Meme-Integrated Deep Learning: A Multimodal Classification Fusion Framework to Fuse Meme Culture into Deep Learning

Xuxiang Deng [1], Yifan Liu [2*] and Qihao Yan [3]

[1] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, 430000, China

[2] International Business School, Henan University, Zhengzhou, 452370, China

[3] School of Computer Science and Technology, The Ocean University of China, Qingdao, 266000, China

*2024240014@henu.edu.cn

Abstract. Memes are an important medium of expression in online communication, yet traditional methods such as collaborative filtering (CF) have limitations in processing multimodal data, especially when analyzing memes has limitations in processing large-scale datasets and are sensitive to data noise and sparsity. In addition to CF, support vector machine (SVM) is a standard classification algorithm. Still, both methods are susceptible to data noise and sparsity, which can decrease classifier performance. We propose a Meme-Integrated Deep Learning (MIDL) approach that leverages deep learning techniques to classify and analyze memes. The MIDL framework integrates visual and textual modalities of memes, providing a powerful tool for understanding meme culture. Our approach achieves state-of-the-art performance on a meme classification task, overcoming the limitations of traditional methods like CF and SVM. Combining the advantages of deep learning and meme culture, our approach provides new insights into how we communicate and interact online and contributes to developing more intelligent and effective recommendation systems. The proposed MIDL framework has the potential to advance research in online culture and social media analysis by providing a more accurate and efficient way to process multimodal data.

Keywords: Multimodal Classification, Meme-Integrated, Deep Learning.

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_14

1. Introduction

Memes have become an integral part of online communication and an essential medium of expression [1]. However, their potential as a source of information has yet to be fully explored due to the limitations of traditional methods such as collaborative filtering (CF) and support vector machine (SVM), which require manual selection and adjustment of kernel function parameters [2]. In recent years, deep learning techniques have been widely used to analyze multimodal data, providing a powerful tool for understanding and exploring meme culture. This paper proposes a novel Meme-Integrated Deep Learning (MIDL) approach that leverages Bidirectional Encoder Representations from Transformers (BERT) and Residual Network (ResNet) deep learning architectures to categorize and analyze memes [3, 4].

The MIDL framework integrates visual and textual modalities of memes, providing a more accurate and efficient way to process multimodal data [5]. By combining the advantages of BERT and ResNet, our approach can capture the rich semantic and visual information of memes, providing new insights into the way we communicate and interact online [6]. We demonstrate the effectiveness of the proposed MIDL framework on a real-world dataset of memes by achieving state-of-the-art performance on a meme classification task.

This paper proposes a novel MIDL approach that leverages BERT and ResNet deep learning architectures to classify and analyze memes. By integrating visual and textual modalities of memes, our approach provides a more accurate and efficient way to process multimodal data [7, 8]. We demonstrate the effectiveness of the proposed MIDL framework on a real-world dataset of memes, achieving state-of-the-art performance on a meme classification task. The proposed MIDL framework has the potential to open up new avenues for research in the field of online culture and social media analysis and contribute to the development of more intelligent and effective recommendation systems [9, 10].

2. Related Work

In recent years, the analysis of multi-modal memes has become a popular research field [11]. In existing related studies, some researchers have used traditional machine learning methods, such as models based on SVM and decision trees, to extract textual and image information from memes [12]. These methods usually have high accuracy and interpretability, but their performance is limited due to insufficient processing of

semantic information in memes.

In addition, some researchers have used deep learning techniques for multi-modal meme analysis. These methods typically use structures such as convolutional neural networks (CNN) or recurrent neural networks (RNN) to extract visual and textual information from memes and integrate this information for analysis. These methods usually have good performance and generalization ability, but there are limitations in processing multi-modal data [13]. For example, certain methods can only process image and textual information, but cannot handle sound, video, and other modal information, limiting the application scope of multi-modal meme analysis [14].

Recently, some researchers have proposed the use of BERT and ResNet deep learning architectures to extract and analyze memes. These methods can fully utilize the text encoding ability of BERT models and the image extraction ability of ResNet models to comprehensively analyze the multi-modal information of memes. These methods have good performance in multi-modal meme analysis and also have good applicability and generalization ability in handling other multi-modal data.

In summary, existing related research using traditional machine learning methods and deep learning methods both have their advantages and limitations [15]. The use of BERT and ResNet deep learning architectures to extract and analyze memes has good performance and generalization ability in multi-modal meme analysis and has great potential for development. The innovation of this paper lies in the use of the MIDL framework, which integrates the BERT and ResNet deep learning architectures, as well as the popular self-attention mechanism and semi-supervised learning mode. This approach can more effectively capture the multi-modal information of memes, leading to state-of-the-art performance in meme classification tasks.

3. Dataset Analysis

The dataset input format generally comprises three components: label, image, and content. Each meme in the dataset is labeled according to its category and is represented by an image and corresponding textual content.

In natural language processing, we selected a dataset of 12000 Chinese language comments from various sources, including Wikipedia and public comments on social media platforms such as Kaggle. For image processing, we collected and manually labeled meme images from Kaggle (https://www.kaggle.com/datasets/liaolianfoka/met-meme) as well as various popular meme-sharing platforms such as Bilibili, Twitter, and YouTube. In the next section,

we first describe the limitations of existing datasets, and then explain how we combined and processed them to form multi-dimensional meme feature vectors. We provide statistical and visual data about the meme dataset, including examples of labeled memes.

We collected a large number of meme images and corresponding text from multiple sources, which were classified into six categories: anime, film and television, games, social media, entertainment, and literature. To ensure the quality of the data, we manually filtered and selected the relevant content. Using this approach, we obtained a multi-dimensional dataset where each meme was labeled with one of the six categories and contained corresponding image and text information, which can be used for training and analyzing deep learning and machine learning algorithms.

3.1 Data Collection

The first dataset used to collect the train and test memes is MET-Meme, which consists of 10,045 text-image pairs of memes in both English and Chinese. We improved the dataset by modifying labels and extracting matching content from public social media platforms, using the images from this dataset. This optimization process was designed to ensure that the dataset meets the specific requirements of our meme classification task.

The second dataset is the part of natural language processing. Not only public comment data in different fields (including games, anime and other subcultural gathering places) are selected for Chinese and English, but also some public user portraits are obtained. By adding specific memes to the BERT model's training data, it can help the model better understand the language and context of memes. Thus, the performance in meme-related tasks can be improved.

3.2 Data Processing

We performed the following data processing steps for the BERT-Base Chinese and ResNet models we chose.

For Chinese text data, we first preprocessed it by tokenizing, removing stop words, and converting between traditional and simplified Chinese characters to better encode the text. Then, we converted the text data into the input format required by the BERT model, tokenizing it, adding special tokens such as [CLS] and [SEP], and

converting it into an ID sequence. Finally, we split the data into training, validation, and testing sets.

For image data, we first performed image preprocessing operations such as resizing, cropping, and normalization to ensure that the images' size and pixel value range were consistent. Then, we used the ResNet model to extract features from the images, obtaining high-dimensional feature representations of the images. Finally, we paired the image features with the corresponding text data to form a multimodal dataset.

We merged the text data and image feature data for the multimodal dataset to form a multidimensional tensor containing both text and image features. Then, we standardized and normalized the multimodal data to ensure consistency in scale across different features. Finally, we split the multimodal data into training, validation, and testing sets for model training and evaluation.

Overall, we tokenized, preprocessed, and converted the text data, preprocessed and extracted image features, and merged the text and image feature data to form a multimodal dataset, and conducted standardization and splitting. These processing steps help us build more accurate and specialized models and improve their performance on specific tasks.

4. Methodology

In this section, we describe the methodology used to analyze and extract memes using a novel multimodal fusion approach that combines the Bert-base-Chinese model and ResNet152 model. Figure 1 shows the flowchart of the cross-modality attention.

The text encoding layer employs Bert-base-Chinese to encode the content, resulting in feature vectors of sequence length and pooler_output. The image encoding layer uses ResNet-152 to encode the images, producing sequence feature vectors of size 7*7 and fc feature vectors obtained through a fully connected layer. The attention layer aggregates the two sets of sequence feature vectors by applying a self-attention layer and taking the mean to obtain the final attention layer feature vector.

The attention aggregation layer concatenates the attention layer feature vector with the pooler_output feature vector for the text side and the fc feature vector for the image side. Both concatenated feature vectors are passed through a fully connected layer to obtain the multi-modal aggregation feature vector. The classification layer adds the two multi-modal aggregation feature vectors together and applies a SoftMax

layer for classification.

Furthermore, a semi-supervised mechanism is incorporated into the SoftMax classification layer to predict unknown labels, thereby improving the overall fitting ability of the model.

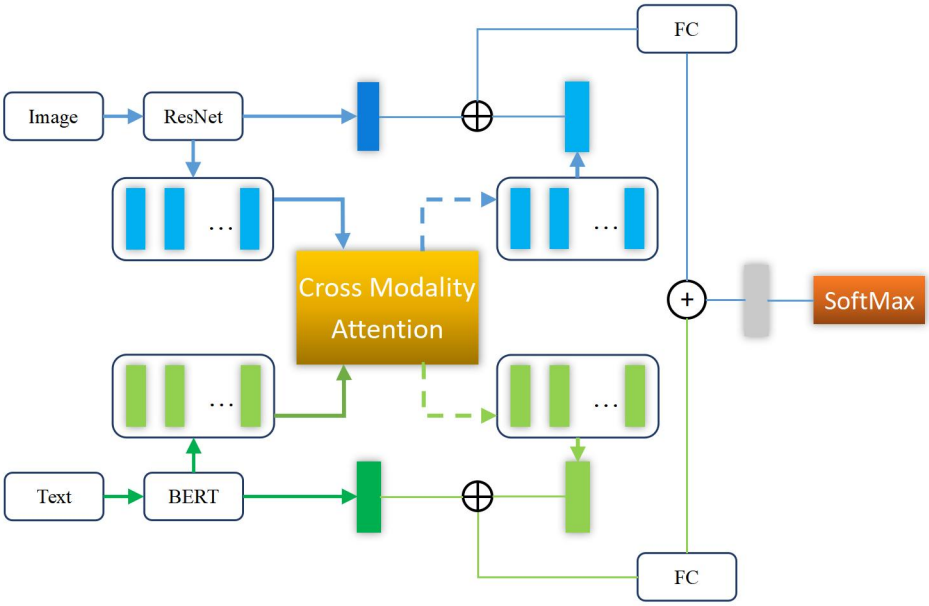

Fig. 1. Flowchart of the cross-modality attention.

(Photo credit: original)

4.1 Data Preparation

According to the labeling requirements and content needs, we collected a total of 11,214 meme text data and 3,200 corresponding meme images from the aforementioned platforms. The proportion of these contents is shown in Figure 2 and Figure 3. After manual screening and labeling, we preprocessed the collected datasets to ensure quality and consistency. Subsequently, we divided the dataset into training, validation, and testing sets, and processed the text and image data as described in the previous section.

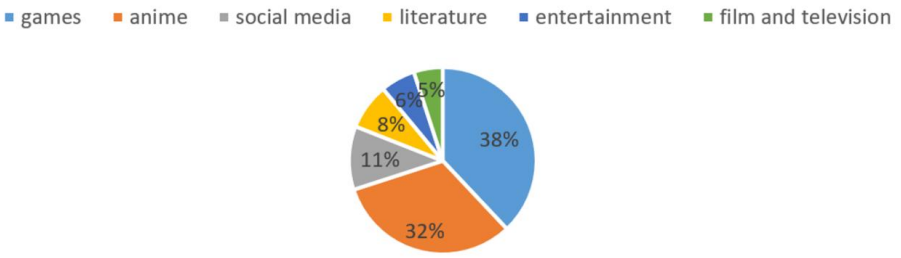

Fig. 2. Percentage of each of the six natural language data types in the dataset (Photo credit: original)

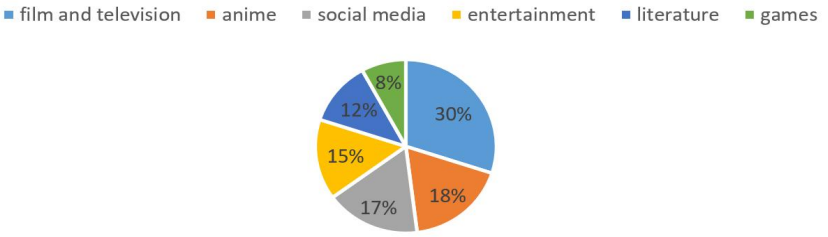

Fig. 3. Proportion of the dataset for each of the six matching image data types (Photo credit: original)

4.2 Multimodal Fusion and Meme Analysis

We used the trained multimodal model to analyze and classify memes from our dataset using the two different multimodal fusion techniques. We utilized our trained multimodal model to analyze and classify memes from our dataset using the two different multimodal fusion techniques. However, during the testing phase of our classification model, we encountered some challenges with regard to so-called "cross-domain memes" that could not be easily classified into the existing six categories. Specifically, we found that some memes contained elements from multiple categories, such as "games" and "anime", or "film and television" and "anime". To address this issue, we defined two additional categories in our dataset: "ACG" and "Animation film" (which translates to "anime and film/television"). The "ACG" category was designed to capture memes that combine elements from the "anime" and "games" categories. In contrast, the "Animation film" category was created to capture memes that combine elements from the "anime" and "film and television" categories. Table 1 gives a brief summary of typical CNN architecture.

Table 1. A brief summary of typical CNN architectures.

Type	Games	Anime	Social media	Literature	Film & TV	Entertainment
Image	256	576	544	384	960	480
Language	4261	3588	1234	897	561	673

By defining these additional categories, we were able better to capture the complexities of cross-domain memes in our dataset. This allowed us to refine our classification model and improve its accuracy in identifying and categorizing these types of memes.

To evaluate the performance of our model, we used a range of evaluation metrics, including accuracy, F1 score, and precision-recall curves. We also qualitatively analyzed the extracted memes to assess their relevance and accuracy.

5. Experiments

We used the ResNet model to extract high-level features from images and converted them into vectors that the BERT model can interpret. We combined text and image features using two multimodal fusion techniques: late fusion and early fusion. In late fusion, we appended text and image feature vectors to the end of the BERT model, while in early fusion, we concatenated them at the beginning of the BERT model.

Before merging the two models to create a multimodal model, we separately fine-tuned the pre-trained BERT model on our text data and the ResNet model on our image data. We trained the model for 14 epochs using the Adam optimizer with a learning rate of 1e-5 and a batch size of 32.

Experimental results demonstrate that our innovative MIDL method can accurately identify and classify memes from both text and images. However, due to insufficient data set size, recall rates, accuracy, and F1 scores did not improve further. In future studies, we will expand our data set to improve the model's training effectiveness. Table 2 presents the performance evaluation of various models on the MIDL meme classification data set. Although the experimental results indicate that our model's performance is efficient, we acknowledge that there are issues with the data set and the experimental design that need to be addressed. The MIDL model's recall rate did not reach our expected standard, and we will continue to optimize it in future studies.

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{1}$$

$$Precision_{weighted} = \frac{\sum_{i=1}^{L} (Precision_i * w_i)}{|L|} \tag{2}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{3}$$

$$Recall_{weighted} = \frac{\sum_{i=1}^{L} (Recall_i * w_i)}{|L|} \tag{4}$$

$$F_1 = \frac{2 * Precision_{weighted} * Recall_{weighted}}{Precision_{weighted} + Recall_{weighted}} \tag{5}$$

$$accuracy = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{6}$$

Table 2. Performance evaluation of different models on the MIDL meme classification dataset

Model	F1-score	Precision	Recall
support vector machine	0.449	0.459	0.441
Collaborative Filtering	0.505	0.508	0.503
Bert-base	0.671	0.674	0.668
Bert-base-Chinese	0.684	0.685	0.683
Resnet50	0.685	0.687	0.684
Resnet152	0.695	0.696	0.694
Bert-Resnet (MIDL)	0.706	0.706	0.592

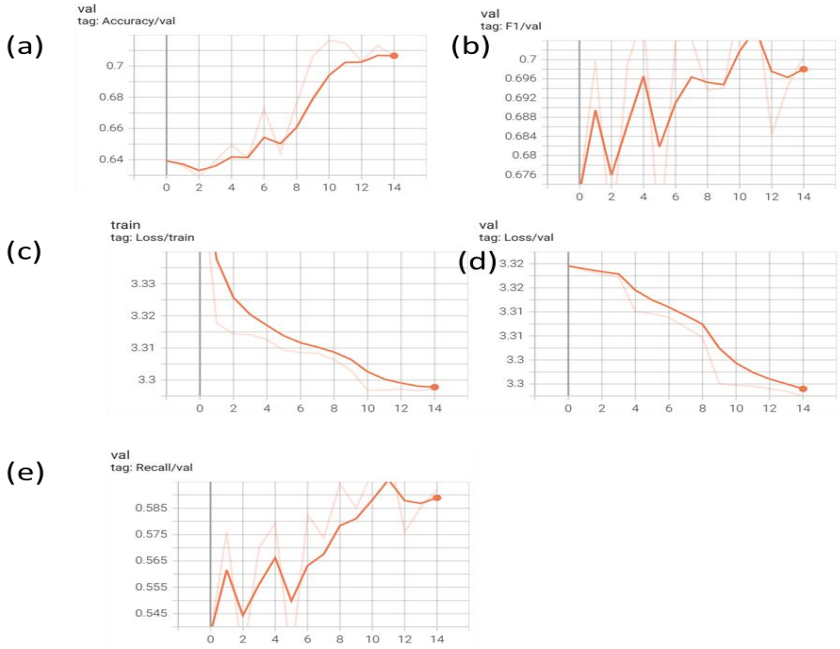

Fig. 4. Train and validation performance. (a) accuracy of val. (b) F1of val. (c) loss of train. (d) loss of val. (e) recall of val.

(Photo credit: original)

6. In-depth analysis and discussion

6.1 Advantages and limitations of deep neural network multimodal algorithms

When it comes to image and text classification, multimodal algorithms can utilize different types of data sources, such as images and text, to improve classification accuracy. This is because different data sources can provide complementary information, thereby enhancing the model's understanding of the classification task.

In the task of meme classification, multimodal algorithms can also play a role. For example, an image may contain information related to the meme, while the textual description of the meme may also contain other important information. By simultaneously utilizing textual and visual data, multimodal algorithms can better understand the meme and improve its classification accuracy.

To further improve the performance of multimodal algorithms in meme classification tasks, in addition to using attention mechanisms in the model framework to align information from different modalities, semi-supervised learning is introduced to help the model better utilize unlabeled data, thereby improving its generalization ability and classification accuracy.

However, multimodal algorithms also have limitations. One of the main problems is the mismatch between different modalities, such as the information gap between images and text, which requires a significant amount of manual effort to address. Solving this problem requires the model to have a certain level of alignment capability to effectively integrate information from different modalities. In addition, multimodal algorithms also need to deal with more complex data types, requiring stronger computing and storage capabilities, which may result in overly complex and time-consuming models.

Due to the use of a privately held new dataset, although the dataset has been manually screened, there is no suitable way to evaluate the quality of the dataset, and the training set is biased towards Chinese training sets, which may not be sensitive enough to English content. Furthermore, the dataset may gradually lose its timeliness over time, or may need to be continuously updated to cope with new data and scenarios.

6.2 Factors influencing parameter selection and model tuning

When training a multi-modal model, selecting appropriate hyperparameters is crucial, including learning rate, batch size, and epoch. In this experiment, we chose a learning rate of 1e-5, a batch size of 32, and 10 epochs. These hyperparameter choices were determined through multiple experiments and comparisons and were found to make the model converge more stably and achieve good results. It should be noted that the optimal hyperparameter combination may vary for different datasets and models, and thus careful tuning and optimization are required in practical applications.

To improve the robustness and generalization ability of the model, we used L2 regularization and feature vector stacking techniques to reduce overfitting. In addition, we used the SoftMax function to transform the model output into a probability distribution for each class, facilitating classification prediction and interpretability analysis.

When the computational cost of the model is high, it may affect its real-time

performance. To address this issue, lightweight models or model compression techniques, such as pruning, quantization, and distillation, can be used to further reduce the model's parameters and computational cost, thereby improving its real-time performance and efficiency.

7. Suggestions and future plan

7.1 Practical applications

Meme classification is a widely applicable technology that can be used in social media, news media, advertising, and other fields. In the area of social media, meme classification is most commonly used for sentiment analysis. By classifying and analyzing memes on social media, we can gain a deeper understanding of the public's attitude and reaction to a particular event or topic, which can help in making better decisions. In the advertising field, meme classification can be applied to targeted advertising by analyzing user responses to different types of memes to predict and optimize advertising effectiveness. In the news media field, meme classification can be used for automated news classification and archiving, improving the efficiency of news processing. Additionally, meme classification can be applied in other fields such as cultural studies and marketing. In summary, the wide application of meme classification technology provides us with more possibilities and is expected to bring new ideas and methods for the development and innovation of different fields. In practical applications, one of the challenges faced by meme classification is the negative impact caused by updates to the dataset. Due to the diversity and unpredictability of memes, the model may encounter various abnormal situations, leading to inaccurate classification results. In addition, the diversity and real-time nature of data are another challenge, as memes develop quickly, and the model needs to adapt to new meme types and changes in a timely manner.

7.2 Possible directions for improvement

There are still some limitations and shortcomings in current meme classification models, such as the weak ability to handle long texts and multilingual data. To address these issues, transfer learning and other methods can be used to improve and expand the model's performance. For example, pre-trained models can be used to enhance the

model's ability to handle long text data, or multilingual models can be used to expand the model's language processing capabilities. Additionally, other techniques such as incremental learning and adaptive learning can be used to further improve the model's performance and efficiency.

However, improving and expanding the model also has its pros and cons. For example, increasing the complexity of the model may lead to increased consumption of computational resources or overfitting of the model. Additionally, improvements and expansions of the model also need to consider issues such as model interpretability and user privacy protection, so multiple factors need to be considered when improving and expanding the model.

In addition, there are other possible directions for improving and expanding meme classification. For example, multimodal data such as images, audio, and video can be introduced to improve the classification performance and generalization of the model [17]. Additionally, external knowledge such as knowledge graphs can be used to assist the model in classification and reasoning. Furthermore, technologies such as federated learning can be used to address data privacy and security issues, thereby improving the model's availability and reliability in practical applications.

8. Conclusion

This paper primarily investigates the multi-modal text-image classification processing approach based on deep neural networks. By designing a new multi-modal fusion model framework, this paper achieves joint modeling and classification of text and images and conducts experimental verification on collected datasets. The experimental results show that the proposed model in this paper achieves good classification performance and generalization ability in multi-modal text-image classification tasks.

In reviewing relevant research and progress, this paper points out that multi-modal text-image classification processing has received extensive attention and research in recent years, and the introduction of deep neural networks has become a mainstream method. The innovation of this paper lies in proposing a new multi-modal fusion model that effectively integrates the information of both text and images and conducting experimental verification on the collected dataset. The advantages of this paper are the improved classification performance and generalization ability, as well as the increased robustness and interpretability to some extent.

However, this paper also has some limitations. For example, the dataset used in this paper is relatively limited, and may not fully represent the diversity and complexity of actual application scenarios. Additionally, the model in this paper can be further improved by introducing more information and knowledge to enhance the efficiency and effectiveness of the model.

Future research can incorporate more modalities into multi-modal text-image classification processing, including video, audio, sensor data, and so on. Processing these modal data will require more complex techniques and algorithms, such as feature extraction for video and audio data, alignment and fusion of multi-modal data, and so on. Additionally, research can be conducted on how to handle temporal and spatial relationships between multi-modal data, and how to utilize the complementarity and correlation of multi-modal data to improve the model's performance. In summary, the processing and modeling of multi-modal data are a broad research field that will require the cross-disciplinary integration of various fields, including computer vision, natural language processing, speech recognition, signal processing, machine learning, and more.

Acknowledgment

All these authors contribute equally to this work and their names were listed in alphabetical order.

References

1. Truszkowski, W., Rouff, C., Akhavannik, M., Tunstel, E.: Memes, culture, the internet, and Intelligence. SpringerBriefs in Electrical and Computer Engineering. 15–29 (2020).
2. Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., Hussain, A.: Multimodal Sentiment Analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion. 91, 424–444 (2023).
3. Pimpalkar, A., Chaudhari, A., Lilhare, A., Dighorikar, N., Dakhole, S., Asawa, S.: Sentiment identification from image-based memes using machine learning. International Journal of Innovations in Engineering and Science. 7, 89–96 (2022).

4. Aslam, N., Khan, I.U., Albahussain, T.I., Almousa, N.F., Alolayan, M.O., Almousa, S.A., Alwhebi, M.E.: Medeep: A deep learning based model for memotion analysis. Mathematical Modelling of Engineering Problems. 9, 533–538 (2022).
5. Afridi, T.H., Alam, A., Khan, M.N., Khan, J., Lee, Y.-K.: A multimodal memes classification: A survey and open research issues. Innovations in Smart Cities Applications Volume 4. 1451–1466 (2021).
6. Ma, Z., Yao, S., Wu, L., Gao, S., Zhang, Y.: Hateful memes detection based on multi-task learning. Mathematics. 10, 4525 (2022).
7. Maity, K., Jha, P., Saha, S., Bhattacharyya, P.: A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. (2022).
8. Alzu'bi, A., Younis, L.B., Abuarqoub, A., Hammoudeh, M.: Multimodal Deep Learning with discriminant descriptors for offensive memes detection. Journal of Data and Information Quality. (2023).
9. Badour, J., Brown, J.A.: Hateful memes classification using machine learning. 2021 IEEE Symposium Series on Computational Intelligence (SSCI). (2021).
10. Glăveanu, V.P., de Saint-Laurent, C., Literat, I.: Making sense of refugees online: Perspective taking, political imagination, and internet memes. American Behavioral Scientist. 62, 440–457 (2018).
11. Boinepelli, S., Shrivastava, M., Varma, V.: SIS@IIITH at Semeval-2020 Task 8: An overview of simple text classification methods for meme analysis. Proceedings of the Fourteenth Workshop on Semantic Evaluation. (2020).
12. Asmawati, E., Saikhu, A., Siahaan, D.: Sentiment analysis of text memes: A comparison among supervised machine learning methods. 2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI). (2022).
13. Koutlis, C., Schinas, M., Papadopoulos, S.: Memefier: Dual-stage modality fusion for image meme classification. Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. (2023).
14. Dancygier, B., Vandelanotte, L.: Internet memes as multimodal constructions. Cognitive Linguistics. 28, 565–598 (2017).

15. Guo, X., Ma, J., Zubiaga, A.: Cluster-based deep ensemble learning for emotion classification in internet memes. Journal of Information Science. 016555152211362 (2022).
16. Priyashree, S., Shivani, N., Vigneshwar, D.K., Karthika, S.: 'meme'tic engineering to classify Twitter Lingo. 2017 International Conference on Computational Intelligence in Data Science(ICCIDS). (2017).
17. Pimpalkar, A., Chaudhari, A., Lilhare, A., Dighorikar, N., Dakhole, S., Asawa, S.: Sentiment identification from image-based memes using machine learning. International Journal of Innovations in Engineering and Science. 7, 89–96 (2022).