Semantic Disambiguation Transformer for Multimodal Emotion Understanding

1st Indhumathi S
Research Scholar
Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India.
indhume11@gmail.com

2nd F. Mary Harin Fernandez,
Professor,
Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India.
mary.fherin@gmail.com

***Abstract*— The fast evolution of social media platforms, multimodal content combining textual and visual information has become a foremost part for expressing emotions and opinions. Precisely understanding such content is challenging due to the presence of implicit semantics, contextual ambiguity, and multimodal relations. A Semantic Disambiguation Transformer for fine-grained multimodal emotion analysis, which integrates contextual semantic learning with multimodal feature fusion is proposed. The framework employs a BERT-based encoder for textual representation and a ResNet-50 architecture for visual feature extraction. To improve semantic understanding, a word sense disambiguation mechanism is combined through pretraining on SemEval datasets. This fusion approach is used to align multimodal features into a shared embedding space, followed by a multi-task learning framework for predicting humour, sarcasm, offensiveness, motivation, sentiment, and hate. The model is assessed using accuracy, macro F1-score, and AUC, supported by visualization techniques such as confusion matrices, ROC curves, precision recall curves, and semantic heatmaps. Experimental results demonstrate the effectiveness of the proposed approach in capturing multimodal semantics and improving emotion classification performance.**

Keywords*— Multimodal emotion analysis, semantic disambiguation transformer, deep learning for affective computing, multimodal fusion

I. INTRODUCTION

Social networking has seen an exponential growth trend in current years, changing the dynamics of human interaction, information broadcasting, and opinion formation in the digital age. Among many types of content that present in online, memes represent one of the most popular ways of expressing feelings, humor, sarcasm, and sociopolitical comment. Memes do not represent any conventional mode of expression, as memes include both text and visual components. This combination makes memes more complex than simple texts or images because memes can carry contextual meaning. Communicating sentiments through memes is relatively easy, but analyzing the same attitudes is a challenge to automatic systems due to the complexity associated with multimodal meme understanding [1]. The current models used for sentiment analysis usually depend on unimodal input such as text-based reviews and single-mode inputs like images. Although this technique is quite successful in dealing with organized domains, it fails when it comes to managing multimodal inputs where the meaning is generated from the interaction between different modes. In some cases, the textual part of a meme might indicate a neutral or positive sentiment, but the added image can be sarcastic or opposite so the sentiment in the textual part, resulting in a different emotional perception. This requires more sophisticated models that can analyze both modalities simultaneously.

With the advancement in deep learning techniques, the ability of natural language processing models to understand contextual semantics in textual input has been improved greatly. The Transformer model is one of the most famous architectures that have gained popularity due to its exceptional abilities in handling the context of words and sentences using a bidirectional attention mechanism [1]. On the other hand, convolutional neural network architectures like ResNet have shown outstanding efficiency in understanding hierarchical visual semantics of an image [2]. However, these advances have been applied using rather simplistic feature fusion methods such as concatenation and late fusion and without considering the semantic ambiguity present in the language used. As pointed out earlier, this presents a major challenge in the case of memes because words usually have different meanings in varying contexts. For instance, a seemingly positive word could be used sarcastically when combined with an image. The use of semantic disambiguation methods would help overcome this difficulty and facilitate accurate understanding of language used.

Word sense disambiguation (WSD) is one of the problems addressed in NLP with regard to understanding the meaning of a word based on its context. Integrating WSD into multimodal systems would greatly improve semantics and help in the identification of the true meaning of texts which might be quite subtle. In this regard, the current multimodal systems used in emotion detection fail to apply semantic disambiguation techniques. This leads to their inability to effectively detect nuances like sarcasm and humor. In response to these challenges, a Semantic Disambiguation Transformer for Multimodal Emotion Analysis is presented in this paper. As the name suggests, this is an innovative system that fuses context-based semantic learning with the feature fusion of multiple modes. The main principle behind this approach is based on improving the quality of textual representations by adding semantic disambiguation using pre-training on SemEval datasets. With the help of such a pre-trained neural network, the task of interpreting potentially ambiguous textual descriptions in combination with other information becomes easier.

The model proposed uses a text encoder, similar to BERT but adapted to extract contextual textual embeddings, along with an image encoder, based on the ResNet neural network. In order to fuse visual and textual data, visual features are first translated to a common embedding space. After that, the combined representation is sent to the multi-task learning framework. It is capable of predicting a number of emotion-related characteristics including humour, sarcasm, offensiveness, motivation, sentiment, and hate detection,

which are commonly studied in multimodal meme understanding tasks [4].

One of the major contributions of this work is through the addition of the auxiliary semantic disambiguation component, using SemEval dataset for word sense disambiguation in the pretraining process. This helps the model understand the semantics of text input in order to capture the subtle differences between the semantic meaning of the text to improve classification accuracy for complicated emotion classes. Unlike other conventional works where separate models are used for processing text and images individually, our framework emphasizes semantic consistency and contextual understanding when it comes to multimodality. Traditional multimodal sentiment analysis approaches have addressed key challenges in combining textual and visual modalities [5].

Summary of the contributions made in this work include:

- A new multimodal framework that incorporates semantic disambiguation to deal with ambiguities in text analysis.
- An innovative multimodal learning architecture involving transformer encoding of text data and CNN-based encoding of visual data.
- Multi-task prediction of emotion attributes in a single learning framework.
- Semantic disambiguation via SemEval pretraining in the framework to improve contextual text comprehension and reasoning abilities.
- Evaluation of the framework on the basis of accuracy, macro F1-score, AUC score, and several visualization techniques including confusion matrix, ROC curve, PR curve, and attention heatmap visualization. It has been shown through extensive experiments that the proposed approach produces excellent results on several emotion classification problems especially for the offensive and motivational classes. Though there are still some difficulties associated with recognition of subjective emotions such as sarcasm and humor, semantic disambiguation helps considerably in achieving better understanding of multimodal interactions. There are several potential real-world applications of the proposed approach. Content moderation, social media and sentiment analysis, and detection of misrepresentation can benefit from the results obtained in this study. By developing an improved approach to multimodal emotion analysis, this research contributes to the creation of more advanced artificial intelligence systems.

II. RELATED WORK

Multimodal emotion analysis in memes is a growing field owing to the elaborate relationship between text and image in meme processing. Some benchmark datasets and approaches based on deep learning techniques have been devised to tackle this problem. The paper “SemEval-2020 Task 8: Memotion Analysis – The Visuo-Lingual Metaphor” [3] presents the MEMotion dataset as a benchmarking dataset for multimodal sentiment and emotion analysis. The paper focuses on the difficulties faced while analyzing memes which use textual sarcasm and visual metaphors. The presence of multiple classes such as humor, sarcasm, offensive language, and sentiment in the dataset makes it useful for fine-grained emotion analysis. This is a pioneering work in the field of meme interpretation. “MemoSYS at SemEval-2020 Task 8: Multimodal Emotion Analysis in Memes” [6] presents a novel hybrid model for classifying emotions in memes using a BERT embedding for textual data and convolutional layers for extracting visual features. It establishes the effectiveness of multimodal feature fusion techniques by proving that multimodal features yield better results than unimodal features. However, it fails to address semantic ambiguity in textual information. Moreover, the MultiOFF dataset, which has been introduced in [7], offers a standard test for identifying offensive language from multimodal memes, including both textual and visual elements. This dataset once again shows the significance of multimodal learning in the recognition of harmful or offensive memes. In the paper “Utilizing BERT and DenseNet for Internet Meme Emotion Analysis” [8], BERT has been utilized in combination with DenseNet for the purpose of text encoding and image feature extraction, respectively. This research places significance on the power of deep visual embeddings in capturing contextual cues. Though there is an improvement in performance, the system is non-transparent and does not account for disambiguation semantics.

The research “Using BERT to Analyse Meme Emotions” [9] investigates the potential of transformer architectures in dealing with textual cues in memes. As seen from the experiment outcomes, BERT is capable of understanding context and semantics efficiently. However, lack of visuals in the process impedes the possibility of understanding multimodal cues. As evidenced by the research “Findings of Memotion 2: Sentiment and Emotion Analysis of Memes” [10], the baseline models employing BERT and ResNet are quite efficient in emotion recognition due to multimodal integration. Despite the improvement of performance, word sense ambiguity is not addressed in the research, which is significant for sarcasm recognition.

Based on the literature, the following gaps can be observed:

- Semantic disambiguation in multimodal systems is missing
- Hardly any efforts to interpret ambiguous texts, like sarcasm, have been made
- Pre-training methods for word sense disambiguation are missing

The current study attempts to fill the gaps mentioned above by using a semantic disambiguation component alongside multimodal transformers.

III. METHODOLOGY

The Semantic Disambiguation Transformer follows the multistage pipeline design which aims to effectively incorporate information from various modes to ensure efficient emotion recognition at fine granular level. The framework consists of four major stages, namely data pre-processing, feature extraction, semantic disambiguation, and multimodal classification.

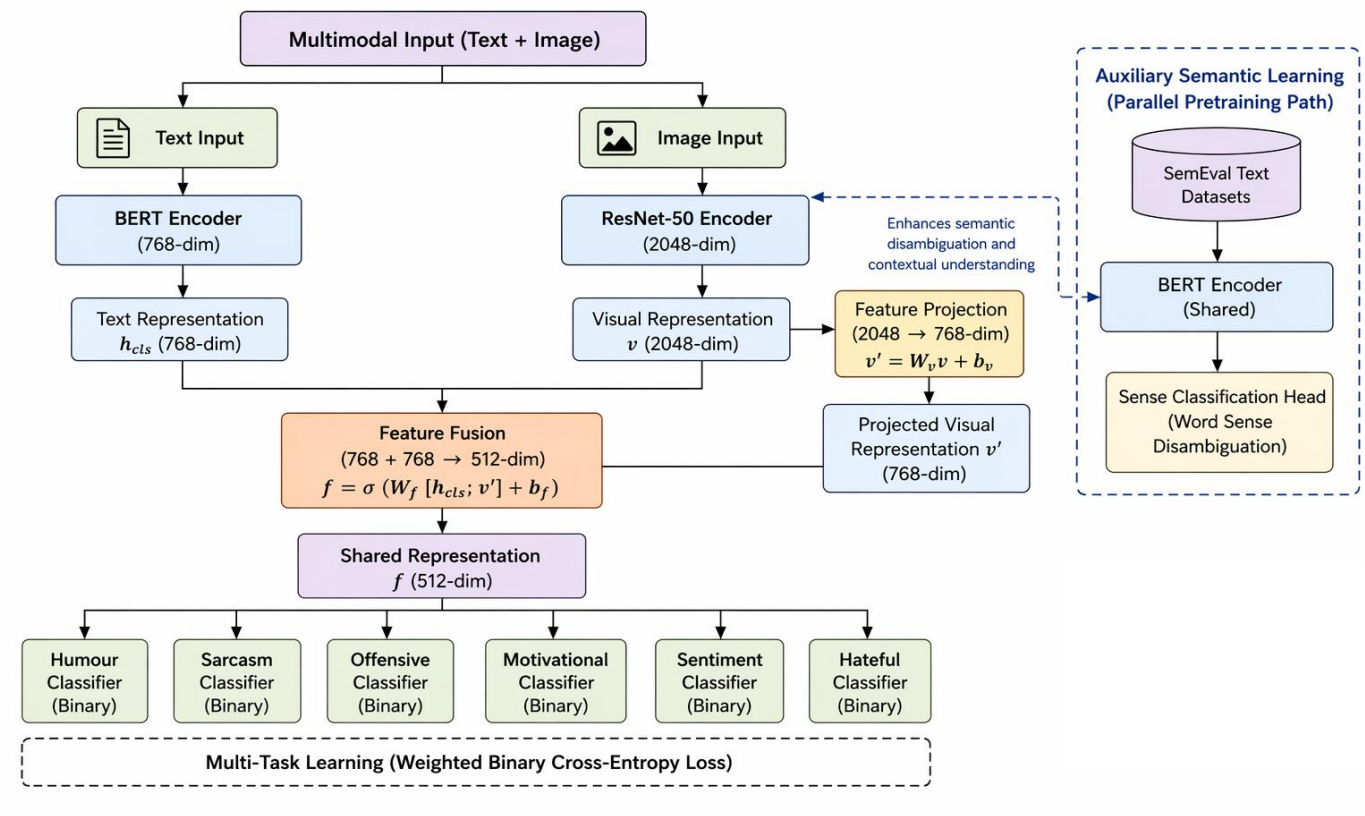

Fig 1. Framework of Semantic Disambiguation Transformer for Multimodal Emotion Analysis

As shown in the figure 1, the framework accepts multimodal input in form of text and images. The text representations extracted using a BERT framework, while the latter involves visual features extracted using a ResNet-50 architecture. The higher dimensional visual features are projected to the same dimension as the dimension of the obtained text features, followed by combining them via a feature fusion layer which produces a unified representation that incorporates semantic and visual characteristics. In addition to that, a semantic learning module is incorporated in the model in form of pre-training of the BERT model on SemEval datasets to enhance word sense interpretation. Multiple classification heads are used in the presented solution to perform tasks related to predicting humour, sarcasm, offensive content, motivational content, sentiment, and hateful content.

A. Data Preprocessing and Feature Extraction

MEMotion dataset is applied as the primary source of data for training and testing purposes, while the hateful memes dataset and SemEval datasets is used for additional pretraining for improving model robustness [11]. Data preprocessing procedure consists of several crucial steps that allow reaching high-quality results. Firstly, textual data undergoes cleaning, namely removal of missing values and normalization of inputs. Secondly, categorical emotions are converted into multi-class or binary format. Emotions such as humor, sarcasm, offensive language, and motivation are encoded in a specific way to perform multi-task learning. To validate our model reliably, the dataset is split into two parts in proportions of 80% and 20%. Moreover, weighted loss function is implemented for class imbalance.

Textual and visual features are obtained with the help of a hybrid approach. In particular, text features learnt by using a pretrained BERT encoder. More precisely, contextualized token-level embeddings is produced, and the CLS token will be chosen to form a sentence-level representation of input sequence. As for visual features, ResNet-50 architecture is used as a feature extractor from meme images. It implies that a convolutional network is trained on images to produce discriminative features. This combination of approach gives effective representation of multimodal data for emotion classification tasks.

B. Semantic Disambiguation

A disambiguation module that makes use of contextual representation of text learned from pre-training on SemEval data sets is included in the proposed architecture [12]. Contextual representation of text enables a system to differentiate among the different senses of a word based on the surrounding context. An input sentence denoted as a sequence of tokens, $X = \{x_1, x_2, \ldots, x_n\}$, as input to a BERT encoder, contextual embeddings $H = \{h_1, h_2, \ldots, h_n\}$, $h_i \in \mathbb{R}^d$ is generated such that each h_i represents the contextual sense of token x_i.

$$P(s \mid x_i) = \text{softmax}(W_s h_i + b_s)$$

where $P(s \mid x_i)$ represents the probability distribution over possible senses s, W_s and b_s are learnable parameters, and h_i is the contextual embedding of token x_i. For sentence-level disambiguation, the [CLS] token embedding h_{cls} is used to denote the overall semantic context. Once this representation is created, it is then included within the framework of the

multimodal system, contributing to an improvement of the model's capabilities to analyze ambiguous statements like sarcasm, humor, and implicit emotions. The result of incorporating this knowledge into the model is a better capability to understand context.

C. Multimodal Fusion

For efficient integration of text and image data, a multimodal fusion technique is used by combining the context embeddings of the BERT model with the features of the ResNet-50 model. Let $h_{cls} \in \mathbb{R}^d$ represents the textual representation obtained from the [CLS] token of BERT, and $v \in \mathbb{R}^k$ denotes the visual feature vector taken from the CNN. Since these features initiate from different modalities, the visual representation is first given into the same embedding space as the textual features using a linear transformation:

$$v' = W_v v + b_v$$

where $W_v \in \mathbb{R}^{dtimesk}$ and $b_v \in \mathbb{R}^d$ are learnable parameters. The projected visual features v' are then fused with the textual representation through concatenation of both:

$$z = [h_{cls}\,;\; v']$$

The fused representation z is passed through a fully connected layer with non-linear activation to obtain a joint multimodal embedding:

$$f = \sigma(W_f z + b_f)$$

where W_f and b_f are trainable parameters, and σ denotes a non-linear activation function such as ReLU. This fused feature vector f captures both semantic and visual cues and helps as the shared representation for all classification tasks [13].

D. Multi-Task Learning Framework

The proposed framework accepts a multi-task learning paradigm, where the fused multimodal representation f is concurrently used to find the multiple emotion-related attributes, including humour, sarcasm, offensive content, motivational tone, sentiment, and hateful content. Each task is modeled with a dedicated classification head, making the network to learn task-specific patterns while sharing a common representation [14]. For a given task t, the prediction is calculated as:

$$P_t(y \mid f) = \text{softmax}(W_t f + b_t)$$

where W_t and b_t are learnable parameters matching to task t. To address class imbalance, a weighted cross-entropy loss is utilized for each task, defined as:

$$\mathcal{L}_t = -\sum_{c=1}^{C} w_c \; y_c \log\left(P_t(y_c \mid f)\right)$$

where w_c denotes the class weight for class c. The overall multi-task objective is framed as the sum of individual task losses:

$$\mathcal{L}_{total} = \sum_{t \in T} \mathcal{L}_t$$

Initially, training is done in two stages. The model is pretrained in the first stage using the SemEval dataset. This helps the model learn semantic disambiguation, helping understand ambiguous words in context. Multimodal training is done on the MEMotion and Hateful Memes datasets in the second stage. The model learns how to perform all tasks in one unified framework using the shared representation. Training is completed using the Adam optimizer with suitable learning rate scheduling and mini-batch training. Evaluating the model performance is done using measures including accuracy, macro F1-score, and the ROC Curve (AUC). These measures provide a complete evaluation of classification performance with class imbalance. Additionally, visualization methods are used to evaluate the model and its interpretability which includes confusion matrix for each task, measuring classification performance, ROC curves measuring discrimination capabilities, precision-recall curves when working under class imbalance, loss curves monitoring model convergence during training, and attention heat maps that measure the role of textual tokens.

IV. RESULT AND DISCUSSION

The Semantic Disambiguation Transformer is proposed to exhibits excellent performance on various tasks related to emotion recognition, as shown through the results obtained. The evaluation process shows that the model is capable of recognizing multimodal semantics where emotional information is explicitly indicated. The accuracy and F1-score show the steady performance for all types of emotions, with motivation and offense achieving the highest accuracy at 95.42% and 87.99%, respectively. Humour and sarcasm, on the other hand, demonstrate relatively less performance due to their subjective and contextual nature, with accuracy scores of 73.63% and 72.55%. The AUC value of 0.98 for hateful content recognition reveals the model's effectiveness in identifying harmful and sensitive information.

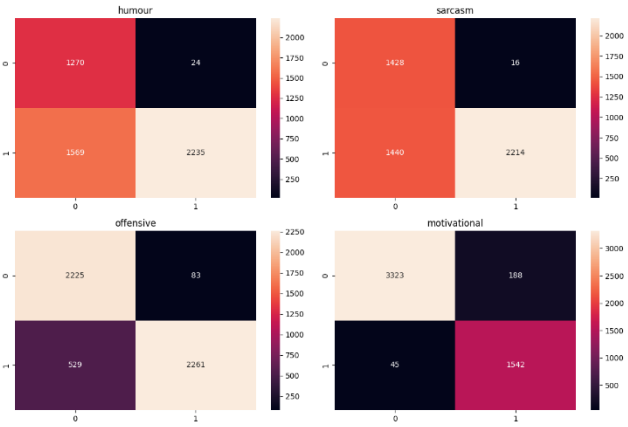

Fig 2. Confusion Matrix for Multimodal Emotion Classification Tasks

The confusion matrix (figure 2) demonstrates the classification accuracy of each category, such as humor, sarcasm, offensive content, and motivational messages. It appears that the model performs well in distinguishing motivational and offensive classes with high true positives. Yet, many false negatives exist in humor and sarcasm classes, implying that subtle semantic implications are still difficult to interpret. This is not surprising because irony and humor are not easily comprehensible in any context. The ROC curves clearly show the ability to classify with great efficiency, as the curves are close to the upper left corner, especially when dealing with motivation and offensive tasks. The relatively straight and steep curves indicate that the model has good generalization ability regardless of the threshold level, while the lesser curve in humor and sarcasm suggests that they have some semantic ambiguity. As with the ROC curves, the precision-recall graphs further show that the proposed method is very efficient in addressing class imbalance issues. Motivational and offensive tasks are consistent at having higher precision values than the other two categories. Humor and sarcasm, on the other hand, have a gentle curve. In the figure 3, False Positive Rate is indicated on the x-axis of the ROC curve, whereas the True Positive Rate is denoted on the y-axis. On the other hand, the Precision–Recall curve has its x-axis labeled with Recall, and the y-axis indicates Precision.

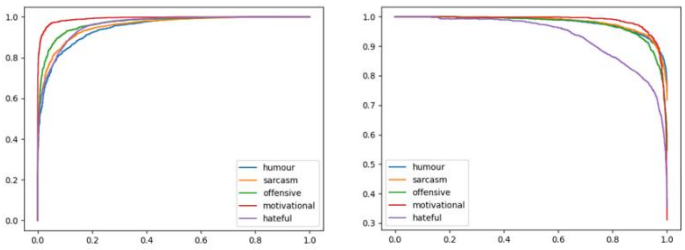

Fig 3. ROC and PR curves

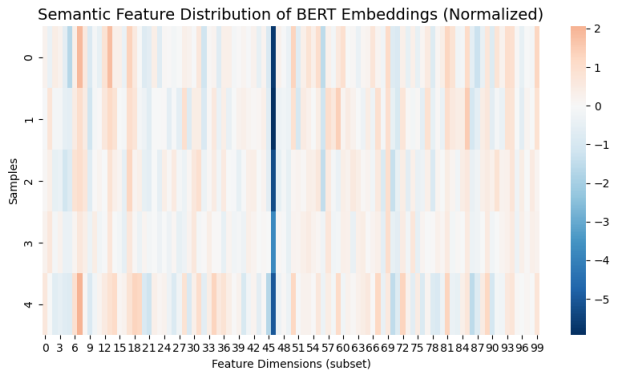

Fig 4. Semantic Feature Distribution of BERT Embeddings

The heatmap (figure 4) demonstrates the normalized semantic representations learned through the BERT model. As different activation patterns emerge along feature dimensions, it is evident that the model learns meaningful semantic representation through contextual embeddings. Despite the apparently smooth distribution of features, clear differences in activation across different examples show that semantic disambiguation helps learn better contextual

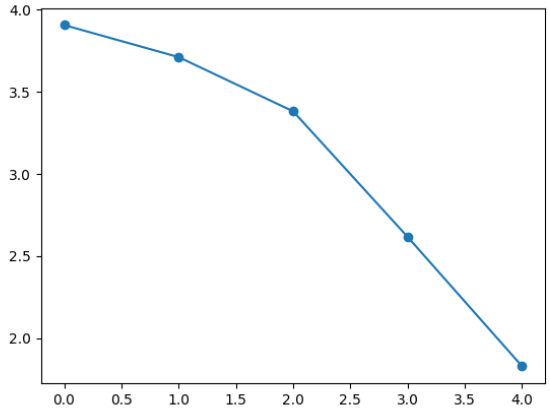

Fig 5. Training Loss Curve

representation. From the training loss graph as shown in figure 5, the results indicates that the training process is stable and that loss decreases steadily over each epoch without any major fluctuations, which signifies that the training procedure and choice of learning rates are efficient. As a result, the model is able to produce consistent F1-scores on all tasks, especially in the offensive and motivational categories (F1-scores = 0.88 and 0.93, respectively). The relatively less performance on humor and sarcasm categories shows how hard it is to capture implicit semantics, emphasizing the usefulness of semantic disambiguation.

V. CONCLUSION

In this paper, Semantic Disambiguation Transformer for multimodal emotion analysis on memes is proposed. The inclusion of semantic disambiguation in the model helps it to resolve ambiguity associated with the textual modality of the input meme. More specifically, our approach is capable of addressing any ambiguity related to sarcasm or indirectness in expressing emotions. The experiments performed on the proposed approach have shown impressive performance, especially in terms of accuracy, macro F1-score, and AUC scores. Moreover, the use of weighted loss functions also increases the performance of our model in situations when there is data imbalance among classes. Confusion matrices, ROC, Precision-Recall and Attention heat maps are useful for interpreting how the model works in different scenarios. In general, the developed model can be applied to solve multimodal emotion analysis problems efficiently. Even though our proposed method performs well in terms of accuracy and other evaluation measures, still, there is scope for improvement. Specifically, one of the ways to improve our approach can be through the use of transformer-based multimodal fusion techniques, where cross-modal attention is used instead of concatenated features. The addition of explainable AI techniques, including frameworks like SHAP or Grad-CAM, can enhance interpretability by providing insights into model decision-making processes. Moreover, adapting the model for cross-lingual meme understanding would broaden its usability across multilingual contexts, allowing it to handle diverse linguistic and cultural variations in online content.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[3] C. Sharma, D. Hazarika, S. Poria, and E. Cambria, “SemEval-2020 Task 8: Memotion Analysis – The Visuo-Lingual Metaphor,” in Proc. 14th Int. Workshop on Semantic Evaluation (SemEval-2020), 2020, pp. 759–773.

[4] D. Kiela, Z. Firooz, A. Mohan, et al., “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes,” in Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.

[5] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal Sentiment Analysis: Addressing Key Issues and Setting up the Baselines,” IEEE Intell. Syst., vol. 33, no. 6, pp. 17–25, 2018.

[6] S. Pramanick, S. Roy, S. Mishra, and A. K. Roy, “MemoSYS at SemEval-2020 Task 8: Multimodal Emotion Analysis in Memes,” in Proc. SemEval-2020, 2020.

[7] A. Suryawanshi, B. Chakravarthi, M. Arcan, and P. Buitelaar, “Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text,” in Proc. 2nd Workshop on Trolling, Aggression and Cyberbullying (TRAC), 2020.

[8] S. K. Saha, S. Pramanick, and S. Roy, “Utilizing BERT and DenseNet for Internet Meme Emotion Analysis,” in Proc. Int. Conf. Pattern Recognition and Machine Intelligence, 2021.

[9] A. Avvaru and S. Vobilisetty, “BERT at SemEval-2020 Task 8: Using BERT to Analyse Meme Emotions,” in Proc. 14th Int. Workshop on Semantic Evaluation (SemEval-2020), Barcelona, Spain (online), 2020, pp. 1094–1099.

[10] C. Sharma, D. Hazarika, S. Poria, and E. Cambria, “Findings of the Memotion 2 Shared Task: Sentiment and Emotion Analysis of Memes,” in Proc. SemEval-2020, 2020.

[11] S. Pramanick, S. Roy, and S. Mishra, “IITK at SemEval-2020 Task 8: Unimodal and Bimodal Sentiment Analysis of Internet Memes,” in Proc. SemEval-2020, 2020.

[12] A. Kumar, R. Singh, and A. Bansal, “NIT-Agartala-NLP-Team at SemEval-2020 Task 8: Multimodal Sentiment Analysis using Deep Learning,” in Proc. SemEval-2020, 2020.

[13] H. Zhang, L. Wang, and J. Liu, “Emotion-Aware Multimodal Fusion for Meme Understanding,” Inf. Process. Manage., vol. 59, no. 2, 2022.

[14] D. Singh, R. Gupta, and A. Sharma, “A Vision-Language Model for Multitask Classification of Memes,” in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), 2022.