb{W}_{\mathbb{g}_{\mathbb{c}}}^{\mathbb{i}}\}\right\}\right)\right) \qquad (\mathbf{6.5})$$

When speaker information $\mathbb{S}_{\mathbb{g}\mathbb{c}}^{\mathbb{i}}$ is available, it is appended to every contextual utterance $\mathbb{G}_{\mathbb{c}}^{\mathbb{i}}$ too. In the end, the features at the context level are also obtained by concatenating all textual utterances that are influenced by the speaker, as represented in **Eqn.**$(\mathbf{6.5})$.

$$\mathbb{G}_{\mathbb{c}} \oplus \mathbb{S}_{\mathbb{gc}} = Concatenation\left(\left(\mathbb{G}_{\mathbb{c}}^{1} \oplus \mathbb{S}_{\mathbb{gc}}^{1}\right), \left(\mathbb{G}_{\mathbb{c}}^{2} \oplus \mathbb{S}_{\mathbb{c}}^{2}\right), \dots \dots \dots, \left(\mathbb{G}_{\mathbb{c}}^{\mathbb{i}} \oplus \mathbb{S}_{\mathbb{gc}}^{\mathbb{i}}\right)\right) \qquad (\mathbf{6.6})$$

Visual Feature Extraction from the Video Frames of the Utterances

To obtain visual features from the video frames [145] is integrated with the depth attention module [146], discussed in the following section.

Lightweight Attention Module

The Convolutional Neural Networks (ConvNets) have demonstrated remarkable representational abilities, leading to significant enhancements in their efficacy for visual tasks. In addition, we explore another aspect of architectural design that has become increasingly prevalent in modern times, namely, attention. Through the utilization of attention mechanisms, which involve prioritising significant attributes while inhibiting irrelevant ones, it is anticipated that the efficacy of representation will be enhanced. Considering this information, a framework known as the "light weighted attention framework" [146], illustrated in **Figure 6.3**, has been developed and integrated into [145] to concentrate on the most salient characteristics from the visual frames while disregarding the others. In order to accomplish this task, we have implemented four different modules, namely, feature grouping, depth attention, spatial attention and aggregation, which constitute [146].

The word "spatial" refers to the encompassing spatial domain of each feature map. By including the spatial attention module to enhance the feature maps, the superior input is then sent to the subsequent levels of convolution, hence increasing the efficacy of the model**.** On the other hand, the phrase "depth" denotes the total number of channels, which are simply a set of feature maps arranged in a tensor. Each and every multidimensional layer inside this tensor represents a feature map with a depth of $\mathbb{H} \times \mathbb{W}$. The depth attention mechanism provides a numerical value associated with every channel, therefore prioritising those channels that have the most impact on the learning process. This prioritisation leads to the optimisation of the most important features, ultimately enhancing the overall performance of the model. The property of feature grouping is characterised by a hierarchical structure consisting of two levels. Suppose that the attention module's input tensor is $\mathbb{X} \in \mathbb{R}^{\mathbb{D} \times \mathbb{H} \times \mathbb{W}}$, where $\mathbb{D}, \mathbb{H}\ and\ \mathbb{W}$ denote the depth, height and width of the feature map, respectively. Initially, $\mathbb{X}$ is partitioned into $\mathbb{P}$ distinct groups, resulting in $\mathbb{X}' \in \mathbb{R}^{\frac{\mathbb{D}}{\mathbb{P}} \times \mathbb{H} \times \mathbb{W}}$ for each group across the depth of the feature maps. The obtained feature groups are then transmitted to the attention components, where they are subsequently segregated into two distinct groups based on the depth dimension. One group is allocated to the spatial attention branch, while the other is assigned to the depth attention branch. And these sub-feature groups that are transmitted across both the spatial or depth attention branches can be represented as $\mathbb{X}^{"} \in \mathbb{R}^{\frac{\mathbb{D}}{2\mathbb{P}}\times\mathbb{H}\times\mathbb{W}}$. The depth attention branch involves reducing the feature maps obtained from the feature grouping phase to $\mathbb{X}^{"} \in \mathbb{R}^{\frac{\mathbb{D}}{2\mathbb{P}}\times\mathbf{1}\times\mathbf{1}}$. This is achieved through the use of a global average pooling operation and gating mechanism, which enables more accurate and versatile decisions. The resulting output is then subjected to a sigmoid activation function which is represented as follows:

$$\boldsymbol{X}^{\wedge}_{\mathbb{k}\mathbf{1}} = \boldsymbol{\sigma}\left(\mathbb{F}_{\mathbb{c}}(\mathbb{t})\right) \cdot \mathbb{X}^{"} = \boldsymbol{\sigma}(\mathbb{V}_{\mathbf{1}\mathbb{t}} \oplus \mathbb{b}_{\mathbf{1}}) \cdot \mathbb{X}^{"} \qquad (\mathbf{6.7})$$

The Group Norm technique is employed to reduce the input $\mathbb{X}'$ in spatial attention, resulting in spatial features. The function $\mathbb{F}(.)$ is subsequently employed to improve the depiction of the diminished tensor. This concept can be expressed through a simple mathematical formula:

$$X^{\wedge}_{\mathbb{k}2} = \sigma\left(\mathbb{V}_2 \cdot GroupNorm\left(\mathbb{X}^{"}\right) \oplus \mathbb{b}_2\right) \cdot \mathbb{X}^{"} \qquad (\mathbf{6.8})$$

The concatenation of the outputs obtained from the Spatial Attention and depth attention is performed initially. Then a depth shuffle technique is implemented, similar to the approach used in ShuffleNet, to facilitate interaction among groups along the depth. Consequently, the resulting output possesses identical dimensions to those of the input tensor that underwent processing in the shuffle attention layer.

$$X^{\wedge}_{\mathbb{k}} = \left[X^{\wedge}_{\mathbb{k}1} \oplus X^{\wedge}_{\mathbb{k}2}\right] \in \mathbb{R}^{\frac{\mathbb{D}}{\mathbb{P}}\times\mathbb{H}\times\mathbb{W}} \qquad (\mathbf{6.9})$$

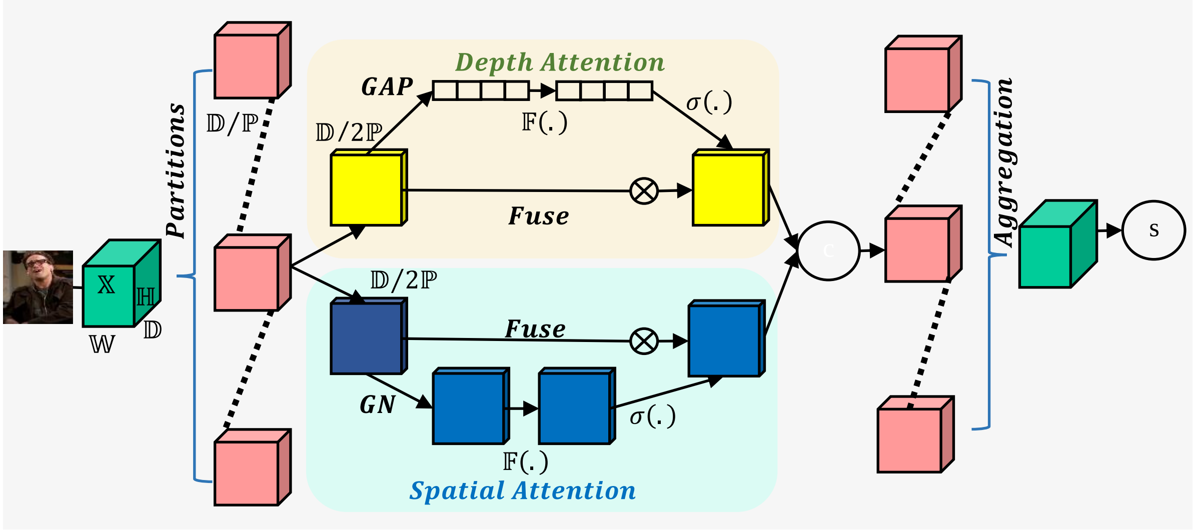


Figure 6.3 Light weighted depth attention module where GAP represents global average pooling operation, GN represents group normalization, “C represents concatenation and “S” represents depth shuffle operation.

Utterance and Context-Level Feature Extraction from Video Frames

In this, utterance-level features are initially extracted, followed by the extraction of context-level features. These two sets of features are then subsequently concatenated to yield the final feature representation. Suppose there is a set of $\mathbb{N}_{\mathbb{u}}$ visual frames at utterance level denoted as $\mathbb{V}_{1:\mathbb{N}_{\mathbb{u}}} = \{\mathbb{V}_1, \mathbb{V}_2, \dots\dots, \mathbb{V}_{\mathbb{N}_{\mathbb{u}}}\}$. Each visual frame is sent to a self-regulatory ConvNet model [145] depicted in **Figure 6.4** that makes use of a light-weighted depth attention module [146] to extract the most prominent features, which has already been explained in detail above. To obtain information pertaining to the level of utterance, the mean value is computed for all frames $\mathbb{V}_{\mathbb{u}}$. In cases where speaker information is present, the utterance $\mathbb{V}_{\mathbb{u}}$ is concatenated with the corresponding speaker information $\mathbb{S}_{\mathbb{vu}}$ given by **Eqn.** $(\mathbf{6.10})$ and **Eqn.**$(\mathbf{6.11})$. The notation $\mathbb{V}_{\mathbb{u}}\oplus\mathbb{S}_{\mathbb{vu}}$ used is where $\mathbb{V}_{\mathbb{u}}$ belongs to the set of real numbers $\mathfrak{R}$ and has a cardinality of 2048.

$$\mathbb{VS}_{\mathbb{u}} = \left((\mathbb{V}_{\mathbb{u}1}\oplus\mathbb{S}_{\mathbb{vu}1}), (\mathbb{V}_{\mathbb{u}2}\oplus\mathbb{S}_{\mathbb{vu}2}), \dots\dots.., (\mathbb{V}_{\mathbb{uN}}\oplus\mathbb{S}_{\mathbb{vuN}})\right) \qquad (\mathbf{6.10})$$

$$\mathbb{V}_{\mathbb{u}}\oplus\mathbb{S}_{\mathbb{vu}} = \boldsymbol{Self - regulated\ ConvNet\oplus Lightweighted\ depth\ attention}(\mathbb{VS}_{\mathbb{u}}) \qquad (\mathbf{6.11})$$

Similarly, to obtain context-level features from the set of $\mathbb{N}_{\mathbb{uc}}$ contextual utterances, the mean value is computed for all frames denoted as $\mathbb{V}_{\mathbb{uc}}$. In cases where speaker information is present, the utterance $\mathbb{V}_{\mathbb{uc}}$ is concatenated with the corresponding speaker information $\mathbb{S}_{\mathbb{vc}}$ defined by **Eqn.**$(\mathbf{6.12})$ and **Eqn.**$(\mathbf{6.13})$.

$$\mathbb{VS}_{\mathbb{c}} = \left((\mathbb{V}_{\mathbb{uc}\mathbf{1}}\oplus\mathbb{S}_{\mathbb{vc}\mathbf{1}}), (\mathbb{V}_{\mathbb{uc}\mathbf{2}}\oplus\mathbb{S}_{\mathbb{vc}\mathbf{2}}), \dots\dots.., (\mathbb{V}_{\mathbb{ucN}}\oplus\mathbb{S}_{\mathbb{vcN}})\right) \qquad (\mathbf{6.12})$$

$$\mathbb{V}_{\mathbb{uc}}\oplus\mathbb{S}_{\mathbb{vc}} = \boldsymbol{Self - regulated\ ConvNet\oplus Lightweighted\ depth\ attention}(\mathbb{VS}) \qquad (\mathbf{6.13})$$

Utterance-Level Feature Extraction from Acoustic Content

Suppose there is a set of $\mathbb{N}_{\mathbb{a}}$ acoustic frames at utterance level denoted as $\mathbb{A}_{1:\mathbb{N}_{\mathbb{a}}} = \{\mathbb{A}_1, \mathbb{A}_2, \dots\dots, \mathbb{A}_{\mathbb{N}_{\mathbb{a}}}\}$. Librosa library has been utilized to extract acoustic information. Similar to the process of extracting visual features, the methodology utilised here also involves the computation of the average value of all frames to extract information pertaining to utterances denoted as $\mathbb{A}_{\mathbb{u}}$. In cases where speaker information is present, the utterance $\mathbb{A}_{\mathbb{u}}$ is concatenated with the corresponding speaker information $\mathbb{S}_{\mathbb{au}}$ given by **Eqn.**$(\mathbf{6.14})$. The notation $\mathbb{A}_{\mathbb{u}}\oplus\mathbb{S}_{\mathbb{au}}$ used is where $\mathbb{A}_{\mathbb{u}}$ belongs to the set of real numbers $\mathfrak{R}$ and has a cardinality of 283.

Also, it is important to keep in mind that audio recordings often consist of a variety of speakers, ambient noise, cues for laughter, and other sounds in the background. As a result, the consideration of contextual factors is not integrated into acoustic analysis, as it might cause challenges in distinguishing it from the laughter portion of the conversation. Therefore, an empty list [ ] is employed within the context of acoustic content.

$$\mathbb{A}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{au}} = \textbf{Librosa tool}\left(\boldsymbol{Concatenation}(\mathbb{A}_{\mathbb{u}}, \mathbb{S}_{\mathbb{au}})\right) \qquad (\mathbf{6.14})$$

Multi-Headed Attention-Based Feature Fusion

In the very first step, all the utterance and context-level features of all the modalities are concatenated together, as represented in **Eqn.**$(\mathbf{6.15})$.

$$\begin{aligned}
\mathbb{G}_{cat} &= \boldsymbol{Concatenation}\left((\mathbb{G}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{gu}}), (\mathbb{G}_{\mathbb{c}} \oplus \mathbb{S}_{\mathbb{gc}})\right)\\
\mathbb{V}_{cat} &= \boldsymbol{Concatenation}\left((\mathbb{V}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{vu}}), (\ \mathbb{V}_{\mathbb{uc}} \oplus \mathbb{S}_{\mathbb{vc}})\right)\\
\mathbb{A}_{cat} &= \boldsymbol{Concatenation}\left((\mathbb{A}_{\mathbb{u}} \oplus \mathbb{S}_{\mathbb{au}}), ([\ ])\right) \qquad (\mathbf{6.15})
\end{aligned}$$

Subsequently, $\mathbb{G}_{cat}$, $\mathbb{V}_{cat}$, and $\mathbb{A}_{cat}$ are individually fed into the linear layer followed by the multi-headed attention layer as defined by **Eqn.** $(\mathbf{6.16})$

$$\begin{aligned}
\mathbb{L}_{\mathbb{G}} &= \boldsymbol{Multi-headed\ attention}\left(\boldsymbol{Linear}(\mathbb{G}_{cat})\right)\\
\mathbb{L}_{\mathbb{V}} &= \boldsymbol{Multi-headed\ attention}\left(\boldsymbol{Linear}(\mathbb{V}_{cat})\right)\\
\mathbb{L}_{\mathbb{A}} &= \boldsymbol{Multi-headed\ attention}\left(\boldsymbol{Linear}(\mathbb{A}_{cat})\right) \qquad (\mathbf{6.16})
\end{aligned}$$

The features derived from various modalities are concatenated and subsequently fed into a linear layer, which is succeeded by a multi-headed attention layer. This process yields a highly significant multi-modal feature vector, as outlined in **Eqn.**$(\mathbf{6.17})$.

$$\begin{aligned}
&\mathbb{GVA} = \boldsymbol{Multi-}\\
&\boldsymbol{headed\ attention}\left(\boldsymbol{Linear}\left(\boldsymbol{Concatenation}(\mathbb{L}_{\mathbb{G}}, \mathbb{L}_{\mathbb{V}}, \mathbb{L}_{\mathbb{A}})\right)\right) \qquad (\mathbf{6.17})
\end{aligned}$$

Finally, the softmax layer is utilized to forecast the classification label as either sarcastic or non-sarcastic, as illustrated in **Eqn.**$(\mathbf{6.18})$.

$$\mathbb{C} = \boldsymbol{Softmax}(\mathbb{GVA}) \qquad (\mathbf{6.18})$$

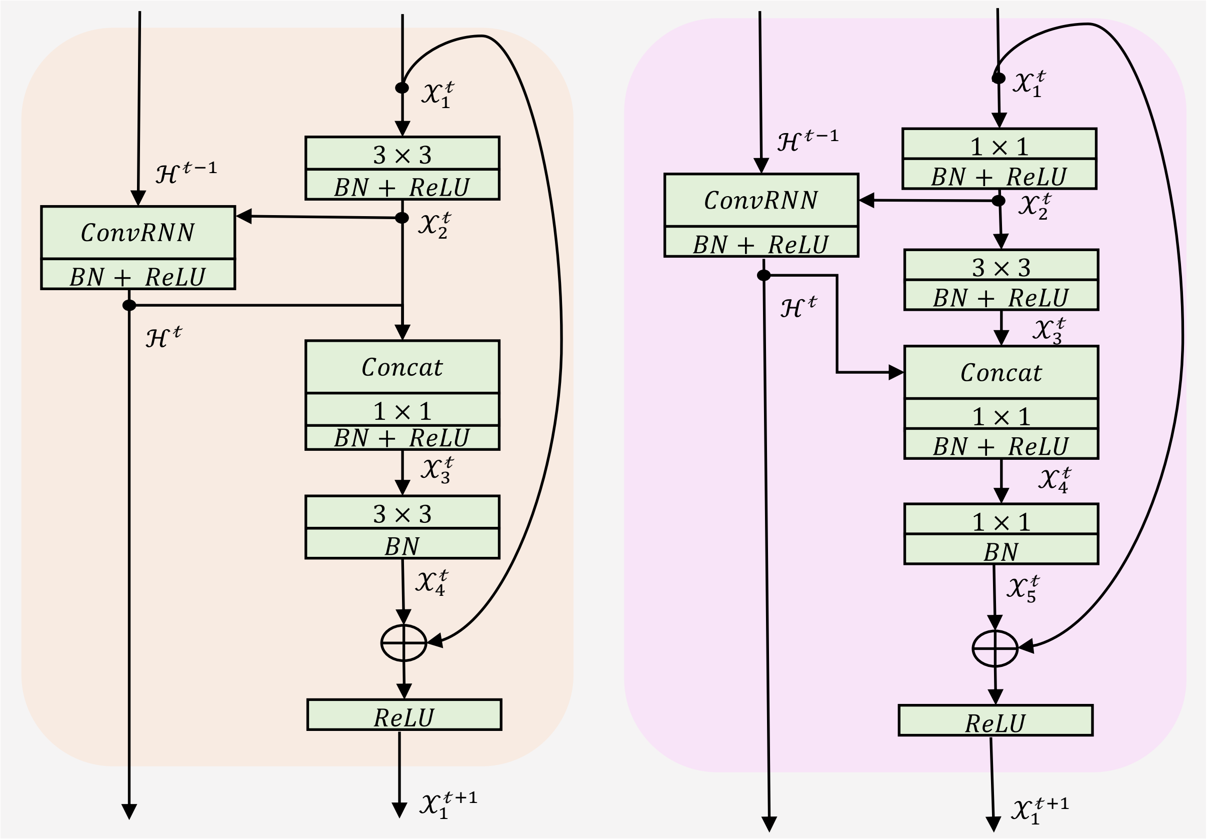


Figure 6.4 Self-regulated ConvNet in which $\mathcal{H}$ refers to the hidden states, $\mathcal{X}$ refers to the input feature map, and t denotes the number of building blocks

6.2.3 Experiment Setup and Result

The following subsection covers comprehensive details related to the dataset used throughout the study, the experimental configurations of the proposed methodology, and evaluations of its performance.

Dataset Used

The MUStARD dataset, as provided by [147], is utilized for the purpose of multimodal sarcasm recognition. This dataset comprises a total of 690 utterances, with 345 being sarcastic and 345 being non-sarcastic. The data was gathered from various well-known television series, including Sarcasmaholics Anonymous, Friends, and The Golden Girls. Similar to previous research, we analyse our proposed framework in two distinct experimental setups.

One of the scenarios is the "speaker independent" configuration. The present study employs utterances obtained from the Friends Series as testing data, while the remaining utterances are utilized as training data. The alternative setup is the "speaker dependent", wherein the dataset is partitioned into five folds. During each of the five iterations, the ith fold is designated as the testing set, while the rest of the sample folds are utilized for training purposes. Subsequently, five datasets can be acquired.

In addition to the MUStARD dataset, we have also used the extended version MUStARD++ [36] to execute a cross-dataset study as part of a generalization research to test the resilience of VyAnG-Net. Our proposed approach, VyAnG-Net, was trained using the MUStARD dataset for this experimental investigation, and its performance was evaluated using an unseen MUStARD++ dataset. Consistent with prior research, we utilize $\mathbb{Accuracy}(\mathrm{A})$, $\mathbb{Precision}(\mathrm{P})$, $\mathbb{Recall}(\mathrm{R})$, and $\mathbb{F}1\ \mathbb{Score}(\mathrm{F1})$ as the metrics for evaluation. Regarding the "speaker dependent" configuration, the outcomes are presented by computing the mean of the results obtained from five distinct evaluation sets.

Experimental Setup

The proposed method was executed by utilising the Keras and PyTorch framework. The evaluation metrics employed for recognising sarcasm are $\mathbb{Accuracy}(A)$, $\mathbb{Precision}(P)$, $\mathbb{Recall}(R)$, an $\mathbb{F}1\ \mathbb{Score}(F1)$. Concerning all experimental procedures, the architectures employed in this study incorporate, Rectified Linear Unit (ReLU) as an activation function, a dropout rate of 0.4, the Adam optimisation algorithm with a learning rate of 0.001, and a batch size of 32. The training process of the model was conducted for a total of 200 epochs because the results were not satisfactory beyond this limit i.e., we pertain to a limit of 200 epochs. The Adam optimizer is employed in conjunction with Softmax as a classifier to identify sarcasm. Furthermore, the implementation of the sigmoid activation function and the optimisation of binary cross-entropy as the loss function were utilized. The model we have proposed is evaluated using the MUStARD [147] dataset. The results of the grid search were used to obtain the optimal hyper-parameters. Our study aims to utilize a consistent hyper-parameter setup across all experimental trials.

Also, it is worth noting that all the experimental procedures were conducted using high-end GPU systems featuring the following specifications: NVIDIA Titan RTX (48 GB), 256 GB of RAM, 10 TB of storage space, and an Intel Xeon Silver 4116 processor. The computational model being analysed demonstrates a relatively low requirement for GPU memory, utilizing approximately 2 gigabytes. Typically, on an average, each epoch requires a duration of approximately 3 to 4 seconds.

To ensure a rigorous evaluation in accordance with current cutting-edge frameworks, we conducted comprehensive experiments encompassing unimodal, bimodal, and trimodal approaches for both "speaker dependent" and "speaker independent" setups.

Results and Discussion

The proposed architecture was assessed through a comprehensive analysis of all potential combinations of input. These include unimodal inputs such as $\mathbb{G}$, $\mathbb{V}$, and $\mathbb{A}$, bimodal inputs such as $\mathbb{G}\oplus\mathbb{V}$, $\mathbb{V}\oplus\mathbb{A}$, and $\mathbb{G}\oplus\mathbb{A}$, as well as trimodal input $\mathbb{G}\oplus\mathbb{V}\oplus\mathbb{A}$. In the context of speaker dependent configuration, our proposed VyAnG-Net (for trimodal) proved its superior performance (**Table 6.2**) with a $\mathbb{Precision}(P)$ of 78.83% (an increase of 6.93 [147], 3.63[148], and 4.63[149] points), $\mathbb{Recall}(R)$ of 78.21% (an increase of 6.81 [147], 3.61[148], and 4.01[149] points), and $\mathbb{F}1 - \mathbb{Score}(F1)$ of 78.52% (an increase of 7.02 [147], 4.02[148], and 4.32[149] points). Experimental evidence indicates that the trimodal approach outperforms both the unimodal and bimodal approaches.

In addition, for speaker $-$ independent configuration, the proposed model VyAnG-Net (for trimodal) exhibited exceptional performance (**Table 6.3**) with $\mathbb{Precision}(P)$ of 75.69% (an increase of 11.39 [147], 4.39[148], and 3.59[149] points), $\mathbb{Recall}(R)$ of 75.52% (an increase of 12.92 [147], 4.22[148], and 3.52[149] points), and $\mathbb{F}1\ \mathbb{Score}(F1)$ of 75.6% (an increase of 12.8[147], 5.6[148], and 3.6[149] points). **Figure 6.5** and **Figure 6.6** illustrate the curves pertaining to testing loss, accuracy, precision, recall, and F1 scores for speaker $-$ dependent and speaker $-$ independent configurations.

Table 6.2 Experimental results for speaker-dependent setup using VyAnG-Net.

Modalities	**Speaker dependent**			
	$\mathbb{Accuracy}(\mathbf{A})$	$\mathbb{Precision}(\mathbf{P})$	$\mathbb{Recall}(\mathbf{R})$	$\mathbb{F}\mathbf{1}\ \mathbb{Score}(\mathbf{F1})$
$\mathbb{G}$ (**unimodal**)	73.16	72.53	72.4	72.45
$\mathbb{V}$ (**unimodal**)	73.81	72.93	71.86	72.4
$\mathbb{A}$ (**unimodal**)	74.42	72.7	72.69	72.7
$\mathbb{G}\oplus\mathbb{V}$ (**bimodal**)	75.68	74.94	74.26	74.61
$\mathbb{V}\oplus\mathbb{A}$ (**bimodal**)	77.37	77.53	77.45	77.49
$\mathbb{G}\oplus\mathbb{A}$ (**bimodal**)	77.14	76.84	76.92	76.87
$\mathbb{G}\oplus\mathbb{V}\oplus\mathbb{A}$ (**trimodal**)	**79.86**	**78.83**	**78.21**	**78.52**

Table 6.3 Experimental results for speaker-independent setup using VyAnG-Net.

Modalities	**Speaker Independent**			
	$\mathbb{Accuracy}(\mathbf{A})$	$\mathbb{Precision}(\mathbf{P})$	$\mathbb{Recall}(\mathbf{R})$	$\mathbb{F}\mathbf{1}\ \mathbb{Score}(\mathbf{F1})$
$\mathbb{G}$ (**unimodal**)	69.47	68.36	68.24	68.29
$\mathbb{V}$ (**unimodal**)	70.15	70.89	70.16	70.52
$\mathbb{A}$ (**unimodal**)	70.92	71.12	71.23	71.17
$\mathbb{G}\oplus\mathbb{V}$ (**bimodal**)	72.42	72.64	72.61	72.61
$\mathbb{V}\oplus\mathbb{A}$ (**bimodal**)	74.74	74.51	74.32	74.41

Modalities	**Speaker Independent**			
	$\mathbb{Accuracy}(\mathbf{A})$	$\mathbb{Precision}(\mathbf{P})$	$\mathbb{Recall}(\mathbf{R})$	$\mathbb{F1\ Score}(\mathbf{F1})$
$\mathbb{G}\oplus\mathbb{A}$ (**bimodal**)	74.64	73.96	73.48	73.72
$\mathbb{G}\oplus\mathbb{V}\oplus\mathbb{A}$ (**trimodal**)	**76.94**	**75.69**	**75.52**	**75.6**

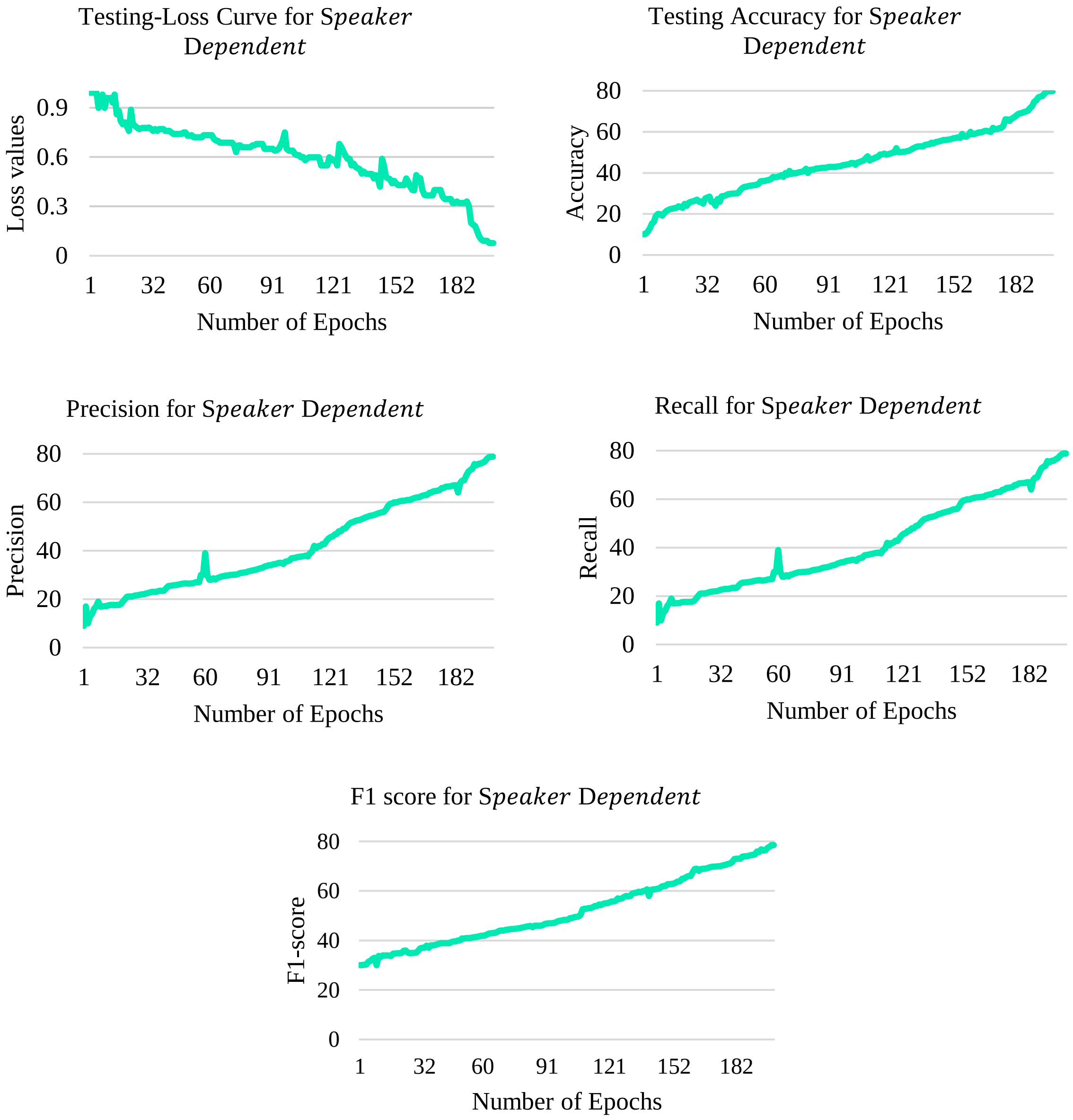


Figure 6.5 Testing curve for loss, accuracy, precision, recall, and F1 scores for $speaker - dependent$.

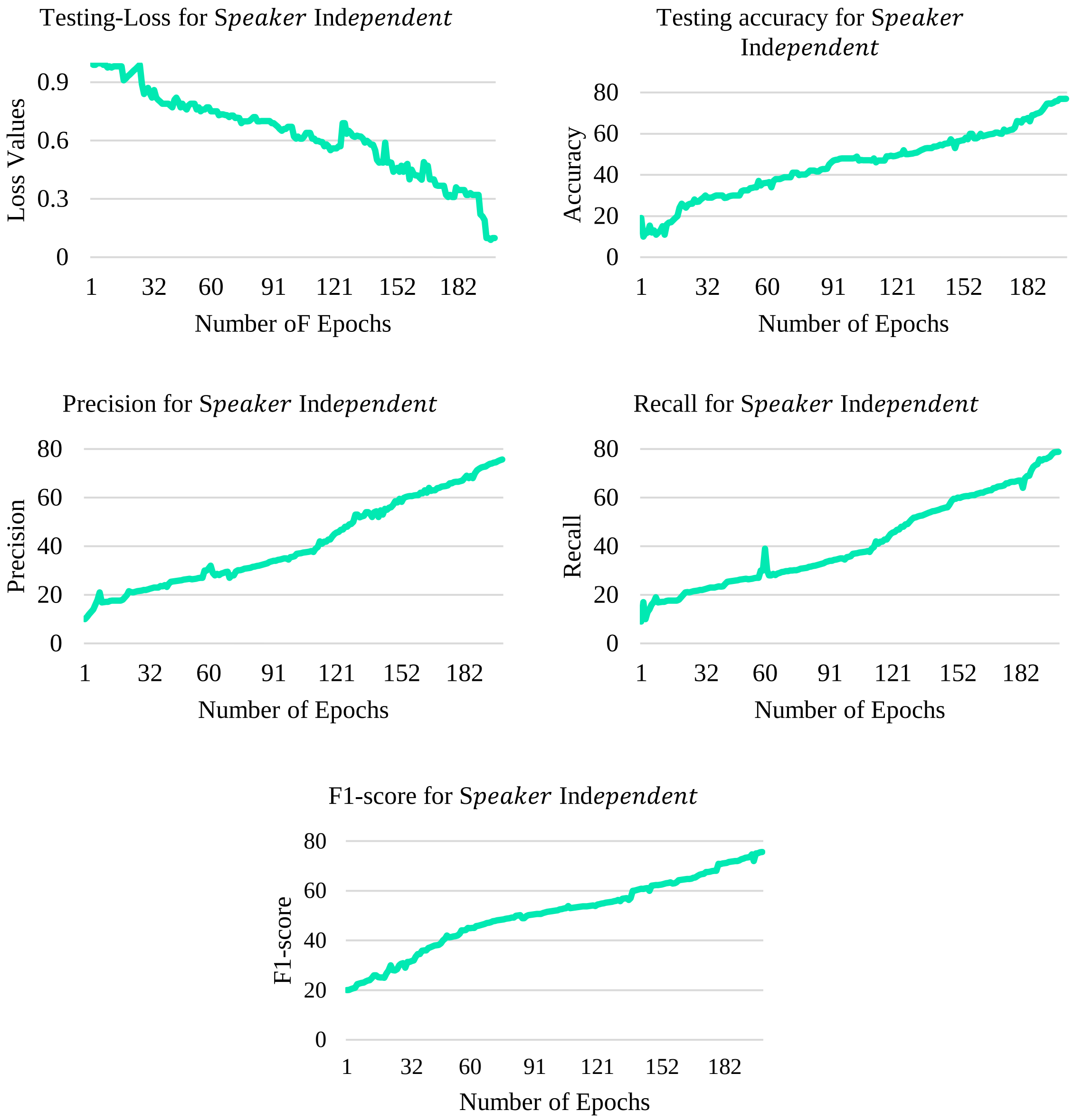


Figure 6.6 Testing curve for loss, accuracy, precision, recall, and F1 scores for $speaker - independent$.

Comparative Analysis with the Baselines

In this study, we conducted a comparative analysis using consistent experimental conditions with the pre-existing models, namely Baseline − 1 [147], Baseline − 2 [148], Baseline − 3 [149], Baseline − 4 [150], and Baseline − 5 [151], which were developed using a similar MUStARD [147] dataset. Baseline − 1 [147] was the first to release and work on the video dataset MUStARD in the field of MSR. This study employed ResNet-152 for the purpose of extracting visual features, BERT-based uncased architecture for extracting textual features, and the librosa library for extracting auditory features. Finally, a Support Vector Machine (SVM) was employed as a classifier to distinguish between instances of sarcasm and non-sarcasm.

Later, the IWAN framework was proposed in $\text{Baseline} - 2$ [148], which incorporates attention mechanisms and employs similar architectures as employed in [147] for extracting visual and textual features. However, for the extraction of auditory features, the OpenSmile tool was utilized. The authors in $\text{Baseline} - 3$ [149] proposed a novel methodology for MSR that integrates ResNet-152, BART, and librosa for the purpose of feature extraction across various modalities. This research study has undertaken the task of emotion recognition in addition to implicit sentiment, specifically sarcasm. The study conducted by $\text{Baseline} - 5$ [151] employed comparable architectures as those utilised in the previous research [147], in conjunction with a late fusion strategy, to detect instances of sarcasm in utterances. In addition to the aforementioned research studies, it is noteworthy to mention that $\text{Baseline} - 4$ [150] was the pioneer in utilising fuzzy logic and quantum theory to detect sarcasm in video utterances.

Our methodology outperforms all of the baseline models that have been addressed in **Table 6.4** and **Table 6.5**. The results indicate that the VyAnG-Net model is advantageous for MSR due to its ability to comprehensively stimulate the relationship between all the modalities at a more profound level. This is achieved through the use of a glossary branch that employs an attention-based tokenization approach and a dedicated attention module to identify the most salient features from the video frames, along with a multi-headed attention-based feature fusion technique. The visualisation of VyAnG-Net on MUStARD dataset versus the cutting-edge baseline approaches in terms of Accuracy, Precision, Recall, and F1 scores for speaker dependent and $\text{speaker independent}$ configuration are presented in **Figure 6.7** and **Figure 6.8**.

Table 6.4 Comparison with baseline models for $speaker - dependent$ setup.

	$\mathbb{Precision}(\mathbf{P})$						$\mathbb{Recall}(\mathbf{R})$						$\mathbb{F}\mathbf{1}\ \mathbb{Score}(\mathbf{F1})$					
Modalities	**VyAnG-Net**	[147]	[148]	[149]	[150]	[151]	**VyAnG-Net**	[147]	[148]	[149]	[150]	[151]	**VyAnG-Net**	[147]	[148]	[149]	[150]	[151]
$\mathbb{G}$	**72.53**	65.1	-	-	-	67.73	**72.4**	64.6	-	-	-	66.8	**72.45**	64.6	-	-	-	67.66
$\mathbb{V}$	**72.93**	68.1	-	-	-	71.6	**71.86**	67.4	-	-	-	71.06	**72.4**	67.4	-	-	-	70.84
$\mathbb{A}$	**74.71**	65.9	-	-	-	73.22	**73.69**	64.6	-	-	-	72.61	**73.7**	64.6	-	-	-	72.44

Modalities	Precision$(\mathbf{P})$						Recall$(\mathbf{R})$						F1 Score$(\mathbf{F1})$					
	VyAnG-Net	[147]	[148]	[149]	[150]	[151]	**VyAnG-Net**	[147]	[148]	[149]	[150]	[151]	**VyAnG-Net**	[147]	[148]	[149]	[150]	[151]
$\mathbb{G}\oplus\mathbb{V}$	**74.94**	72	69.1	-	-	73.44	**74.26**	71.6	68.9	-	-	73.33	**74.61**	71.6	68.9	-	-	73.3
$\mathbb{V}\oplus\mathbb{A}$	**77.53**	66.2	-	-	-	73.81	**77.45**	65.7	-	-	-	73.33	**77.49**	65.7	-	-	-	73.19
$\mathbb{G}\oplus\mathbb{A}$	**76.84**	66.6	70.8	-	-	71.19	**76.92**	66.2	70.2	-	-	70.87	**76.87**	66.2	70.2	-	-	70.87
$\mathbb{G}\oplus\mathbb{V}\oplus\mathbb{A}$	**78.83**	71.9	75.2	74.2	75.3	73.8	**78.21**	71.4	74.6	74.2	75.5	73.62	**78.52**	71.5	74.5	74.2	75.4	73.58

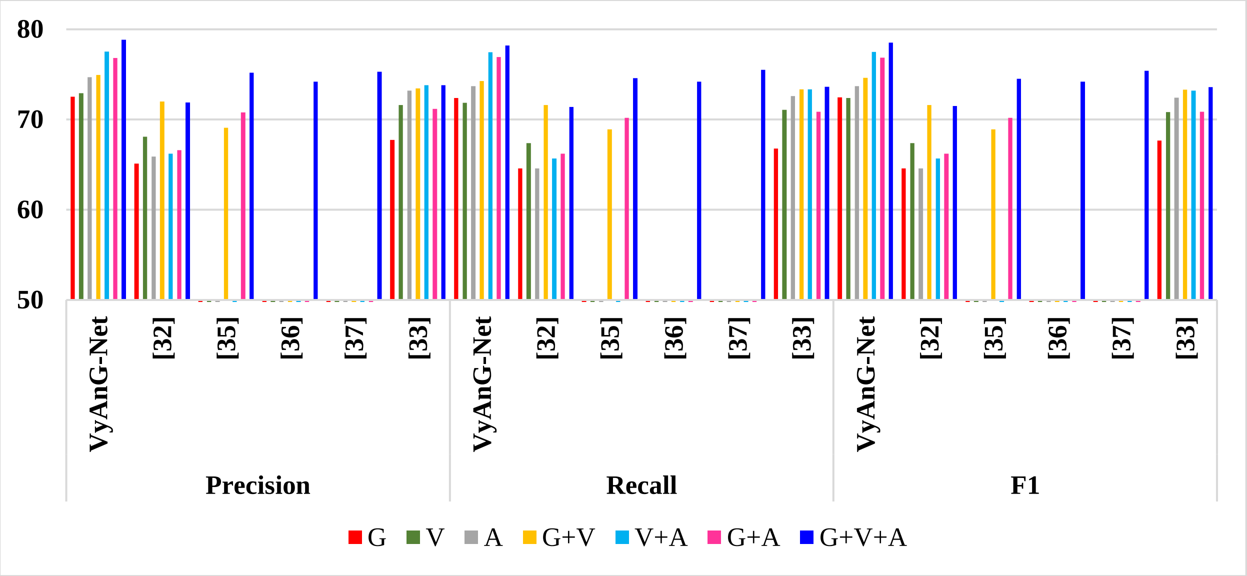


Figure 6.7 Evaluation of our proposed framework VyAnG-Net on MUStARD dataset versus the cutting-edge baseline approaches in terms of Accuracy, Precision, Recall, and F1 scores for $speaker - dependent$ configuration.

Table 6.5 Comparison with baseline models for $speaker - independent$ setup.

Modalities	Precision$(\boldsymbol{P})$						Recall$(\boldsymbol{R})$						F1 Score$(\boldsymbol{F1})$					
	VyAnG-Net	[147]	[148]	[149]	[150]	[151]	**VyAnG-Net**	[147]	[148]	[149]	[150]	[151]	**VyAnG-Net**	[147]	[148]	[149]	[150]	[151]
$\mathbb{G}$	**68.36**	60.9	-	-	-	58.18	**68.24**	59.6	-	-	-	57.75	**68.29**	59.8	-	-	-	57.84

Modalities	Precision ($\boldsymbol{P}$)						Recall ($\boldsymbol{R}$)						F1 Score ($\boldsymbol{F1}$)					
	VyAnG-Net	[147]	[148]	[149]	[150]	[151]	**VyAnG-Net**	[147]	[148]	[149]	[150]	[151]	**VyAnG-Net**	[147]	[148]	[149]	[150]	[151]
$\mathbb{V}$	**70.89**	54.9	-	-	-	70.37	**70.16**	53.4	-	-	-	70.56	**70.52**	53.6	-	-	-	70.13
$\mathbb{A}$	**73.92**	65.1	-	-	-	72.93	**73.23**	62.6	-	-	-	71.4	**73.58**	62.7	-	-	-	71.21
$\mathbb{G}\oplus\mathbb{V}$	**72.64**	62.2	63.2	-	-	64.7	**72.61**	61.5	63.1	-	-	64.72	**72.61**	61.5	63.1	-	-	63.98
$\mathbb{V}\oplus\mathbb{A}$	**74.51**	64.1	-	-	-	72.44	**74.32**	61.8	-	-	-	71.57	**74.41**	61.9	-	-	-	70.27
$\mathbb{G}\oplus\mathbb{A}$	**73.96**	64.7	59.6	-	-	62.75	**73.48**	62.9	60.0	-	-	61.24	**73.72**	63.1	59.7	-	-	61.33
$\mathbb{G}\oplus\mathbb{V}\oplus\mathbb{A}$	**75.69**	64.3	71.9	72.1	75.3	71.55	**75.52**	62.6	71.3	72	75.5	71.52	**75.6**	62.8	70.0	72	75.4	70.99

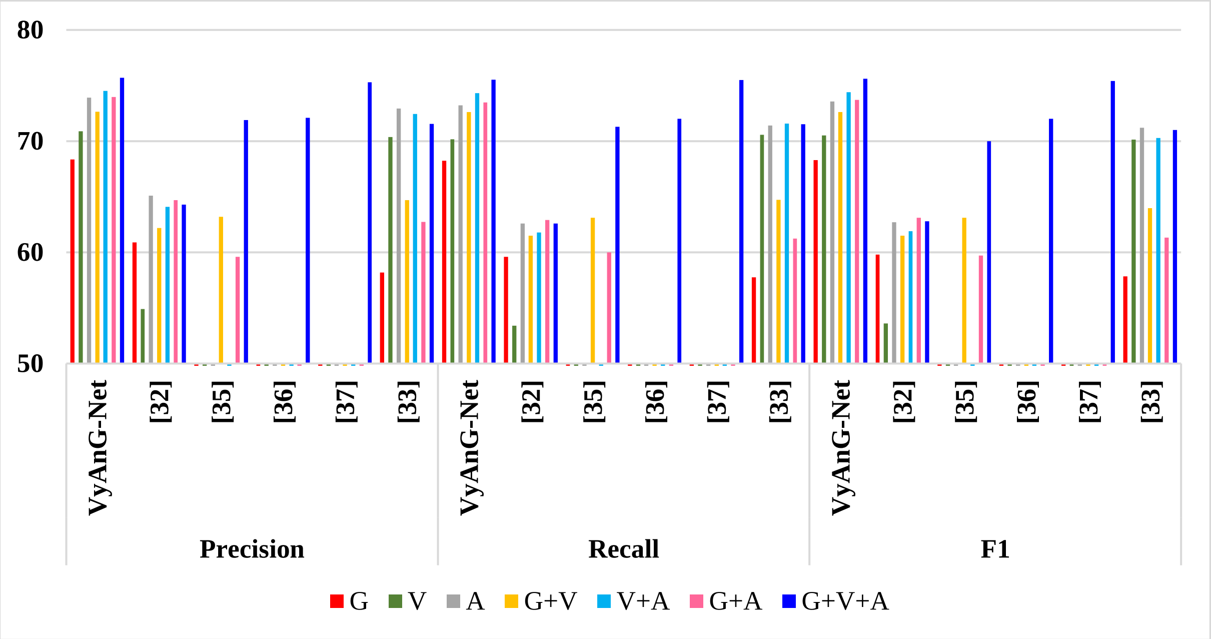


Figure 6.8 Evaluation of our proposed framework VyAnG-Net on MUStARD dataset versus the cutting-edge baseline approaches in terms of Accuracy, Precision, Recall, and F1 scores for $speaker - independent$ configuration.

Ablation Study

In this section, a series of ablation trials were carried out on the MUStARD dataset to more accurately evaluate the effectiveness of each proposed module specifically for trimodal $(\mathbb{G}\oplus \mathbb{V}\oplus\mathbb{A})$. The VyAnG-Net framework was employed to generate three distinct variants, namely “VyAnG-Net w/o attention-seeking tokenizer”, “VyAnG-Net w/o lightweight depth attention”, and “VyAnG-Net w/o multi-headed attention” as illustrated in **Table 6.6** and **Table 6.7**. The aforementioned modifications entail the extraction of features from the textual content of video utterance’s subtitles by eliminating the attention-seeking tokenizer in the glossary branch module, removal of the lightweight depth attention module in the visual branch, and instead of utilising multi-headed attention for fusion, a method of directly concatenating the obtained feature representation from multiple branches was employed. **Table 6.6** and **Table 6.7** summarise the outcomes obtained from the ablation experiments.

Based on these observations, the following conclusions can be drawn: The VyAnG-Net, as the proposed model, encompasses all modules and exhibits superior performance on the MUStARD dataset. The removal of a single module would lead to inadequate predictive outcomes. Based on these observations, it can be inferred that every proposed module is crucial and plays a significant role in the overall performance. Also, it can be observed that in the speaker-dependent configuration, the proposed model is performing better because of the intermixing of a variety of videos in the training set as compared to the speaker-independent configuration, resulting in enhanced generalizability of speaker-dependent configuration.

Table 6.6 The results of the ablation trials carried out on the MUStARD dataset for a $speaker - dependent$ configuration.

Methods	**MUStARD**			
	$\mathbb{Accuracy}(\boldsymbol{A})$	$\mathbb{Precision}(\boldsymbol{P})$	$\mathbb{Recall}(\boldsymbol{R})$	$\mathbb{F}\mathbf{1}\ \mathbb{Score}(\boldsymbol{F1})$
VyAnG-Net w/o attention-seeking tokenizer	76.49	75.87	75.62	75.74
VyAnG-Net w/o lightweight depth attention	75.24	75.53	74.97	75.47
VyAnG-Net w/o multi-headed attention	73.86	72.63	72.2	72.41
VyAnG-Net	**79.86**	**78.83**	**78.21**	**78.52**

Table 6.7 The results of the ablation trials carried out on the MUStARD dataset for a $speaker - independent$ configuration.

Methods	**MUStARD**			
	$\mathbb{Accuracy}(\boldsymbol{A})$	$\mathbb{Precision}(\boldsymbol{P})$	$\mathbb{Recall}(\boldsymbol{R})$	$\mathbb{F}\mathbf{1}\ \mathbb{Score}(\boldsymbol{F1})$
VyAnG-Net w/o attention-seeking tokenizer	73.96	73.32	72.84	73.08

Methods	**MUStARD**			
	$\mathbb{Accuracy}(\boldsymbol{A})$	$\mathbb{Precision}(\boldsymbol{P})$	$\mathbb{Recall}(\boldsymbol{R})$	$\mathbb{F}\mathbf{1}\ \mathbb{Score}(\boldsymbol{F1})$
VyAnG-Net w/o lightweight depth attention	71.42	72.03	71.99	72.00
VyAnG-Net w/o multi-headed attention	73.85	72.74	72.62	72.68
VyAnG-Net	**76.94**	**75.69**	**75.52**	**75.6**

Generalization study

In the recent years the MUStARD dataset has shown remarkable results for various emerging multimodal sarcasm recognition frameworks; however, these approaches do not possess the generalizability that algorithms need to examine samples from other domains or datasets. They are more concerned in conducting thorough architectural in-dataset analyses. Consequently, rather than limiting ourselves to assessing VyAnG-Net on a single dataset, we provide a cross-dataset evaluation that puts a model trained on one dataset to the test on another. Therefore, in addition to the experiment discussed in the above sections, we undertake a cross-dataset study as part of a generalization research to test the resilience of VyAnG-Net. Our proposed approach, VyAnG-Net, was trained using the MUStARD dataset for this experimental investigation, and its performance was evaluated using the MUStARD++ dataset. Throughout the training stage of the cross-dataset study, 80% of the samples from the MUStARD dataset were chosen at random for training, while 10% were allocated for validation. For the testing phase, 10% of the samples had been picked at random from the MUStARD++ dataset to assess the predictive power of the proposed method using various parameters including $\mathbb{Accuracy}(\boldsymbol{A})$, $\mathbb{Precision}(\boldsymbol{P})$**,** $\mathbb{Recall}(\boldsymbol{R})$**,** $\mathbb{F}\mathbf{1}\ \mathbb{Score}(\boldsymbol{F1})$. The efficiency of VyAnG-Net, which is based on lightweight depth attention module and the attention-based tokenization approach, was tested through cross-dataset study. From **Table 6.8** It is evident that our technique could identify instances from datasets other than the ones it was trained on. Hence, it can be proved that our suggested model, VyAnG-Net, is effective, generalizable and more reliable than earlier state-of-the-art solutions.

Table 6.8 Evaluating the resilience of VyAnG-Net using cross-dataset study that uses MUStARD dataset for training and validation, whereas MUStARD++ is employed for testing purposes.

Dataset utilized	**VyAnG-Net**			
	$\mathbb{Accuracy}(\boldsymbol{A})$	$\mathbb{Precision}(\boldsymbol{P})$	$\mathbb{Recall}(\boldsymbol{R})$	$\mathbb{F}\mathbf{1}\ \mathbb{Score}(\boldsymbol{F1})$
MUStARD (for training and validation) and MUStARD ++ (for testing)	**73.93**	**72.41**	**72.05**	**72.23**

6.2.4 Conclusion & Future Scope

This paper proposes a novel VyAnG-Net for multi-modal sarcasm recognition by uncovering visual, acoustic and glossary features. Our proposed methodology analyses the interaction between all three modalities more effectively than earlier innovative methods. The sentiment evoked by a particular phrase may vary across different contexts. Therefore, it is vital to utilise auditory and visual cues to boost prediction accuracy. Motivated by this, we have introduced an innovative visual branch incorporating a lightweight depth attention module to extract the most salient features from video frames. Additionally, a glossary branch utilises an attention-based tokenization approach to capture the most critical contextual features from the textual content provided by video subtitles. Furthermore, an utterance-level feature extraction method for acoustic content has been implemented along with a multi-headed attention-based feature fusion technique to combine features obtained from each of the distinct modalities. Our extensive experiments on the publicly available dataset MUStARD indicate that our proposed model outperforms other competitive baseline models.

6.3 Significant Outcomes of this Chapter

The significant outcomes of this chapter are as follows:

- We proposed VyAnG-Net, a novel multi-modal sarcasm recognition framework, by uncovering visual, acoustic and glossary (textual) features. This framework includes the glossary branch that uses the attention-based tokenization approach to acquire the most significant contextual features from the textual content provided by the subtitles of the video utterances, a visual unit with a dedicated lightweight depth attention module to acquire the most prominent features from the video frames, an utterance-level feature extraction from acoustic content and lastly multi-headed attention based feature fusion has been employed to blend features acquired from each of the separate modalities.
- In recent years, remarkable advancements have been made in sarcasm identification frameworks on the MUStARD dataset, but there are concerns about their generalizability. Consequently, rather than limiting ourselves to assessing VyAnG-Net on a single dataset, we undertake a cross-dataset study as part of generalization research to test the resilience of VyAnG-Net. Our proposed approach, VyAnG-Net, was trained using the MUStARD dataset for this experimental investigation, and its performance was evaluated using an unseen MUStARD++ dataset.

The following research works form the basis of this chapter:

- **A. Pandey** and D. K. Vishwakarma, “VyAnG-Net: A Novel Multi-Modal Sarcasm Recognition Model by Uncovering Visual, Acoustic and Glossary Features.” Under Minor Revisiom in ***Intelligent Data Analysis*** (Pub: IOS Press). https://doi.org/10.48550/arXiv.2408.10246.
- **A. Pandey** and D. K. Vishwakarma, “Multimodal Sarcasm Detection (MSD) in Videos using Deep Learning Models,” in ***2023 IEEE International Conference in Advances in Power, Signal, and Information Technology (APSIT)***, Institute of Electrical and Electronics Engineers (IEEE), Jun. 2023, pp. 811-814. doi: 10.1109/APSIT58554.2023.10201731.

Chapter 7: Conclusion & Future Scope

7.1 Conclusion

This chapter serves as the finalization of the research conducted in this thesis. In summary, we conduct four research contributions dealing with different aspects of identifying sentiments in multimedia data. These approaches are summarized as follows:

- ✓ In the first research work, the proposed VABDC-Net, integrates an attention module with the convolutional neural network to focus on the most relevant information from the visual modality and attentional tokenizer-based method to extract the most relevant contextual information from the caption modality. Thorough experimentation on two benchmark datasets, Twitter-15, with an accuracy of 83.80%, and Twitter-17, with an accuracy of 72.42%, indicates that our technique outperforms existing methods for Visual-Caption Sentiment Recognition.

- ✓ In the second approach, the proposed model VECT-Net includes: the facial emotion description module, the target alignment and refinement module for face description, and the fusion module. The facial emotion description unit is responsible for generating a face description that includes various features such as age, gender, and emotion. The target alignment and refinement module estimates the cosine similarity between the visual input and the face descriptions with the target. In the fusion component, two robustly optimised pre-trained language models are utilised to simulate images, captions and face descriptions by a gating mechanism for feature fusion and noise reduction. The experimental results show that the suggested model achieves an accuracy of 81.23% and a macro-F1 of 80.61% on the Twitter-15 dataset, while 77.42% and 75.19% on the Twitter-17 dataset, respectively.

- ✓ In the third approach, Contrastive Learning-based Multi-Modal Architecture has been developed to predict emoticons by Employing Image-Text Pairs. The proposed model employs the joint training of dual-branch encoder along with the contrastive learning to accurately map text and images into a common latent space. Our key finding is that by integrating the principle of contrastive learning with that of the other two branches yields superior results. The experimental results demonstrate that our suggested methodology surpasses existing multimodal approaches in terms of accuracy and robustness. The proposed model attained an accuracy of 91% and an MCC-score of 90% while assessing emoticons using the Multimodal-Twitter Emoticon dataset acquired from Twitter.

- ✓ Lastly, a novel approach has been proposed to detect sarcasm in videos that combines a self-regulated Convolutional neural network to concentrate on the most crucial features of visual data and an attentional tokenizer-based strategy to extract the most critical context-specific information from the textual data. Extensive testing on one of the benchmark video datasets, MUStARD, yielded an accuracy of 78.52% for speaker-dependent configuration.

7.2 Future Scope

Despite attaining promising results, there are still numerous avenues open for further research as highlighted below:

- ✓ **Fusion Strategy:** Our research has focused on the feature fusion approach by utilising a multi-headed attention mechanism for MSR. Subsequent research endeavours may explore advanced spatiotemporal fusion techniques to effectively capture the correlation between different modalities, such as tensor-based fusion. A potential alternative strategy entails formulating more fusion methodologies that can better encapsulate the discrepancies among diverse modalities to identify occurrences of sarcasm with greater efficacy.
- ✓ **Neural Baseline:** Further research studies should aim to implement transfer learning, pre-training, low-parameter, or domain adaptation models as potential solutions. Also, the architectures with fewer trainable parameters are more advantageous for facilitating real-time deployment.
- ✓ **Visual sarcasm recognition:** So far, there has been a lack of research in the domain of sarcasm detection utilising solely visual cues that include the embedded text, primarily due to the absence of an appropriate dataset. Therefore, developing a visual sample corpus centred on memes is essential to work effectively in this field.
- ✓ **Integration of Attention-based strategies:** In the modern era, there has been an explosion of attention modules that need further study for their potential integration with diverse forms of Convolutional Neural Networks (ConvNets) and Recurrent Neural Networks (RNNs). This integration could facilitate the efficient extraction of pertinent information from both textual and visual inputs.

References

[1] "Social Network Usage & Growth Statistics (2024)," Backlinko. Accessed: Aug. 15, 2024. [Online]. Available: https://backlinko.com/social-media-users

[2] L. Teijeiro-Mosquera, J.-I. Biel, J. L. Alba-Castro, and D. Gatica-Perez, "What Your Face Vlogs About: Expressions of Emotion and Big-Five Traits Impressions in YouTube," *IEEE Trans. Affective Comput.*, vol. 6, no. 2, pp. 193–205, Apr. 2015, doi: 10.1109/TAFFC.2014.2370044.

[3] E. Cambria, "Affective Computing and Sentiment Analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar. 2016, doi: 10.1109/MIS.2016.31.

[4] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment Analysis Is a Big Suitcase," *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 74–80, Nov. 2017, doi: 10.1109/MIS.2017.4531228.

[5] A. Pandey and D. K. Vishwakarma, "VABDC-Net: A framework for Visual-Caption Sentiment Recognition via spatio-depth visual attention and bi-directional caption processing," *Knowledge-Based Systems*, vol. 269, pp. 1105–11015, Jun. 2023, doi: https://doi.org/10.1016/j.knosys.2023.110515.

[6] P. Gupta, A. Pandey, A. Kumar, and D. K. Vishwakarma, "Attention-free based dual-encoder mechanism for Aspect-based Multimodal Sentiment Recognition," in *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*, Bhubaneswar, India: IEEE, Jun. 2023, pp. 534–539. doi: 10.1109/APSIT58554.2023.10201711.

[7] A. Pandey and D. K. Vishwakarma, "Multimodal Sarcasm Detection (MSD) in Videos using Deep Learning Models," in *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*, Jun. 2023. doi: 10.1109/APSIT58554.2023.10201731.

[8] S. Aggarwal, A. Pandey, and D. K. Vishwakarma, "Multimodal Sarcasm Recognition by Fusing Textual, Visual and Acoustic content via Multi-Headed Attention for Video Dataset," in *2023 World Conference on Communication & Computing (WCONF)*, Jul. 2023, pp. 1–5. doi: 10.1109/WCONF58270.2023.10235179.

[9] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*, Alicante Spain: ACM, Nov. 2011, pp. 169–176. doi: 10.1145/2070481.2070509.

[10] R. V. Perez, M. Rada, and M. Louis-Philippe, "Utterance-Level Multimodal Sentiment Analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Aug. 2013, pp. 973–982.

[11] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency, “Multimodal Analysis and Prediction of Persuasiveness in Online Social Multimedia,” *ACM Trans. Interact. Intell. Syst.*, vol. 6, no. 3, pp. 1–25, Oct. 2016, doi: 10.1145/2897739.

[12] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, “Multi-attention recurrent network for human communication comprehension,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, in AAAI’18/IAAI’18/EAAI’18. New Orleans, Louisiana, USA: AAAI Press, Feb. 2018, pp. 5642–5649.

[13] D. Gkoumas, Q. Li, C. Lioma, Y. Yu, and D. Song, “What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis,” *Information Fusion*, vol. 66, pp. 184–197, Feb. 2021, doi: 10.1016/j.inffus.2020.09.005.

[14] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, “Deep multimodal fusion for persuasiveness prediction,” in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, Tokyo Japan: ACM, Oct. 2016, pp. 284–288. doi: 10.1145/2993148.2993176.

[15] N. Xu and W. Mao, “A residual merged neutral network for multimodal sentiment analysis,” in *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*, Beijing, China: IEEE, Mar. 2017, pp. 6–10. doi: 10.1109/ICBDA.2017.8078794.

[16] P. Kumar, V. Khokher, Y. Gupta, and B. Raman, “Hybrid Fusion Based Approach for Multimodal Emotion Recognition with Insufficient Labeled Data,” in *2021 IEEE International Conference on Image Processing (ICIP)*, Anchorage, AK, USA: IEEE, Sep. 2021, pp. 314–318. doi: 10.1109/ICIP42928.2021.9506714.

[17] A. Kumar, K. Srinivasan, W.-H. Cheng, and A. Y. Zomaya, “Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data,” *Information Processing & Management*, vol. 57, no. 1, p. 102141, Jan. 2020, doi: 10.1016/j.ipm.2019.102141.

[18] S. Mai, S. Xing, and H. Hu, “Analyzing Multimodal Sentiment Via Acoustic- and Visual-LSTM With Channel-Aware Temporal Convolution Network,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1424–1437, 2021, doi: 10.1109/TASLP.2021.3068598.

[19] Z. A. Bagher, L. P. Pu, P. Soujanya, C. Erik, and M. Louis-Philippe, “Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Jul. 2018, pp. 2236–2246.

[20] A. Shenoy and A. Sardana, “Multilogue-Net: A Context-Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in Conversation,” in *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, Seattle, USA: Association for Computational Linguistics, 2020, pp. 19–28. doi: 10.18653/v1/2020.challengehml-1.3.

[21] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, “A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis,” in *Second Grand-Challenge and*

Workshop on Multimodal Language (Challenge-HML), Seattle, USA: Association for Computational Linguistics, 2020, pp. 1–7. doi: 10.18653/v1/2020.challengehml-1.1.

[22] Z. Wang, Z. Wan, and X. Wan, "TransModality: An End2End Fusion Method with Transformer for Multimodal Sentiment Analysis," in *Proceedings of The Web Conference 2020*, Taipei Taiwan: ACM, Apr. 2020, pp. 2514–2520. doi: 10.1145/3366423.3380000.

[23] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM," *Multimed Tools Appl*, vol. 80, no. 9, pp. 13059–13076, Apr. 2021, doi: 10.1007/s11042-020-10285-x.

[24] M. Chen and X. Li, "SWAFN: Sentimental Words Aware Fusion Network for Multimodal Sentiment Analysis," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 1067–1077. doi: 10.18653/v1/2020.coling-main.93.

[25] T. Wu *et al.*, "Video sentiment analysis with bimodal information-augmented multi-head attention," *Knowledge-Based Systems*, vol. 235, p. 107676, Jan. 2022, doi: 10.1016/j.knosys.2021.107676.

[26] K. Zhang, Y. Zhu, W. Zhang, and Y. Zhu, "Cross-modal image sentiment analysis via deep correlation of textual semantic," *Knowledge-Based Systems*, vol. 216, p. 106803, Mar. 2021, doi: 10.1016/j.knosys.2021.106803.

[27] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in Translation: Learning Robust Joint Representations by Cyclic Translations between Modalities," *AAAI*, vol. 33, no. 01, pp. 6892–6899, Jul. 2019, doi: 10.1609/aaai.v33i01.33016892.

[28] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 6558–6569. doi: 10.18653/v1/P19-1656.

[29] Y. Zhang, J. Wang, and X. Zhang, "Conciseness is better: Recurrent attention LSTM model for document-level sentiment analysis," *Neurocomputing*, vol. 462, pp. 101–112, Oct. 2021, doi: 10.1016/j.neucom.2021.07.072.

[30] S. Liu and I. Lee, "Sequence encoding incorporated CNN model for Email document sentiment classification," *Applied Soft Computing*, vol. 102, p. 107104, Apr. 2021, doi: 10.1016/j.asoc.2021.107104.

[31] C. Yang, X. Chen, L. Liu, and P. Sweetser, "Leveraging semantic features for recommendation: Sentence-level emotion analysis," *Information Processing & Management*, vol. 58, no. 3, p. 102543, May 2021, doi: 10.1016/j.ipm.2021.102543.

[32] H. Wu, Z. Zhang, S. Shi, Q. Wu, and H. Song, "Phrase dependency relational graph attention network for Aspect-based Sentiment Analysis," *Knowledge-Based Systems*, vol. 236, p. 107736, Jan. 2022, doi: 10.1016/j.knosys.2021.107736.

[33] B. Liang, H. Su, L. Gui, E. Cambria, and R. Xu, “Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks,” *Knowledge-Based Systems*, vol. 235, p. 107643, Jan. 2022, doi: 10.1016/j.knosys.2021.107643.

[34] Y.-C. Chang, C.-H. Ku, and D.-D. L. Nguyen, “Predicting aspect-based sentiment using deep learning and information visualization: The impact of COVID-19 on the airline industry,” *Information & Management*, vol. 59, no. 2, p. 103587, Mar. 2022, doi: 10.1016/j.im.2021.103587.

[35] L. Vu and T. Le, “A lexicon-based method for Sentiment Analysis using social network data,” in *The 2017 WorldComp International Conference Proceedings*, Jul. 2017, pp. 11–16.

[36] S. Pang, B., Lee, L., & Vithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques,” in *Proceedings of the Institution of Civil Engineers - Transport*, 2019.

[37] L. Barbosa and J. Feng, “Robust sentiment detection on Twitter from biased and noisy data,” in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, in COLING ’10. USA: Association for Computational Linguistics, Aug. 2010, pp. 36–44.

[38] R. Mao and X. Li, “Bridging Towers of Multi-task Learning with a Gating Mechanism for Aspect-based Sentiment Analysis and Sequential Metaphor Identification,” *AAAI*, vol. 35, no. 15, pp. 13534–13542, May 2021, doi: 10.1609/aaai.v35i15.17596.

[39] K. He, R. Mao, T. Gong, C. Li, and E. Cambria, “Meta-Based Self-Training and Re-Weighting for Aspect-Based Sentiment Analysis,” *IEEE Trans. Affective Comput.*, vol. 14, no. 3, pp. 1731–1742, Jul. 2023, doi: 10.1109/TAFFC.2022.3202831.

[40] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, “SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, Eds., Marseille, France: European Language Resources Association, Jun. 2022, pp. 3829–3839. Accessed: Jul. 28, 2024. [Online]. Available: https://aclanthology.org/2022.lrec-1.408

[41] E. Cambria, R. Mao, S. Han, and Q. Liu, “Sentic Parser: A Graph-Based Approach to Concept Extraction for Sentiment Analysis,” in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, Orlando, FL, USA: IEEE, Nov. 2022, pp. 1–8. doi: 10.1109/ICDMW58026.2022.00060.

[42] R. Mao, X. Li, M. Ge, and E. Cambria, “MetaPro: A computational metaphor processing model for text pre-processing,” *Information Fusion*, vol. 86–87, pp. 30–43, Oct. 2022, doi: 10.1016/j.inffus.2022.06.002.

[43] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, “The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection,” *IEEE Trans. Affective Comput.*, vol. 14, no. 3, pp. 1743–1753, Jul. 2023, doi: 10.1109/TAFFC.2022.3204972.

[44] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. doi: 10.3115/v1/D14-1181.

[45] D. Tang, B. Qin, and T. Liu, “Document Modeling with Gated Recurrent Neural Network for Sentiment Classification,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 1422–1432. doi: 10.18653/v1/D15-1167.

[46] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical Attention Networks for Document Classification,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California: Association for Computational Linguistics, 2016, pp. 1480–1489. doi: 10.18653/v1/N16-1174.

[47] Y. Wang, M. Huang, X. Zhu, and L. Zhao, “Attention-based LSTM for Aspect-level Sentiment Classification,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas: Association for Computational Linguistics, 2016, pp. 606–615. doi: 10.18653/v1/D16-1058.

[48] J. Dai, H. Yan, T. Sun, P. Liu, and X. Qiu, “Does syntax matter? A strong baseline for Aspect-based Sentiment Analysis with RoBERTa,” in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, 2021, pp. 1816–1829. doi: 10.18653/v1/2021.naacl-main.146.

[49] X. Yang, S. Karaman, J. Tetreault, and A. Jaimes, “Journalistic Guidelines Aware News Image Captioning,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 5162–5175. doi: 10.18653/v1/2021.emnlp-main.419.

[50] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, “RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network,” *IEEE Access*, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/ACCESS.2022.3152828.

[51] V. R. Revathy, A. S. Pillai, and F. Daneshfar, “LyEmoBERT: Classification of lyrics’ emotion and recommendation using a pre-trained model,” *Procedia Computer Science*, vol. 218, pp. 1196–1208, 2023, doi: 10.1016/j.procs.2023.01.098.

[52] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, “From Show to Tell: A Survey on Deep Learning-Based Image Captioning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 539–559, Jan. 2023, doi: 10.1109/TPAMI.2022.3148210.

[53] J. Joo, W. Li, F. F. Steen, and S.-C. Zhu, “Visual Persuasion: Inferring Communicative Intents of Images,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA: IEEE, Jun. 2014, pp. 216–223. doi: 10.1109/CVPR.2014.35.

[54] S. Jindal and S. Singh, “Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning,” in *2015 International Conference on Information*

Processing (ICIP), Pune, India: IEEE, Dec. 2015, pp. 447–451. doi: 10.1109/INFOP.2015.7489424.

[55] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[56] L. Wu, S. Liu, M. Jian, J. Luo, X. Zhang, and M. Qi, “Reducing noisy labels in weakly labeled data for visual sentiment analysis,” in *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing: IEEE, Sep. 2017, pp. 1322–1326. doi: 10.1109/ICIP.2017.8296496.

[57] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Mar. 31, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[58] S. Ruan, K. Zhang, L. Wu, T. Xu, Q. Liu, and E. Chen, “Color Enhanced Cross Correlation Net for Image Sentiment Analysis,” *IEEE Trans. Multimedia*, vol. 26, pp. 4097–4109, 2024, doi: 10.1109/TMM.2021.3118208.

[59] A. Yadav and D. K. Vishwakarma, “A deep learning architecture of RA-DLNet for visual sentiment analysis,” *Multimedia Systems*, vol. 26, no. 4, pp. 431–451, Aug. 2020, doi: 10.1007/s00530-020-00656-7.

[60] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, “Weakly Supervised Coupled Networks for Visual Sentiment Analysis,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 7584–7592. doi: 10.1109/CVPR.2018.00791.

[61] Z. Li, H. Lu, C. Zhao, L. Feng, G. Gu, and W. Chen, “Weakly supervised discriminate enhancement network for visual sentiment analysis,” *Artif Intell Rev*, vol. 56, no. 2, pp. 1763–1785, Feb. 2023, doi: 10.1007/s10462-022-10212-6.

[62] G. Meena and K. K. Mohbey, “Sentiment analysis on images using different transfer learning models,” *Procedia Computer Science*, vol. 218, pp. 1640–1649, 2023, doi: 10.1016/j.procs.2023.01.142.

[63] H. Negi, T. Bhola, M. S Pillai, and D. Kumar, “A Novel Approach for Depression Detection Using Audio Sentiment Analysis,” *International Journal of Information Systems & Management Science*, vol. 1, 2018, Accessed: Jul. 28, 2024. [Online]. Available: https://papers.ssrn.com/abstract=3363837

[64] S. Luitel and M. Anwar, “Audio Sentiment Analysis using Spectrogram and Bag-of-Visual- Words,” in *2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI)*, San Diego, CA, USA: IEEE, Aug. 2022, pp. 200–205. doi: 10.1109/IRI54793.2022.00052.

[65] X. Chen, Y. Wang, and Q. Liu, “Visual and textual sentiment analysis using deep fusion convolutional neural networks,” in *2017 IEEE International Conference on Image Processing (ICIP)*, Beijing: IEEE, Sep. 2017, pp. 1557–1561. doi: 10.1109/ICIP.2017.8296543.

[66] R. A. Deshmukh, V. Amati, A. Bhamare, and A. Jadhav, "Visual Sentiment Analysis: An Analysis of Emotions in Video and Audio," in *IoT Based Control Networks and Intelligent Systems*, vol. 789, P. P. Joby, M. S. Alencar, and P. Falkowski-Gilski, Eds., in Lecture Notes in Networks and Systems, vol. 789. , Singapore: Springer Nature Singapore, 2024, pp. 313–326. doi: 10.1007/978-981-99-6586-1_21.

[67] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, "Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis," in *2017 IEEE International Conference on Data Mining (ICDM)*, New Orleans, LA: IEEE, Nov. 2017, pp. 1033–1038. doi: 10.1109/ICDM.2017.134.

[68] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-Dependent Sentiment Analysis in User-Generated Videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 873–883. doi: 10.18653/v1/P17-1081.

[69] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, "Contextual Inter-modal Attention for Multi-modal Sentiment Analysis," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 3454–3466. doi: 10.18653/v1/D18-1382.

[70] A. Agarwal, A. Yadav, and D. K. Vishwakarma, "Multimodal Sentiment Analysis via RNN variants," in *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD)*, Honolulu, HI, USA: IEEE, May 2019, pp. 19–23. doi: 10.1109/BCD.2019.8885108.

[71] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, Sep. 2017, doi: 10.1016/j.imavis.2017.08.003.

[72] N. Xu and W. Mao, "MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Singapore Singapore: ACM, Nov. 2017, pp. 2399–2402. doi: 10.1145/3132847.3133142.

[73] N. Xu, W. Mao, and G. Chen, "A Co-Memory Network for Multimodal Sentiment Analysis," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, Ann Arbor MI USA: ACM, Jun. 2018, pp. 929–932. doi: 10.1145/3209978.3210093.

[74] Z. Zhao *et al.*, "An image-text consistency driven multimodal sentiment analysis approach for social media," *Information Processing & Management*, vol. 56, no. 6, p. 102097, Nov. 2019, doi: 10.1016/j.ipm.2019.102097.

[75] M. Wang, D. Cao, L. Li, S. Li, and R. Ji, "Microblog Sentiment Analysis Based on Cross-media Bag-of-words Model," in *Proceedings of International Conference on Internet Multimedia Computing and Service*, Xiamen China: ACM, Jul. 2014, pp. 76–80. doi: 10.1145/2632856.2632912.

[76] Q. You, J. Luo, H. Jin, and J. Yang, “Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia,” in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, San Francisco California USA: ACM, Feb. 2016, pp. 13–22. doi: 10.1145/2835776.2835779.

[77] N. Xu, “Analyzing multimodal public sentiment based on hierarchical semantic attentional network,” in *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, Beijing, China: IEEE, Jul. 2017, pp. 152–154. doi: 10.1109/ISI.2017.8004895.

[78] N. Xu, W. Mao, and G. Chen, “Multi-Interactive Memory Network for Aspect Based Multimodal Sentiment Analysis,” *AAAI*, vol. 33, no. 01, pp. 371–378, Jul. 2019, doi: 10.1609/aaai.v33i01.3301371.

[79] Y. Ling, J. Yu, and R. Xia, “Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 2149–2159. doi: 10.18653/v1/2022.acl-long.152.

[80] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, “Select-additive learning: Improving generalization in multimodal sentiment analysis,” in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, Hong Kong, Hong Kong: IEEE, Jul. 2017, pp. 949–954. doi: 10.1109/ICME.2017.8019301.

[81] N. Cummins, S. Amiriparian, S. Ottl, M. Gerczuk, M. Schmitt, and B. Schuller, “Multimodal Bag-of-Words for Cross Domains Sentiment Analysis,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB: IEEE, Apr. 2018, pp. 4954–4958. doi: 10.1109/ICASSP.2018.8462660.

[82] F. Chen, R. Ji, J. Su, D. Cao, and Y. Gao, “Predicting Microblog Sentiments via Weakly Supervised Multimodal Deep Learning,” *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 997–1007, Apr. 2018, doi: 10.1109/TMM.2017.2757769.

[83] Z. Li, Y. Fan, W. Liu, and F. Wang, “Image sentiment prediction based on textual descriptions with adjective noun pairs,” *Multimed Tools Appl*, vol. 77, no. 1, pp. 1115–1132, Jan. 2018, doi: 10.1007/s11042-016-4310-5.

[84] D. Zhang, S. Li, Q. Zhu, and G. Zhou, “Modeling the Clause-Level Structure to Multimodal Sentiment Analysis via Reinforcement Learning,” in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, Shanghai, China: IEEE, Jul. 2019, pp. 730–735. doi: 10.1109/ICME.2019.00131.

[85] A. Kumar and G. Garg, “Sentiment analysis of multimodal twitter data,” *Multimed Tools Appl*, vol. 78, no. 17, pp. 24103–24119, Sep. 2019, doi: 10.1007/s11042-019-7390-1.

[86] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, “Image–text sentiment analysis via deep multimodal attentive fusion,” *Knowledge-Based Systems*, vol. 167, pp. 26–37, Mar. 2019, doi: 10.1016/j.knosys.2019.01.019.

[87] J. Xu *et al.*, “Visual-textual sentiment classification with bi-directional multi-level attention networks,” *Knowledge-Based Systems*, vol. 178, pp. 61–73, Aug. 2019, doi: 10.1016/j.knosys.2019.04.018.

[88] J. Yu, J. Jiang, and R. Xia, "Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 429–439, 2020, doi: 10.1109/TASLP.2019.2957872.

[89] K. Yang, H. Xu, and K. Gao, "CM-BERT: Cross-Modal BERT for Text-Audio Sentiment Analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle WA USA: ACM, Oct. 2020, pp. 521–528. doi: 10.1145/3394171.3413690.

[90] A. Kumar and J. Vepa, "Gated Mechanism for Attention Based Multi Modal Sentiment Analysis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain: IEEE, May 2020, pp. 4477–4481. doi: 10.1109/ICASSP40776.2020.9053012.

[91] J. Xu, Z. Li, F. Huang, C. Li, and P. S. Yu, "Social Image Sentiment Analysis by Exploiting Multimodal Content and Heterogeneous Relations," *IEEE Trans. Ind. Inf.*, vol. 17, no. 4, pp. 2974–2982, Apr. 2021, doi: 10.1109/TII.2020.3005405.

[92] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-Text Multimodal Emotion Classification via Multi-View Attentional Network," *IEEE Trans. Multimedia*, vol. 23, pp. 4014–4026, 2021, doi: 10.1109/TMM.2020.3035277.

[93] J. He, S. Mai, and H. Hu, "A Unimodal Reinforced Transformer With Time Squeeze Fusion for Multimodal Sentiment Analysis," *IEEE Signal Process. Lett.*, vol. 28, pp. 992–996, 2021, doi: 10.1109/LSP.2021.3078074.

[94] D. Gu *et al.*, "Targeted Aspect-Based Multimodal Sentiment Analysis: An Attention Capsule Extraction and Multi-Head Fusion Network," *IEEE Access*, vol. 9, pp. 157329–157336, 2021, doi: 10.1109/ACCESS.2021.3126782.

[95] V. Lopes, A. Gaspar, L. A. Alexandre, and J. Cordeiro, "An AutoML-based Approach to Multimodal Image Sentiment Analysis," in *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China: IEEE, Jul. 2021, pp. 1–9. doi: 10.1109/IJCNN52387.2021.9533552.

[96] M. Arjmand, M. J. Dousti, and H. Moradi, "TEASEL: A Transformer-Based Speech-Prefixed Language Model," Sep. 12, 2021, *arXiv*: arXiv:2109.05522. Accessed: Jul. 28, 2024. [Online]. Available: http://arxiv.org/abs/2109.05522

[97] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[98] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[99] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, doi: 10.1109/78.650093.

[100] M. Jaderberg, K. Simonyan, A. Zisserman, and koray kavukcuoglu, “Spatial Transformer Networks,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2015. Accessed: Jul. 28, 2024. [Online]. Available: https://papers.nips.cc/paper_files/paper/2015/hash/33ceb07bf4eeb3da587e268d663aba1a-Abstract.html

[101] V. Mnih, N. Heess, A. Graves, and koray kavukcuoglu, “Recurrent Models of Visual Attention,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2014. Accessed: Jul. 28, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/hash/09c6c3783b4a70054da74f2538ed47c6-Abstract.html

[102] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018, pp. 3–19. doi: 10.1007/978-3-030-01234-2_1.

[103] J. Yu and J. Jiang, “Adapting BERT for Target-Oriented Multimodal Sentiment Classification,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, pp. 5408–5414. doi: 10.24963/ijcai.2019/751.

[104] Z. Zhang, Z. Wang, X. Li, N. Liu, B. Guo, and Z. Yu, “ModalNet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network,” *World Wide Web*, vol. 24, no. 6, pp. 1957–1974, Nov. 2021, doi: 10.1007/s11280-021-00955-7.

[105] J. Yu, K. Chen, and R. Xia, “Hierarchical Interactive Multimodal Transformer for Aspect-Based Multimodal Sentiment Analysis,” *IEEE Trans. Affective Comput.*, vol. 14, no. 3, pp. 1966–1978, Jul. 2023, doi: 10.1109/TAFFC.2022.3171091.

[106] Z. Khan and Y. Fu, “Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation,” in *Proceedings of the 29th ACM International Conference on Multimedia*, Virtual Event China: ACM, Oct. 2021, pp. 3034–3042. doi: 10.1145/3474085.3475692.

[107] “Georgios” “Chochlakis,” “Tejas” “Srinivasan,” “Jesse” “Thomason,” and “Shrikanth” “Narayanan,” “VAuLT: Augmenting the vision-and-language transformer for sentiment classification on social media,” *arxiv*, Oct. 2022, Accessed: Jan. 07, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2208.09021 Focus to learn more

[108] F. Z. Canal *et al.*, “A survey on facial emotion recognition techniques: A state-of-the-art literature review,” *Information Sciences*, vol. 582, pp. 593–617, Jan. 2022, doi: 10.1016/j.ins.2021.10.005.

[109] Y. Khaireddin and Z. Chen, “Facial Emotion Recognition: State of the Art Performance on FER2013”.

[110] A. Azcarate, F. Hageloh, K. V. D. Sande, and R. Valenti, “Automatic facial emotion recognition,” *Universiteit van Amsterdam*, no. June, 2005.

[111] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors (Switzerland)*, vol. 18, no. 2, 2018, doi: 10.3390/s18020401.

[112] H. Jangid, S. Singhal, R. R. Shah, and R. Zimmermann, "Aspect-Based Financial Sentiment Analysis using Deep Learning," in *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, Lyon, France: ACM Press, 2018, pp. 1961–1966. doi: 10.1145/3184558.3191827.

[113] J. Wang *et al.*, "GIT: A Generative Image-to-text Transformer for Vision and Language | Semantic Scholar," *Transactions on Machine Learning Research*, Accessed: Oct. 16, 2023. [Online]. Available: https://www.semanticscholar.org/paper/GIT%3A-A-Generative-Image-to-text-Transformer-for-and-Wang-Yang/60ee030773ba1b68eb222a265b052ca028353362

[114] S. I. Serengil and A. Ozpinar, "LightFace: A Hybrid Deep Face Recognition Framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, Oct. 2020, pp. 1–5. doi: 10.1109/ASYU50717.2020.9259802.

[115] S. I. Serengil and A. Ozpinar, "HyperExtended LightFace: A Facial Attribute Analysis Framework," in *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, Istanbul, Turkey: IEEE, Oct. 2021, pp. 1–4. doi: 10.1109/ICEET53442.2021.9659697.

[116] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, PMLR, Jul. 2021, pp. 8748–8763. Accessed: Oct. 31, 2023. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[117] L. Zhuang, L. Wayne, S. Ya, and Z. Jun, "A Robustly Optimized BERT Pre-training Approach with Post-training," in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, Huhhot, China: Chinese Information Processing Society of China, Aug. 2021, pp. 1218–1227. Accessed: Oct. 25, 2023. [Online]. Available: https://aclanthology.org/2021.ccl-1.108

[118] J. Yu and J. Jiang, "Adapting BERT for Target-Oriented Multimodal Sentiment Classification," 2019. [Online]. Available: https://github.com/jefferyYu/TomBERT.

[119] J. Yu, J. Jiang, and R. Xia, "Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 429–439, 2020, doi: 10.1109/TASLP.2019.2957872.

[120] D. Gu *et al.*, "Targeted Aspect-Based Multimodal Sentiment Analysis: An Attention Capsule Extraction and Multi-Head Fusion Network," *IEEE Access*, vol. 9, pp. 157329–157336, 2021, doi: 10.1109/ACCESS.2021.3126782.

[121] Z. Zhang, Z. Wang, X. Li, N. Liu, B. Guo, and Z. Yu, "ModalNet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network," *World Wide Web*, vol. 24, no. 6, 2021, doi: 10.1007/s11280-021-00955-7.

[122] F. Zhao, C. Li, Z. Wu, S. Xing, and X. Dai, "Learning from Different text-image Pairs: A Relation-enhanced Graph Convolutional Network for Multimodal NER," in *Proceedings of*

the 30th ACM International Conference on Multimedia, Lisboa Portugal: ACM, Oct. 2022, pp. 3983–3992. doi: 10.1145/3503161.3548228.

[123] J. Yu, K. Chen, and R. Xia, "Hierarchical Interactive Multimodal Transformer for Aspect-Based Multimodal Sentiment Analysis," *IEEE Transactions on Affective Computing*, 2022, doi: 10.1109/TAFFC.2022.3171091.

[124] "Target-level Sentiment Analysis Based on Image and Text Fusion | IEEE Conference Publication | IEEE Xplore." Accessed: Oct. 16, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9953248

[125] Z. Khan and Y. Fu, "Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation," in *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia,* Association for Computing Machinery, Inc, Oct. 2021, pp. 3034–3042. doi: 10.1145/3474085.3475692.

[126] J. Wang *et al.*, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 43, no. 10, pp. 3349–3364, Oct. 2021, doi: 10.1109/TPAMI.2020.2983686.

[127] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Robust Face Recognition for Uncontrolled Pose and Illumination Changes," *IEEE Trans. Syst. Man Cybern, Syst.*, vol. 43, no. 1, pp. 149–163, Jan. 2013, doi: 10.1109/TSMCA.2012.2192427.

[128] A. Castiglione, P. Vijayakumar, M. Nappi, S. Sadiq, and M. Umer, "COVID-19: Automatic Detection of the Novel Coronavirus Disease From CT Images Using an Optimized Convolutional Neural Network," *IEEE Trans. Ind. Inf.*, vol. 17, no. 9, pp. 6480–6488, Sep. 2021, doi: 10.1109/TII.2021.3057524.

[129] A. Pandey and D. K. Vishwakarma, "Attention-based Model for Multi-modal sentiment recognition using Text-Image Pairs," in *2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT)*, IEEE, Mar. 2023. doi: 10.1109/ICITIIT57246.2023.10068626.

[130] A. Yadav and D. K. Vishwakarma, "AW-MSA: Adaptively weighted multi-scale attentional features for DeepFake detection," *Engineering Applications of Artificial Intelligence,* vol. 127, p. 107443, Jan. 2024, doi: 10.1016/j.engappai.2023.107443.

[131] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," presented at the ICLR 2021, ICLR, Jun. 2021.

[132] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, p. 140:5485-140:5551, Jan. 2020.

[133] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-Attention with Relative Position Representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers),* New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 464–468. doi: 10.18653/v1/N18-2074.

[134] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning,*

PMLR, Jul. 2021, pp. 8748–8763. Accessed: Apr. 05, 2024. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[135] Z. Ebrahimian, R. Toosi, and M. A. Akhaee, “Multinomial Emoji Prediction Using Deep Bidirectional Transformers and Topic Modeling,” in *2022 30th International Conference on Electrical Engineering (ICEE)*, May 2022, pp. 272–277. doi: 10.1109/ICEE55646.2022.9827247.

[136] F. Barbieri, M. Ballesteros, F. Ronzano, and H. Saggion, “Multimodal Emoji Prediction,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 2018, pp. 679–686. doi: 10.18653/v1/N18-2107.

[137] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of Tricks for Efficient Text Classification,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Association for Computational Linguistics, Apr. 2017, pp. 427–431. Accessed: Apr. 08, 2024. [Online]. Available: https://aclanthology.org/E17-2068

[138] F. Barbieri, M. Ballesteros, and H. Saggion, “Are Emojis Predictable?,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Association for Computational Linguistics, Apr. 2017, pp. 105–111. Accessed: Apr. 08, 2024. [Online]. Available: https://aclanthology.org/E17-2017

[139] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[140] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” Sep. 11, 2020, *arXiv*: arXiv:1905.11946. Accessed: May 22, 2024. [Online]. Available: http://arxiv.org/abs/1905.11946

[141] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, “Aggregated Residual Transformations for Deep Neural Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 5987–5995. doi: 10.1109/CVPR.2017.634.

[142] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollar, “Designing Network Design Spaces,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, Jun. 2020, pp. 10425–10433. doi: 10.1109/CVPR42600.2020.01044.

[143] “A Robustly Optimized BERT Pre-training Approach with Post-training - ACL Anthology.” Accessed: May 11, 2023. [Online]. Available: https://aclanthology.org/2021.ccl-1.108/

[144] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, 1997, doi: 10.1109/78.650093.

[145] J. Xu, Y. Pan, X. Pan, S. Hoi, Z. Yi, and Z. Xu, “RegNet: Self-Regulated Network for Image Classification,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–6, 2022, doi: 10.1109/TNNLS.2022.3158966.

[146] Q.-L. Zhang and Y.-B. Yang, “SA-Net: Shuffle Attention for Deep Convolutional Neural Networks,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 2235–2239. doi: 10.1109/ICASSP39728.2021.9414568.

[147] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, “Towards multimodal sarcasm detection (an obviously perfect paper),” in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2020. doi: 10.18653/v1/p19-1455.

[148] Y. Wu *et al.*, “Modeling Incongruity between Modalities for Multimodal Sarcasm Detection,” *IEEE Multimedia*, vol. 28, no. 2, 2021, doi: 10.1109/MMUL.2021.3069097.

[149] A. Ray, S. Mishra, A. Nunna, and P. Bhattacharyya, “A Multimodal Corpus for Emotion Recognition in Sarcasm,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 6992–7003. Accessed: May 09, 2023. [Online]. Available: https://aclanthology.org/2022.lrec-1.756

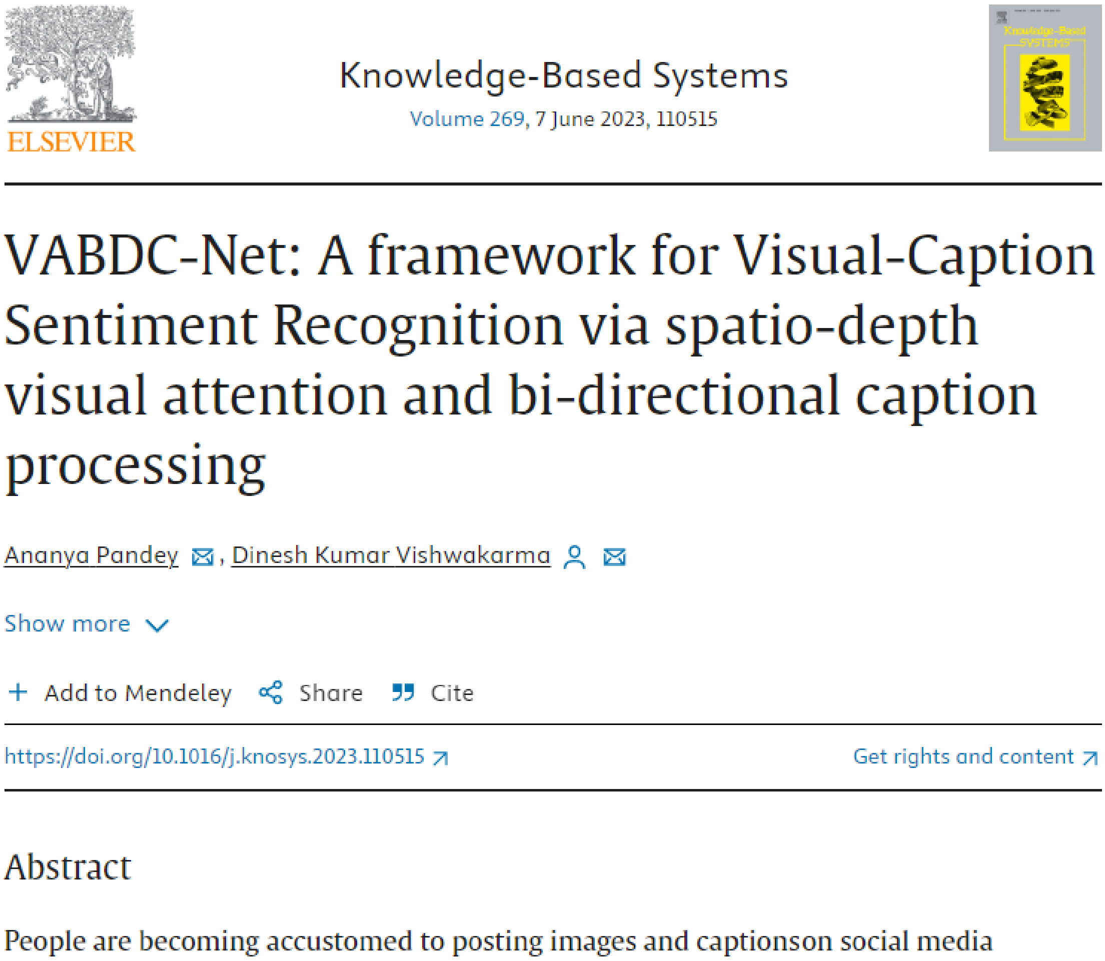
[150] Y. Zhang *et al.*, “CFN: A Complex-Valued Fuzzy Network for Sarcasm Detection in Conversations,” *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 12, 2021, doi: 10.1109/TFUZZ.2021.3072492.

[151] N. Ding, S. wei Tian, and L. Yu, “A multimodal fusion method for sarcasm detection based on late fusion,” *Multimedia Tools and Applications*, vol. 81, no. 6, 2022, doi: 10.1007/s11042-022-12122-9.

PROOF OF PUBLICATIONS

SCIE Journal Paper 1:

❖ **A. Pandey** and D. K. Vishwakarma, “VABDC-Net: A framework for Visual-Caption Sentiment Recognition via spatio-depth visual attention and bi-directional caption processing,” ***Knowledge-Based Systems***, vol. 269, June. 2023, doi: https://doi.org/10.1016/j.knosys.2023.110515.



Knowledge-Based Systems
Volume 269, 7 June 2023, 110515

VABDC-Net: A framework for Visual-Caption Sentiment Recognition via spatio-depth visual attention and bi-directional caption processing

Ananya Pandey, Dinesh Kumar Vishwakarma

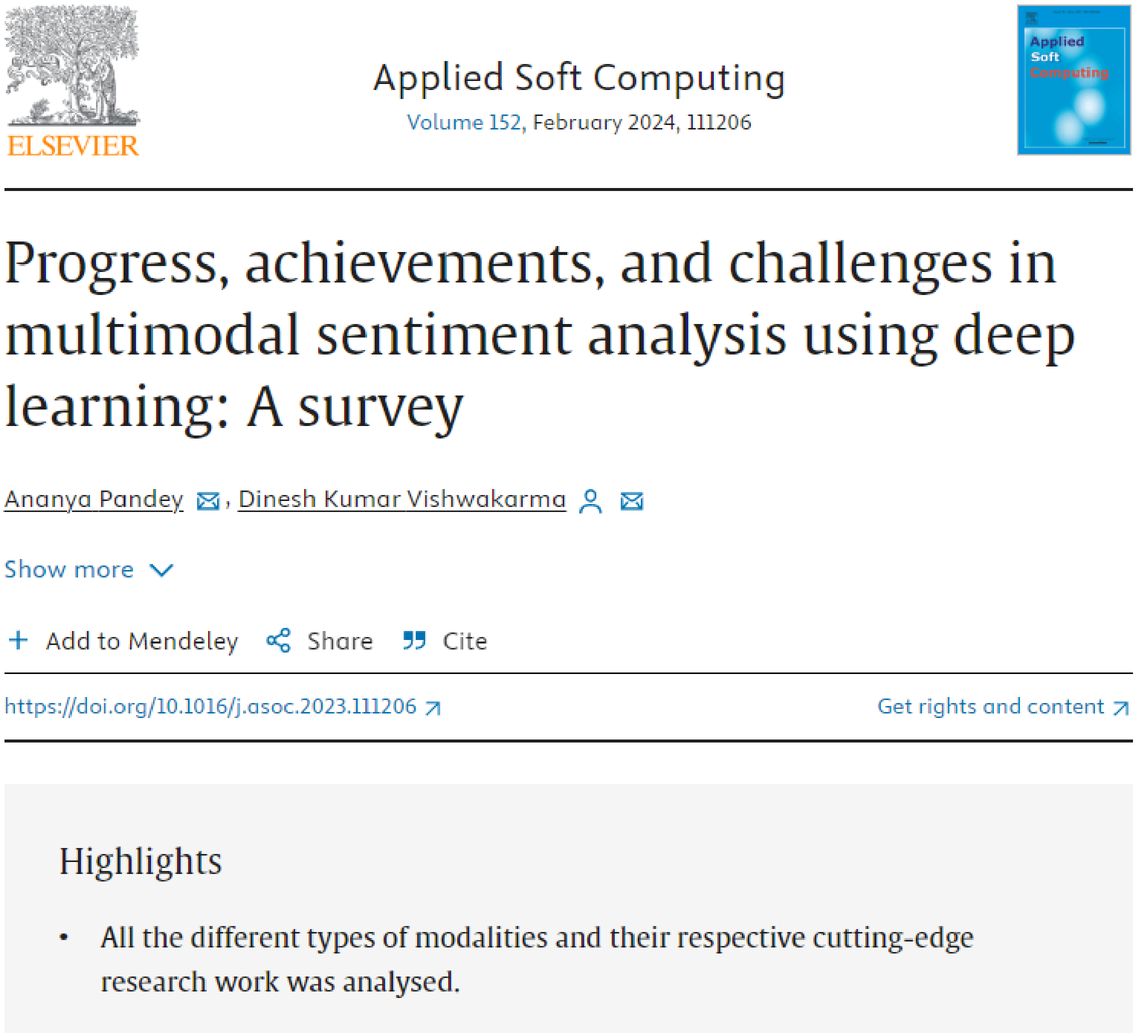
https://doi.org/10.1016/j.knosys.2023.110515

Abstract

People are becoming accustomed to posting images and captionson social media

SCIE Journal Paper 2:

❖ **A. Pandey** and D. K. Vishwakarma, "Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey," ***Applied Soft Computing***, vol. 152, November. 2024, doi: https://doi.org/10.1016/j.asoc.2023.111206.



ELSEVIER

Applied Soft Computing

Volume 152, February 2024, 111206

Progress, achievements, and challenges in multimodal sentiment analysis using deep learning: A survey

Ananya Pandey ✉, Dinesh Kumar Vishwakarma 👤 ✉

Show more ⌄

+ Add to Mendeley   Share   Cite

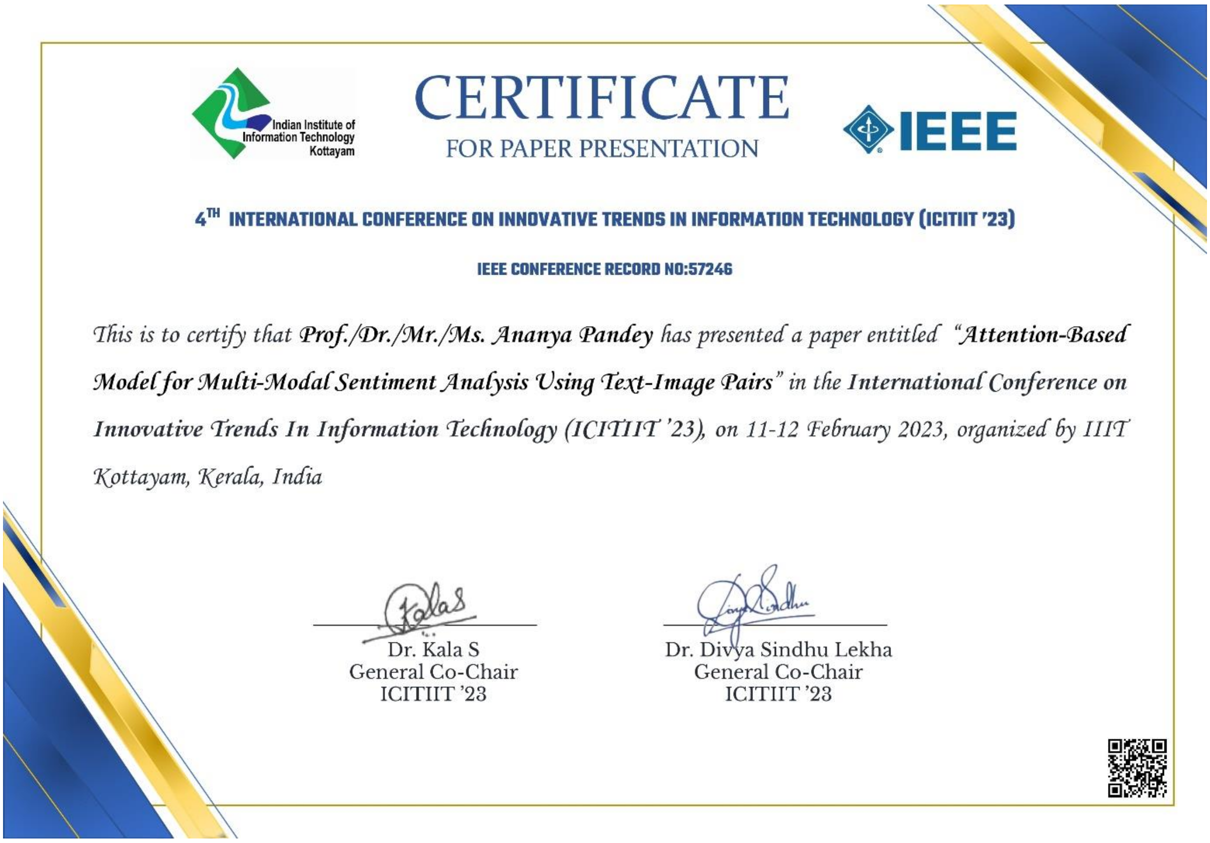
https://doi.org/10.1016/j.asoc.2023.111206 ↗   Get rights and content ↗

Highlights

- All the different types of modalities and their respective cutting-edge research work was analysed.

Conference Paper 1:

- **A. Pandey** and D. K. Vishwakarma, "Attention-based Model for Multi-modal sentiment recognition using Text-Image Pairs," in *2023* ***4th International Conference on Innovative Trends in Information Technology (ICITIIT)***, Institute of Electrical and Electronics Engineers (IEEE), March. 2023, pp. 1–5. doi: 10.1109/ICITIIT57246.2023.10068626.



Indian Institute of Information Technology Kottayam

CERTIFICATE

FOR PAPER PRESENTATION

IEEE

4TH INTERNATIONAL CONFERENCE ON INNOVATIVE TRENDS IN INFORMATION TECHNOLOGY (ICITIIT '23)

IEEE CONFERENCE RECORD NO:57246

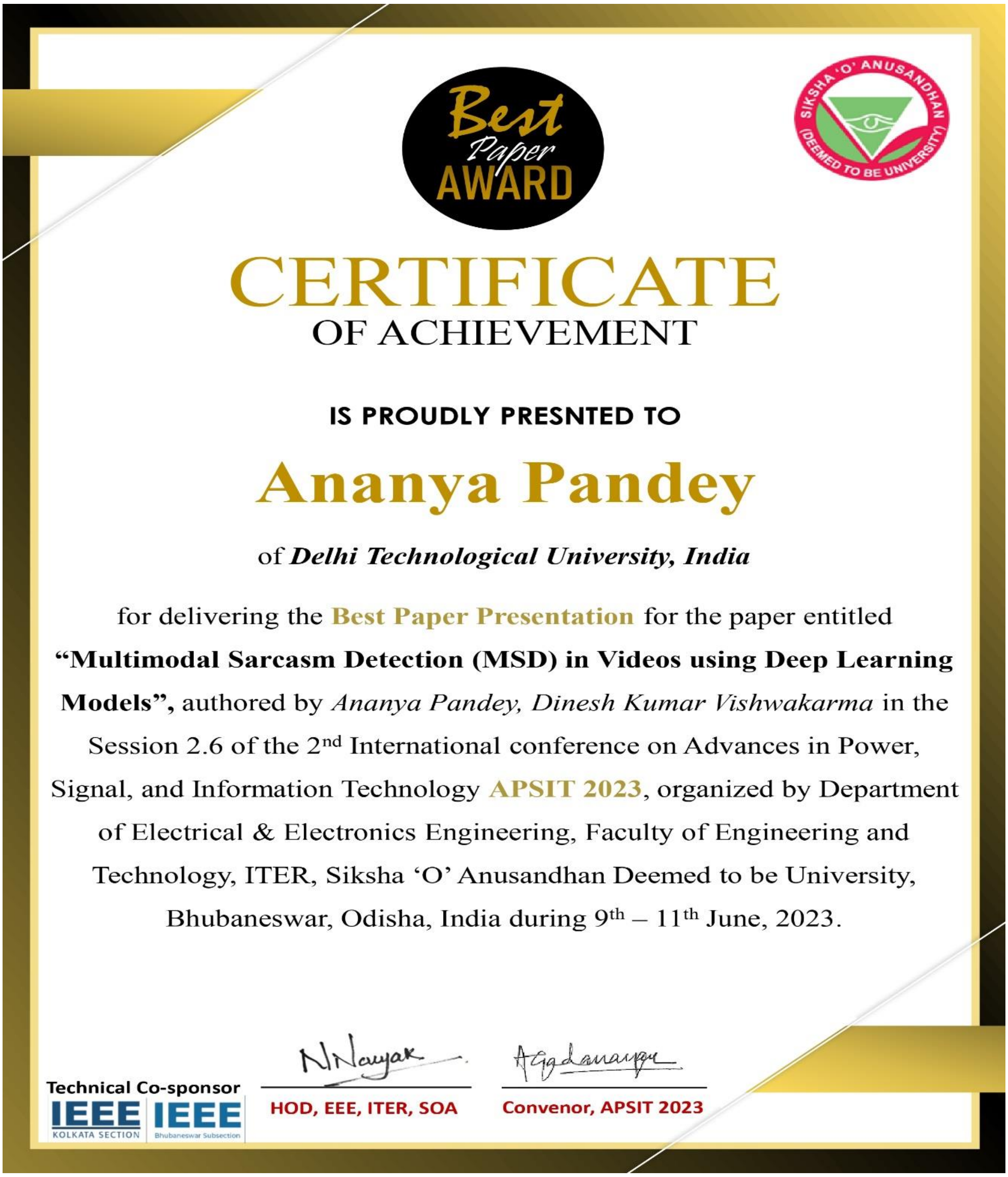
This is to certify that ***Prof./Dr./Mr./Ms. Ananya Pandey*** *has presented a paper entitled* ***"Attention-Based Model for Multi-Modal Sentiment Analysis Using Text-Image Pairs"*** *in the* ***International Conference on Innovative Trends In Information Technology (ICITIIT '23),*** *on 11-12 February 2023, organized by IIIT Kottayam, Kerala, India*

Dr. Kala S
General Co-Chair
ICITIIT '23

Dr. Divya Sindhu Lekha
General Co-Chair
ICITIIT '23

Conference Paper 2:

- **A. Pandey** and D. K. Vishwakarma, “Multimodal Sarcasm Detection (MSD) in Videos using Deep Learning Models,” in ***2023 IEEE International Conference in Advances in Power, Signal, and Information Technology (APSIT)***, Institute of Electrical and Electronics Engineers (IEEE), Jun. 2023, pp. 811-814. doi: 10.1109/APSIT58554.2023.10201731.



Best Paper AWARD
SIKSHA 'O' ANUSANDHAN (DEEMED TO BE UNIVERSITY)

CERTIFICATE

OF ACHIEVEMENT

IS PROUDLY PRESNTED TO

Ananya Pandey

of ***Delhi Technological University, India***

for delivering the **Best Paper Presentation** for the paper entitled **“Multimodal Sarcasm Detection (MSD) in Videos using Deep Learning Models”,** authored by *Ananya Pandey, Dinesh Kumar Vishwakarma* in the Session 2.6 of the 2nd International conference on Advances in Power, Signal, and Information Technology **APSIT 2023**, organized by Department of Electrical & Electronics Engineering, Faculty of Engineering and Technology, ITER, Siksha ‘O’ Anusandhan Deemed to be University, Bhubaneswar, Odisha, India during 9th – 11th June, 2023.

NNayak

HOD, EEE, ITER, SOA

Convenor, APSIT 2023

Technical Co-sponsor

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis: DESIGN OF FRAMEWORK FOR SENTIMENT ANALYSIS USING DEEP LEARNING

Total Pages: 143

Name of the Scholar: Ananya Pandey

Supervisor: Prof. Dinesh Kumar Vishwakarma

Department: Information Technology

This is to report that the above thesis was scanned for similarity detection. The process and outcome are given below:

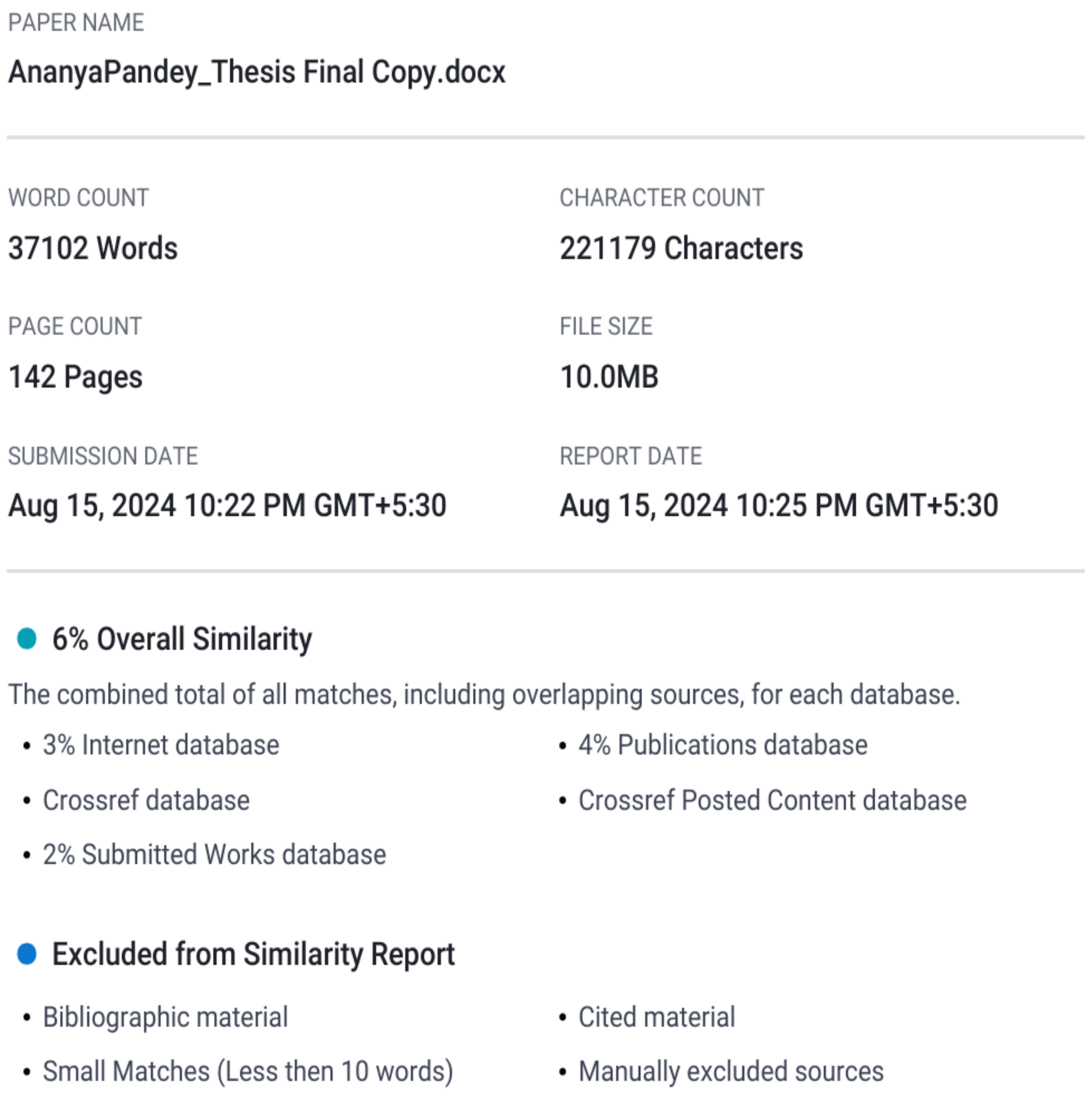
Software used: Turnitin Similarity Index: 6% Word Count: 37102 Words

Date: 10/08/2024

Ananya

Candidate's Signature **Signature of Supervisor**

PLAGIARISM REPORT



PAPER NAME

AnanyaPandey_Thesis Final Copy.docx

WORD COUNT

37102 Words

CHARACTER COUNT

221179 Characters

PAGE COUNT

142 Pages

FILE SIZE

10.0MB

SUBMISSION DATE

Aug 15, 2024 10:22 PM GMT+5:30

REPORT DATE

Aug 15, 2024 10:25 PM GMT+5:30

● 6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 3% Internet database
- Crossref database
- 2% Submitted Works database
- 4% Publications database
- Crossref Posted Content database

● Excluded from Similarity Report

- Bibliographic material
- Small Matches (Less then 10 words)
- Cited material
- Manually excluded sources

Author Biography

Ananya Pandey received Bachelor of Technology in Computer Science & Engineering degree in 2018 from Jamia Hamdard University, New Delhi and Master of Technology in Computer Engineering degree in 2020 from Jamia Millia Islamia University, New Delhi, India. She has successfully qualified UGC-NET with JRF in 2020 and GATE 2021, demonstrating her exceptional academic and research capabilities. She is a senior research scholar at the Department of Information Technology, Delhi Technological University, New Delhi. The topic of her doctoral dissertation is "***Design of Framework for Sentiment Analysis using Deep Learning***". She is primarily interested in deep-learning and computer vision-based multimodal tasks, such as, Multimodal Sentiment Analysis, Multimodal Sarcasm Detection, etc. The primary objective is to identify sentiments by employing multiple modalities.