IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 9, No. 2, April, 2022, pp. 533-538
Journal homepage: http://iieta.org/journals/mmep

MEDeep: A Deep Learning Based Model for Memotion Analysis

Nida Aslam*, Irfan Ullah Khan, Teaf I. Albahussain, Nouf F. Almousa, Mizna O. Alolayan, Sara A. Almousa, Modhi E. Alwhebi

College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 34221, Saudi Arabia

Corresponding Author Email: naslam@iau.edu.sa

https://doi.org/10.18280/mmep.090232

Received: 2 December 2021
Accepted: 18 January 2022

Keywords:
memes, classification, deep learning (DL), convolutional neural network (CNN), emotion analysis

ABSTRACT

A meme is an idea or an expression that becomes a trend and usually spreads within a culture through imitation, carrying a significant meaning representing a specific concept or theme. It represents a particular message by the combination of image and text. An enormous number of memes are spread every day via social media platforms to share sarcastic, humorous, and offensive messages. Therefore, to control the spread of offensive language and propaganda, there is a need for an automated mechanism to classify the meme. Considering this, the proposed study aimed to investigate the impact of using different data modality i.e., text, images, combined (text and images) for memotion analysis using Deep Learning (DL) models. Several pre-trained DL models such as ResNet152V2, VGG19, EfficientNetB7 were used for the images. While, Convolutional Neural Network (CNN) and CNN+ Long short-term memory (LSTM) were implemented for the text classification. Experimental results reveal that the CNN model outperformed for the text, and EfficientNetB7 achieved the best performance on the images. However, for the multimodal analysis, early fusion technique was used and classification was performed using CNN and EfficientNetB7 model. The study found that Glove embedding and CNN model using text produced the highest results among all the experiments conducted. The model achieved an accuracy, F1-macro, precision, and recall of 0.8387, 0.8352, 0.8361, 0.8487, respectively. The results exhibit that the proposed model outperforms other baseline studies.

1. INTRODUCTION

Nowadays, the ubiquity of the internet and social media has given the freedom to the individual to share their views and perception easily. Besides, the omnipresence of internet memes on social media sites, and the usage in online communication. Therefore, it is not confined to one language only and led to many benefits such as easy to read, a quick laugh [1]. However, some of the companies used in digital marketing to grab individual's attention. Despite of the advantages, it sometimes leads to the source of spreading fake news or publicizing hate speech. Recently, memes have gained great attention in social media and can get viral easily [2].

Memes come in every form of social media networks that consist of text and images. It is nonverbal communication and typically refers to cultural impressions gained from TV shows or movies and resonates with people's thoughts and feelings more frequently. Furthermore, the proliferating usage of these memes requires an automated process that is more scalable compared to using manual analysis [3]. The implication of Machine Learning (ML), specifically DL models, has greatly enhanced the automated analysis of multimedia data such as audio, visual, textual, and videos. Studies have been made on the automated generation, sentiment analysis, emotion analysis of memes [4]. Therefore, we aimed to develop a DL-based model that can analyze features used for categorizing a meme to their intended semantic meaning.

The SemEval 2020 shared task on memotion analysis. Sharma et al. [5] draw attention to the analysis of memes conveyed by sentiment and humor. The challenge involves three tasks, first task (A): the sentiment classification of a meme (positive, negative, or neutral); second task (B): emotion humor classification (humor, sarcasm, offensive, and motivational), third task (C) is to scale the level of semantic classes sarcasm, humor, and offense conveyed in a meme.

This paper intends to propose a memotion analysis bimodal model for task A to perform memotion analysis as positive, negative, and neutral. We aim to develop a model with enhanced performance compared to the previous studies related to meme classification. Therefore, we compared several models using the visual and textual features. Finally, we selected and combined the model that achieved the best performance. The proposed system will help in many fields, such as marketing. Some companies tend to market their products through social media, and the customers may share their experiences using memes. Therefore, using the proposed model, companies can know customer satisfaction. Furthermore, it can also help in identifying and controlling toxic memes.

The remaining part of this work is organized as follows: Section 2 contains a review of related literature. Section 3 contains the materials and methods that include dataset description, preprocessing, description of the classifiers Section 4 presents the experimental setup and results. While section 5 contains the conclusion and recommendations emanating from this work.

2. REVIEW OF RELATED LITERATURE

Nowadays, a meme is one of the buzzwords [6]. A meme is a particular behavior, idea, or stylistic method that can be conveyed amongst people [7]. These people demonstrate similar characteristics, such as a culture that can be represented using a meme. Peirson V and Tolunay [4] proposed a novel meme generation model that can produce relevant information related to the selected caption. This model further enhances the meme analysis process by ensuring that the system is conditioned not only to images. To this effect, the system can generate user-defined information related to the chosen meme template resulting in a better understanding of the underlying meme content among users. The study also suggests that the meme analysis procedure can utilize the pre-trained Inception-v3 network to an image related to any caption. These outcomes are achieved by passing the data through an attention-related deep-layer LSTM (Long Short-Term Memory Recurrent Neural Network) model. The identified model ensures that the system generates authentic memes that match the real ones. Few attempts have been made to classify the meme using ML and DL. However, some of the studies proposed the unimodal model considering either the text or image features, while others used a multimodal analysis using the textual and visual features to classify memes. The section below contains some recent studies for memes classification using uni and multimodal data.

Gundapu and Mamidi [8] present a multimodal sentiment analysis system to classify the memes as neutral, negative, and positive and identify the humor type expressed and measure the degree to which a specific effect is expressed, using DL techniques such as Convolutional Neural Network (CNN) and LSTM. The model combined Computer Vision (CV) and NLP (Natural Language Processing) techniques, presents three DNN architectures, and compared the result. They found that a multimodal NN always performed better for sentiment classification and achieved the highest F1-macro of 0.3391. Similarly, Bonheme and Grzes [9] developed an efficient model that can be used to analyze the various meanings of memes. The model was presented in the SemEval 2020 task 8 competition seeking to provide a concrete explanation of algorithms that can be integrated into understanding the underpinning models. This competition is geared towards understanding three core areas. These areas are determining the polarity associated with specific memes, predicting the humor linked to a meme, and evaluating the multi-output characteristics of a meme, such as offense, sarcasm, and humor. Moreover, the selection of the most effective model depends on conducting appropriate tuning of the given algorithms. The key strength of the proposed work was that they tried to explain the importance of these algorithms and their instrumental role in their models. To this effect, the authors sought to understand the implications of the different multimodal representative tactics or data applied during meme analysis. Another strength is that it indicates the best way to deal with crucial multimodal information. They used datasets that contained meme images selected through optical character recognition (OCR) and found that alignment-based strategies are not effective in meme analysis.

Furthermore, they found that there is no correlation between texts and images. Several ML models were trained, such as Random Forest (RF), Gaussian Naïve Bayes (GNB), K Nearest Neighbor (KNN), and Multi-Layer Perceptron (MLP), along with the embedding technique for the text. Experiments were conducted using unimodal and multimodal analysis. For the multimodal analysis, fusion, concatenation, and Canonical correlation analysis (CCA) were used. KNN achieved the highest F1-macro of 0.35 for the memotion of memes using images only.

Keswani et al. [10] presented their approaches for the problem in SemEval-2020 Task 8 to classify memes according to their emotional sentiment and content. They employed unimodal (text only) and bimodal (text and image) with different techniques. For multimodal analysis, two models were developed, such as Feed Forward Neural Network (FFNN) for text and CNN for image, and Multimodal Bitransformer (MMBT). However, for the text only, Naïve Bayes (NB), FFNN, and BERT techniques were used. The results were compared based on the F1-macro, which is the official evaluation metric for Memotion Analysis. The study achieved the highest result of 0.35 in the unimodal (text only) using FFNN with word2vec embedding. However, for the multimodal, the best performance was achieved using MMBT with the F1-macro of 0.30.

Similarly, Bejan [11] presents a solution for emotion analysis using the MemoSys system submitted in Task 8 of SemEval 2020. The study aimed to classify the sentiment of internet memes using four different DL models. Among all the model combinations, using fine-tuned the BERT base uncased model along with the VGG16 using softmax and fusion the extracted features, got the highest accuracy result 0.4988, and F1-macro of 0.3513 for task A.

Additionally, a system was introduced by Gupta et al. [12] that uses various bimodal fusion strategies to exploit the inter-modal dependence for emotion and humor classification tasks by training each task separately. The system focused on the meme's textual content, visual part, and the fusion of both. Furthermore, they used different models and fusion strategies to achieve each task's best performance using F1-macro as the evaluation metric. Task A and task B's best performance was achieved using the RoBERTa+ResNet model with early fusion strategy, while task C's best performance was achieved with the same model but using late fusion strategy. The highest F1-macro of 0.357 was achieved using early fusion for the sentiment analysis of meme.

Correspondingly, Singh et al. [13] also emphasize on creating a unified network for multimodal meme analysis to classify the sentiment, classify the humor, and scale the semantic classes. The proposed multi-model can perform all three tasks simultaneously by creating an independent network for each task. Each task will use a single-layer linear NN for classification, then use a linear layer as an aggregator to combine the embeddings from different modalities and achieve a 0.5915 F1-macro.

Despite of the studies mentioned above related to the meme classification, there is still room for improvement and further investigation. There are very few studies in the sentiment analysis of memes using the images [14]. The highest result was achieved in the study [13] with an F1-macro of 0.5915. They found that the memotion analysis of the memes can be performed using textual and visual data. The results achieved in the study were not satisfactory and needed to be further enhanced. Therefore, in the proposed study, we aimed to develop a model for the memotion analysis of the memes using the SemEval 2020 task 8 dataset with the enhanced outcome compared with the baseline studies. Furthermore, the study will also investigate the impact of data modalities on the memotion analysis of memes.

3. MATERIAL AND METHODS

This section will describe the dataset used in the study. Moreover, the description of the preprocessing and deep learning models used for images, textual and multimodal data are also discussed. Lastly, we describe the evaluation measures used in the study. Figure 1 contains the block diagram of the proposed study.

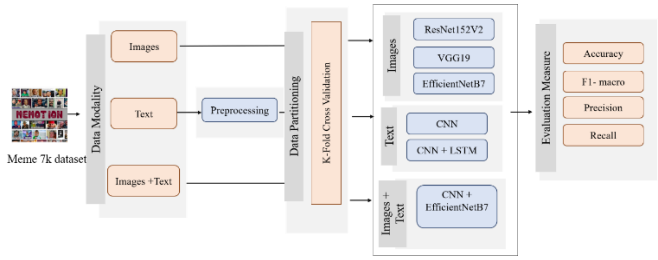

Figure 1. Block diagram for the proposed methodology

3.1 Description of the dataset

The study used SemEval 2020 task 8 dataset [15]. The dataset consists of 7000 manually annotated memes. In the dataset, 6992 memes images with a simple background and embedded textual material are present. However, seven memes do not contain any text. Therefore, they were not considered during the study. Initially, the memes were categorized as very positive/ negative, positive, negative, neutral. The distribution of memes per category is shown in Figure 2. The dataset is highly skewed towards the positive class and contains a minimal number of negative classes. In order to deal with the huge class imbalance among the 5 categories in the dataset, it was converted into three categories as positive, negative, or neutral. Similar strategy is adopted in all benchmark studies [8-13]. After the combination, the number of samples per category for the negative class is 631, neutral class 2201 and positive class 4160. Moreover, combining the positive and very positive as positive and negative and very negative as negative class has reduced the class imbalance and also the risk of model overfitting.

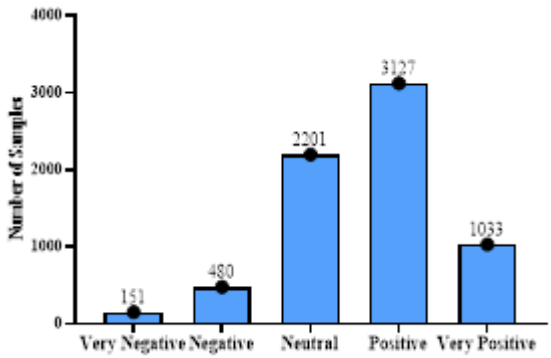

Figure 2. Distribution of samples per category in the dataset

3.2 Data preprocessing

During the preprocessing first, the dataset was converted into three classes by combining (very positive and positive) as positive and (very negative and negative) as negative. After the combination, the number of samples for the negative, positive, and neutral classes is 631, 4160, and 2201, respectively. Moreover, dropna() function was used to remove records that contain any null values. For the text preprocessing, LabelEncoder from sklearn library was used to encode the label class. Additionally, Tokenizer() was used to split the sentences into smaller units as tokens and Glove.6.B.100d pre-trained word vector global vectors for word representation.

3.3 Description of the classifiers

Several classifiers were used for text feature, images and multimodal data (text and images). The section below provides the description of classifiers used in the study for all the three scenarios.

3.3.1 Text classifiers

For the text classification, two classifiers were used such as CNN and CNN+LSTM.

(1) Convolutional Neural Network

One kind of Feedforward Neural Network (FFNN) is the CNN model. It is a multilayer perceptron with five layers, i.e., convolutional layer, pooling layer, and three fully connected dense layers. CNN is used for pattern recognition, text classification, and image processing. It has several potential benefits, including a straightforward layout, fewer training parameters, and adaptability [16].

In our study we used CNN with the input of 100-dimensional vectors, embedding, and an output layer with 'softmax' as an activation function for multi-classification. The details of CNN five hidden layers are dropout layer with 0.2 as the rate of frequency, Conv1D layer with five arguments 250 filters, three kernels, valid padding, ReLU activation function, and one stride, then GlobalMaxPooling1D layer, dense layer with 250 units, dropout layer with 0.2 as the rate of frequency, ReLU activation layer and two dense layers with 100 and 64 units with ReLU as an activation function. The dense layer with three units and 'softmax' as an activation function is used for the prediction layer. The model is compiled with 'adam' as an optimizer, while 'categorical_crossentropy' is set for loss. The model is run with epoch 50 and batch size 128. Figure 3 shows the structure of the CNN model used in the current study.

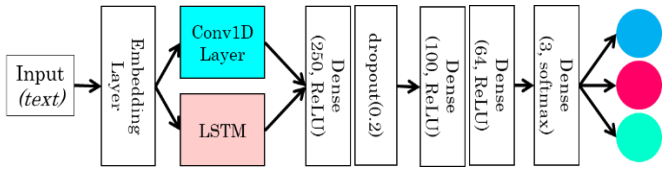

Figure 3. Structure of the CNN model for the textual Data

(2) Convolutional Neural Network and Long Short-Term Memory

CNN+LSTM integrates CNN to extract a series of higher-level phrase definitions, which are then fed into an LSTM to produce the sentence representation. It can recognize phrase local features as well as global and temporal sentence semantics. Using a convolutional layer, CNN+LSTM can learn phrase-level features. Sequences of these higher-level expressions are then fed into the LSTM to learn long-term implementations. When compared to LSTM with approximately the same weights and less training time, CNN with LSTM achieved better test accuracy [17].

Two classifiers were trained and tested the models

(CNN+LSTM) using the same input with 100-dimensional vectors, embedding, and an output layer with ReLU activation function. For the combination of CNN and LSTM, six hidden layers are used, i.e., Conv1D layer with five arguments 250 filters, five kernels, valid padding, ReLU activation function, MaxPooling1D layer with a pool size of five, dropout layer with 0.2 as the rate of frequency, LSTM layer with 128 units, and dense layers with 100 and 64 units having ReLU as an activation function. The output layer is a dense layer with three units, and 'softmax' as an activation function is used for the prediction output. The model is compiled with 'adam' as an optimizer, while 'categorical_crossentropy' is set of loss.

3.3.2 Image classifiers

This section presents three pre-trained models used in the study for image classification, i.e., ResNet152V2, VGG19, and EfficientNetB7.

(1) ResNet152V2

ResNet152V2, a residual neural network, is a pre-trained model for image classification tasks. It has initial weights that can help to achieve better accuracy when compared with the CNN models. The Residual Network (ResNet) is based on a CNN structure consisting of convolutional layers in hundreds or even in thousands. A large number of convolutional layers make the ResNet more efficient in terms of prediction performance [18]. ResNetV2 is using batch normalization prior to each weight layer which makes it distinct from ResNetV1.

(2) VGG19

VGG is a very deep CNN for large-scale image recognition that has two versions VGG16 and VGG19. Each version has different depths, layers, and several parameters. However, VGG19 is deeper, expensive, and larger than VGG16, but it helps improve the performance of the model. The VGG19 model architecture consists of 19 trainable layers in the network, including three layers: convolutional layers with very small receptive fields, hidden layers that use ReLU activation as well as max pooling, three fully connected layers, and a final layer for softmax function [19].

(3) EfficientNetB7

EfficientNet is one of the pre-trained CNN based on a scaling model. It consists of 8 versions from B0 to B7, the number symbolizing more parameters, obtaining a very small number of parameters to keep it more accurate. The EfficientNet model is usually more efficient and accurate than other CNN, work by uniformly scaling and balancing all three dimensions of neural network depth, width, and resolution. Scaling up to deep block layers rather than build a new one as other CNN models [20]. Figure 4 explains the structure of the EfficientNetB7 model for the visual data.

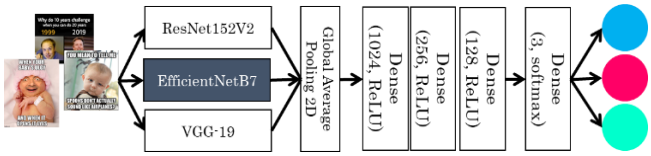

Figure 4. Structure of the EfficientNetB7 model for the visual data

3.3.3 Fusion model

The fusion technique is used to merge the data sources soon after the features are extracted. This type of fusion technique is known as early fusion. The features from the image are extracted with EfficientNetB7, while Conv1D is used to extract features from the text corpus. The extracted features from both sources are then fused by concatenating them to create a single feature vector. It was ensured that the feature vectors are correctly aligned and suitable for further processing. The fused features are then fed into the fully connected feed-forward neural network. The structure of the neural feed-forward network comprises two hidden dense layers with 128 and 64 neurons and ReLU as an activation function. The output layer defines with 3 neurons to perform classification using softmax as an activation function. The configuration of the model includes the Adamx optimizer with a learning rate of 0.001, with categorical cross-entropy as a loss function and accuracy is an evaluation metric. Figure 5 represents the structure of the fusion model for the multimodal data.

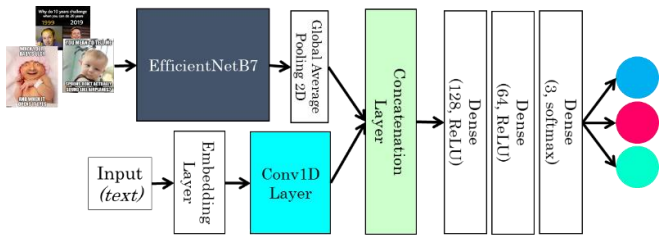

Figure 5. Structure of the early fusion model for the multimodal data

3.4 Evaluation measures

The performance of the proposed models for all the three scenarios i.e., textual feature, visual feature, and multimodal features (text and image) are compared in terms of accuracy, precision, recall and F1-macro. The main evaluation measure is F1-macro for SemEval 2020 shared task on memotion analysis. Therefore, the performance of the proposed model with the baseline studies are compared in terms of F1-macro.

4. RESULTS AND DISCUSSION

The proposed models were implemented using python 3.9.1, using several libraries such as sklearn, keras, pandas, NumPy, and cv2 libraries K-fold cross validation is a technique used for partition the data into training and testing. In this technique the dataset is divided into K subsets. In each iteration one subset is used for testing the model and the remaining subsets were used for training the model. However, in the stratified K-fold cross validation the ratio of the samples in each fold is similar to the ratio of the samples in the original dataset. Due to the class imbalance in the dataset we used stratified k fold cross validation in the current study with the value of K equal to 5. The large value of K indicates that most of the data is utilized in training the model and less in testing the model. In such case the testing error of the model will be low but the model will not be robust. Conversely, if the value of k is very small then most of the data will utilized in testing and the model will be well trained. Therefore, we need to select the optimal value of K. During the experiments several values of K was investigated and found that k=5 produces the best result. For implementing stratified we used StratifiedGroupKFold method of sklearn library with n_split parameter is five and shuffle is set to false and random_state is equal to none. The

training set was further divided into two segments i.e., training and validation. The validation segment was used for parameter optimization.

The study used different models for images and text. These models were concatenated using fusion to make the prediction for multimodal data. Experiments were conducted using unimodal and multimodal to investigate the impact of the data modalities on the sentiment classification of memes. We conducted three sets of experiments:

Scenario I: Unimodal – The models were trained and tested using textual data.

Scenario II: Unimodal -The models were trained and tested using image data.

Scenario III: Multimodal – The models were trained and tested using textual data and image data.

The performance of the models was compared in terms of accuracy, F1-macro, precision, and recall. Experimental results indicate that this study has achieved excellent performance for unimodal and multimodal data models. Table 1 shows the performance of the proposed models using unimodal and multimodal data on a 7k meme dataset. Initially the experiments were conducted using unimodal data to find out the best performing model and later in the third scenario those models were concatenated to train and test the model with the multimodal data. The results in Table 1 indicate that CNN model using textual data from the meme outperformed the other model i.e., CNN + LSTM. Similarly, EfficientNetB7 produces the highest results compared with the other two pre-trained models i.e., ResNet152V2, and VGG19 for the image data. Therefore, the CNN and EfficientNetB7 were concatenated for the third scenario using the multimodal data. The models were combined using early fusion technique.

As indicated by Table 1 results, the models using merely images for memotion analysis shows the worst performance. However, the performance of both models for the textual data is similar. The integration of the LSTM model with the memotion analysis using text has not enhanced the results. Nonetheless, it has a little lower performance. The contribution of the textual data in the memotion analysis is significant. As the result of the multimodality and unimodal (textual) is higher than the unimodal (image). However, the recall of EfficientNetB7 using visual data is higher than the CNN+EfficientNetB7 using combined data.

Additionally, the performance of the proposed study was compared with the baseline studies. The baseline studies listed in the Table 2 were all the previous studies that used a 7k meme dataset for the sentiment analysis (Task A). The results were compared based on the F1-macro, which is the official evaluation metric for Memotion Analysis. The findings of the current study confirmed the hypothesis by Keswani et al. [10] that memotion analysis of the memes can be effectively performed using textual data. While Gundapu and Mamidi [8] achieved the highest F1-macro of 0.3391 using the Multi-model Neural Network (MNN). They found that extracting the image features using inception-V3 and concatenating with the textual features extracting using the embedding produced the highest outcome. Conversely, in the current study, the results produced by using the fusion technique has not enhanced the results as compared to results produced using the textual features.

However, the studies [9-12] have achieved similar results, although their findings were different. Bonheme and Grzes [9] found no correlation among the features extracted from the images and text. Therefore, concatenation of the features didn't enhance the model performance. The study found that sentiment analysis of the memes can be effectively performed using the features extracted from the images rather than using the textual caption or the combined modality data. On the other side [10] found that features extracted from text using Word2Vec and performed classification using FFNN produced the best results. However, interms of the data modalities the findings of [11–13] were similar, they found that combining the multimodal data enhanced the prediction of the meme sentiment. Bejan [11] has used the BERT for embedding the textual features while optimized version of the BERT model i.e., RoBERTa was used by Gupta et al. [12]. While for the image feature different pretrained models were used. However, the results achieved in both the studies were similar. It can be concluded from the results achieved in studies [11], that RoBERTa didn't enhanced the F1-macro of the models [12]. Similarly, Singh et al. [13] also found that using the BERT embedding for textual features and concatenating with the image features produced the highest results. However, in the current study we found that processing the textual features using Glove embedding outperformed the results achieved by using either image or multimodal data.

Table 1. Memotion Classification result for 7k meme dataset using unimodal and multimodal data

Scenario	Model	Modality	Accuracy	F1- macro	Precision	Recall
I	CNN	Text	**0.8387**	**0.8352**	**0.8361**	**0.8487**
	CNN + LSTM		0.8244	0.8229	0.8223	0.8328
II	ResNet152V2	Image	0.4954	0.3907	0.4932	0.4954
	VGG19		0.4182	0.4289	0.4582	0.4182
	EfficientNetB7		**0.5728**	**0.4410**	**0.4477**	**0.5728**
III	CNN+ EfficientNetB7	Text+ Image	**0.5966**	**0.7473**	**0.5966**	**0.5005**

Table 2. Comparison between the benchmark and the proposed study

Study	Modality	Technique	F1-macro
[8]	Text + image	NN	0.3391
[9]	Images	KNN	0.35
[10]	Text	FFNN	0.35
[11]	Text + Image	BERT + VGG using Softmax	0.3513
[12]	Text + Image	RoBERTa+ResNet	0.357
[13]	Text + Image	BERT+NN	0.5915
Proposed study	Text	CNN	0.8352

5. CONCLUSION AND RECOMMENDATION

To conclude, in this study, we developed several models for memotion analysis using different deep learning models for unimodal (text, image) data and multimodal data. The study aimed to investigate the impact of the textual and visual data on the memotion analysis, and we develop a model that can effectively perform the sentiment classification. The study found that text can be used to perform sentiment analysis more effectively as compared to the other data modality i.e., only image and multimodal (image and text). CNN model with Glove embedding outperformed the other models in terms of accuracy, F1-macro, precision, and recall. Furthermore, the proposed model outperformed the baseline studies for SemEval 2020 Task 8, A, i.e., memotion analysis (positive, neutral, negative). Despite these advantages, there is a further need for improvement. As the dataset is highly skewed towards the positive class, there is a need to investigate the performance of the proposed model using a balanced dataset. Lastly, topic modeling needs to be integrated into social images for domain awareness.

REFERENCES

[1] Bargh, J.A., McKenna, K.Y. (2004). The Internet and social life. Annual Review of Psychology, 55: 573-590. https://doi.org/10.1146/annurev.psych.55.090902.141922

[2] Williams, A., Oliver, C., Aumer, K., Meyers, C. (2016). Racial microaggressions and perceptions of Internet memes. Computers in Human Behavior, 63: 424-432. https://doi.org/10.1016/j.chb.2016.05.067

[3] Das, S.D., Mandal, S. (2020). Team neuro at SemEval-2020 task 8: Multi-modal fine grain emotion classification of memes using multitask learning. arXiv preprint arXiv: 2005.10915. https://arxiv.org/abs/2005.10915.

[4] Peirson V, A.L., Tolunay, E.M. (2018). Dank learning: Generating memes using deep neural networks. arXiv preprint arXiv: 1806.04510. https://arxiv.org/abs/1806.04510.

[5] Sharma, C., Bhageria, D., Scott, W., Pykl, S., Das, A., Chakraborty, T., Pulabaigari, V., Gamback, B. (2020). SemEval-2020 Task 8: Memotion Analysis--The Visuo-Lingual Metaphor! arXiv preprint arXiv: 2008.03781. https://arxiv.org/abs/2008.03781.

[6] Sonnad, N. (2018). The world's biggest meme is the word 'meme' itself. https://qz.com/1324334/memes-around-the-world-the-worlds-biggest-meme-is-the-word-meme-itself/, accessed on Jun. 01, 2021.

[7] Gal, N., Shifman, L., Kampf, Z. (2016). "It gets better": Internet memes and the construction of collective identity. New Media & Society, 18(8): 1698-1714. https://doi.org/10.1177/1461444814568784

[8] Gundapu, S., Mamidi, R. (2020). Gundapusunil at SemEval-2020 task 8: Multimodal memotion analysis. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1112-1119.

[9] Bonheme, L., Grzes, M. (2020). SESAM at SemEval-2020 task 8: Investigating the relationship between image and text in sentiment analysis of memes. Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 804-816. https://doi.org/10.18653/v1/2020.semeval-1.102

[10] Keswani, V., Singh, S., Agarwal, S., Modi, A. (2020). IITK at SemEval-2020 task 8: Unimodal and bimodal sentiment analysis of internet memes. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1135-1140.

[11] Bejan, I. (2020). MemoSYS at SemEval-2020 task 8: Multimodal emotion analysis in memes. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1172-1178.

[12] Gupta, P., Gupta, H., Sinha, A. (2020). DSC IIT-ISM at SemeVal-2020 task 8: Bi-fusion techniques for deep meme emotion analysis. arXiv preprint arXiv: 2008.00825. https://arxiv.org/abs/2008.00825.

[13] Singh, P., Bauwelinck, N., Lefever, E. (2020). LT3 at SemeVal-2020 task 8: multi-modal multi-task learning for memotion analysis. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1155-1162.

[14] French, J.H. (2017). Image-based memes as sentiment predictors. In 2017 International Conference on Information Society (i-Society), Dublin, Ireland, pp. 80-85. https://doi.org/10.23919/i-Society.2017.8354676

[15] CodaLab Competition. https://competitions.codalab.org/competitions/20629, accessed on Jun. 01, 2021.

[16] O'Shea, K., Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458. https://arxiv.org/abs/1511.08458.

[17] Zhou, C., Sun, C., Liu, Z., Lau, F. (2015). A C-LSTM neural network for text classification. arXiv preprint arXiv:1511.08630. https://arxiv.org/abs/1511.08630.

[18] Gulli, A., Pal, S. (2017). Deep Learning with Keras. Packt Publishing Ltd.

[19] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556. https://arxiv.org/abs/1409.1556.

[20] Tan, M., Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning, pp. 6105-6114. https://doi.org/10.48550/arXiv.1905.11946