ENSEMBLE MODELS FOR MULTIMODAL SENTIMENT ANALYSIS USING TEXTUAL AND IMAGE FUSION

Radu-Daniel BOLCAȘ[1], Mihai CIUC[2], Eduard-Cristian POPOVICI[3]

Sentiment analysis is an evolving field attracting significant research interest. Multimodal sentiment analysis (MSA) integrates various data forms, like text for emotion recognition and images for facial emotion recognition (FER), to process diverse input modalities. This paper introduces ImaText, a novel dataset for emotion recognition combining text and images from DailyDialog and FER2013. By leveraging these datasets, the study aims to improve model accuracy and robustness against noise and missing data. The proposed multimodal model and dataset provide a fresh perspective on classifying text and image data simultaneously.

Keywords: Multimodal learning, Convolutional Neural Networks, Image Classification, Text Classification

1. Introduction

Sentiment analysis is an evolving field that has won the attention of many researchers. It is a complex task which spans on multiple categories. There is the psychological aspect and also the technical aspect. Depending on which form the emotions are presented a new category can be taken in consideration. In this paper the main two forms of emotions recognition will be facial emotion recognition (FER) and text emotion recognition.

From a psychological perspective, the research conducted by Ekman and Friesen has significantly influenced the development of sentiment models [1]. In their 1971 study, Ekman and Friesen identified a limited set of basic emotions consistently expressed across various cultures and societies. These emotions include anger, happiness, disgust, surprise, sadness, and fear.

From a technical perspective, convolutional neural networks (CNNs) have shown considerable promise in achieving strong results in models using image and text. Consequently, most current research utilizes CNNs. Despite their excellent

[1] Ph.D. student, National University of Science and Technology POLITEHNICA Bucharest, Romania, e-mail: radu_daniel.bolcas@stud.etti.upb.ro

[2] Professor, National University of Science and Technology POLITEHNICA Bucharest, Romania, e-mail: mihai.ciuc@upb.com

[3] Assoc. Professor, National University of Science and Technology POLITEHNICA Bucharest, Romania, e-mail: eduard.popovici@upb.ro

performance, the training process is time-intensive, requiring large datasets and numerous layers and neurons to effectively extract information from the data.

This paper introduces a fresh perspective on emotion recognition through the utilization of a multimodal deep learning strategy to augment performance. While multimodal learning is not entirely novel, its application in sentiment analysis has recently captured the attention of researchers. Multimodal sentiment analysis (MSA) entails incorporating various forms of data, including images, text, audio, or video, to process multiple modalities of input or output. Through the integration of diverse modalities, the model's capabilities can be significantly enriched.

In one paper, Nawaz et al. [2] combined an image model extracted from Navigator-Teacher Scrutinizer Network (NTS-Net) [3] and a model obtained through Bidirectional Encoder Representations from Transformers (BERT) [4] which is a Large Language Model (LLM) to recognize birds on the Caltech-UCSD Birds (CUB-200-2011) dataset [5]. The image model achieved an accuracy of 87.50%, while the text model obtained 65% accuracy. Upon combining both, the authors achieved an improved accuracy of 96.81% by bringing the multiple information sources together.

In the context of multimodal learning, the process begins by setting up or developing a model for each modality. Subsequently, the focus shifts to determining the most effective approach for integrating these diverse modalities. Existing literature commonly outlines two fusion levels: feature level, often termed early fusion, and decision level, also known as late fusion.

In another study conducted by Gallo et al. [6], multimodal learning was applied to classify the UPMC Food-101 dataset [7], which comprises images and text descriptions of foods. Employing BERT [8] and CNNs, they explored various multimodal strategies. Through a fusion of BERT and Long short-term memory (LSTM) for the text component and an InceptionV3 based model for images [9], they achieved promising outcomes using an early fusion approach.

A study conducted by Liu et al. [10], presents an ensemble approach to MSA by employing pretrained models on textual by using BERT and ChatGPT-2 [11] and video data using ResNet [12] and VGG [13]. The study demonstrates that combining these modalities enhances the accuracy and reliability of sentiment predictions. The ensemble models show improved versatility and precision in emotion recognition tasks, highlighting the effectiveness of multimodal data fusion in sentiment analysis.

In a paper by Pereira et al., multimodal emotion recognition is integrated with conversational agents to enhance human-computer interaction [14]. The study combines text, voice, and vision data to develop an empathetic conversational agent capable of understanding and responding to human emotions. The research emphasizes the advantages of using multiple modalities to improve the performance and empathy of conversational agents.

In another study a different approach was used by Tzirakis et al. [15], where they propose an advanced emotion recognition system that utilizes a CNN for auditory features and a 50-layer deep residual network for visual features, combined with a LSTM network to model context and handle outliers, achieving good performance on the RECOLA dataset [16].

This paper presents a novel approach to multimodal sentiment analysis by merging two distinct datasets: DailyDialog [17], which contains text data, and FER2013 [18], which includes image data. The resulting dataset, named ImaText, is combined with a multimodal model that achieves an accuracy of 70.19% on this newly created database. To the best of the authors' knowledge, ImaText is the first multimodal dataset designed specifically for sentiment analysis that includes only text and images.

2. Datasets and data preprocessing

For multimodal learning, it is essential to have a dataset encompassing various modalities of input or output. Most existing datasets include video (both image and speech) and perform various operations on these to create Multimodal Sentiment Analysis (MSA) based on video and the human voice, which can convey information through tone, vocal inflections, and other features. The requirement for a large amount of data for training necessitates the use of powerful hardware and high-end dedicated graphical chips. Simple datasets containing only images and text are scarce to non-existent.

This paper proposes a fusion between two datasets, one containing emotion labelled text and a second one containing labelled images of facial emotions. The resulting database consists of a CSV file and also a directory structure in which the images are located.

DailyDialog [17] is a text dataset that comprises 13,118 dialogues, divided into a training set of 11,118 dialogues, and validation and test sets each containing 1,000 dialogues. On average, each dialogue features approximately 8 speaker turns, with around 15 tokens per turn. The DailyDialog dataset includes: anger, disgust, fear, happiness, sadness, surprise, and neutral. Each sentence is labelled with an emotion regardless of the speaker or dialogue.

FER2013 [18] is the second dataset used for facial emotion recognition (FER), and consists of 35,888 images depicting seven distinct emotions: anger, neutral, disgust, fear, happiness, sadness, and surprise. These images are originally categorized into three subsets: training, validation, and testing.

The pre-processing steps were as follows: First, the DailyDialog dataset was read, and each sentence along with its corresponding emotion was extracted.

Neutral emotions, which constituted 83% of all sentences, were removed as they introduced significant bias. This resulted in 17,407 texts distributed as follows: 1,022 “angry”, 353 “disgust”, 174 “fear”, 12,885 “happy”, 1,150 “sad”, and 1,823 “surprise”. Next, the data was augmented using two strategies: synonym replacement and random word swapping within sentences. These augmentation strategies increased the total number of entries to 90,332, with the following distribution: 5,313 “angry”, 1,899 “disgust”, 950 “fear”, 66,410 “happy”, 6,357 “sad”, and 9,403 “surprise”. The two databases were merged based on labels, and slight modifications to the labels in the DailyDialog dataset were necessary to match FER2013 such as “sad” and “sadness”. Corresponding image names were then assigned, to all text emotions until all images or texts were parsed, and the remaining entries were discarded. The resulting dataset comprised 25,780 entries across six emotions and was saved for the multimodal learning. The final distributions is 4,865 “angry”, 555 “disgust”, 1,255 “fear”, 8,910 “happy”, 6,190 “sad”, and 4,005 “surprise” and the data is split using train_test_split. The newly created dataset named ImaText, consists of a CSV file that includes image names, sentences, and their corresponding labels as it can be seen in Table 1. As stated above, it has a second component resembling the FER2013 directory structure, which contains the actual images.

Table 1

ImaText Dataset

Image Path	Text	Emotion
Training_67023235.jpg	the kitchen stinks	disgust
Training_33607647.jpg	oh let us come in and enjoy yourself	happy
Training_38060810.jpg	that is unfair mom	sad
Training_15431320.jpg	i m very well thank you and you ?	happy
Training_31752247.jpg	what wrong with that ? cigarette is the thing ...	angry

This new dataset ImaText was created by aggregating two existing databases in an innovative way. Consequently, the reported accuracy may differ from that found in the literature of each individual data collection.

3. Architectures and proposed model

The objective of this paper is to explore the development of an effective emotion recognition model using multimodal learning, with experiments tailored to this new dataset. Initially, the process involves reading the defined database, followed by loading and normalizing the text data and image values.

The chosen architecture is a CNN, suitable for both text and images. For the text model, a tokenizer maps words and their occurrences to indexes, followed by padding to the maximum text length. The data is then split into images, texts, and labels, as well as into training and testing sets.

The text model begins with an “input” layer necessary for setting the text tensor in the multimodal later on. Typically, in simpler models, this layer is not required as data is implicitly set in the “fit” method of the model. Following this, an “embedding” layer converts tokens into continuous data representations, enabling feature extraction in subsequent layers.

The subsequent layers consist of a block that includes a one-dimensional convolutional layer, an activation function using ReLU, and a dropout layer. The convolutional layer extracts features, while the activation function introduces non-linearity. Without non-linearity, the model would revert to a linear regression model, limiting its ability to perform complex tasks. The dropout layer helps prevent overfitting by randomly dropping neurons with a certain probability, ensuring that each neuron is less likely to overfit to the data. This block of convolutional, activation, and dropout layers is repeated four times, followed by a max pooling layer that reduces the feature map size while retaining the most significant features. A flatten layer then reshapes the spatial features into a one-dimensional vector, preparing it for the subsequent fully connected dense layer, which classifies the data into six classes. The structures of the text model as well as the parameters of the layers are illustrated in Fig. 1.

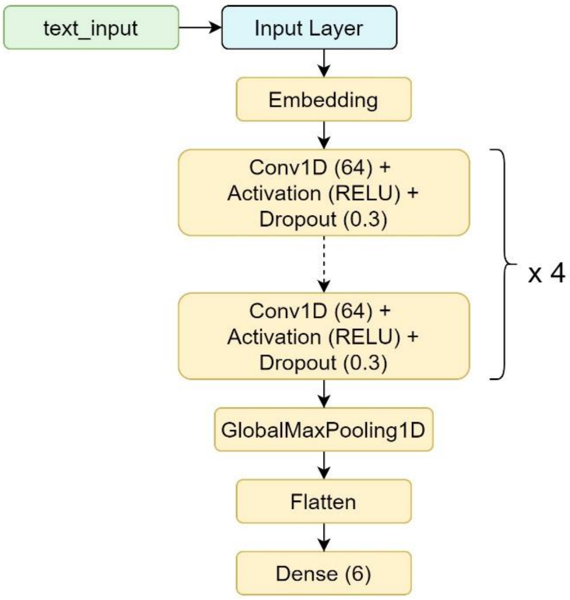
text_input
Input Layer
Embedding
Conv1D (64) + Activation (RELU) + Dropout (0.3)
Conv1D (64) + Activation (RELU) + Dropout (0.3)
x 4
GlobalMaxPooling1D
Flatten
Dense (6)

Fig. 1. The text component of the multimodal model

The image model begins with an “input” layer similar to the text model for the same idea. It’s useful to define to be able to set the image tensor to the model as “layers are recursively composable: If you assign a Layer instance as an attribute of another Layer, the outer layer will start tracking the weights created by the inner layer" [19].

The subsequent layers include a block composed of a two-dimensional convolutional layer followed by a max pooling layer. The convolutional layer extracts features from the normalized images, while the pooling layer reduces and compresses the data from the convolutional layer. This block is repeated three times. A flatten layer at the end restructures the data, preparing it for the fully connected dense layer, which classifies the emotions. These layers combinations have proven effective in literature, as demonstrated by Khaireddin and Chen [20], who used a similar block with two convolutional layers followed by pooling, achieving good accuracy on the FER2013 dataset. Their model includes an additional convolutional layer compared to the image model presented in this paper. Another distinction is in the selection of fully connected layers at the end of the network; their model proposes three layers, whereas this paper suggests a single layer. The complete architecture as well as the parameters used can be seen in Fig. 2.

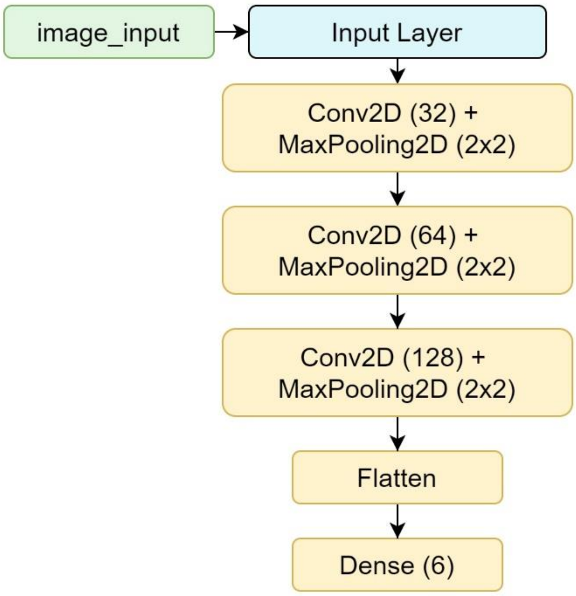

Fig. 2. The image component of the multimodal model

To develop a multimodal model, the individually created text and image models need to be integrated. This is accomplished by a "concatenate" layer that merges the outputs of the two models, forming a unified output within a single model. Subsequently, two fully connected layers are added, with the final layer consisting of six classes, each representing a category of emotion. The activation function used is RELU, except for the last layer, where Softmax provides better performance. The final model structure can be seen in Fig. 3.

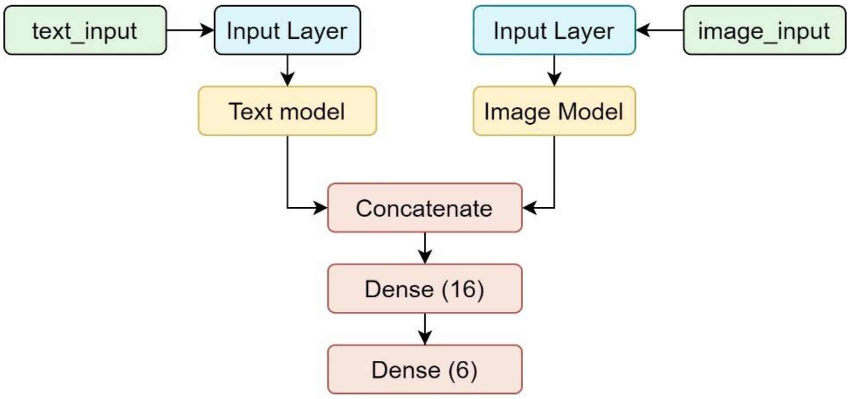

Fig. 3. The proposed multimodal

4. Results and Analysis

An initial approach involved incorporating high-performing image models into a multimodal framework, but this did not lead to improved performance. In one experiment, incorporating the image model created by Khaireddin and Chen [20] not only failed to enhance performance but proved incompatible with the dataset and the multimodal approach; thus resulting in fast overfitting and the loss function returning NAN (Not A Number). This typically occurs when training doesn't converge, leading the cost to explode to infinity, or when an invalid operation, such as divide-by-zero or taking the log of zero, is performed during the processing of the cost or activation function. Since log of zero is negative infinity, when training a model a highly skewed output distribution can be calculated as a result. To address this, adding a small number like 1e-8 to the output probability could help prevent this issue. Another possible cause could be a high learning rate; however, since the Adam optimizer, an adaptive algorithm, was used, this is unlikely the issue. As this approach began to exhibit a poor performance and errors becoming more frequent, the decision to conclude the research for the Khaireddin and Chen model was taken. It was decided to transition to the proposed image model.

The proposed multimodal model performed well on the dataset created. Various experiments were conducted during which the model underwent adaptations and improvements, including testing different optimizers and performing hyperparameter tuning. The multimodal was trained using early-level fusion, combining raw data or features from different modalities at the initial stages, and achieved a peak validation accuracy of 70.19% as illustrated in Fig. 4, while the loss graph is presented in Fig. 5.

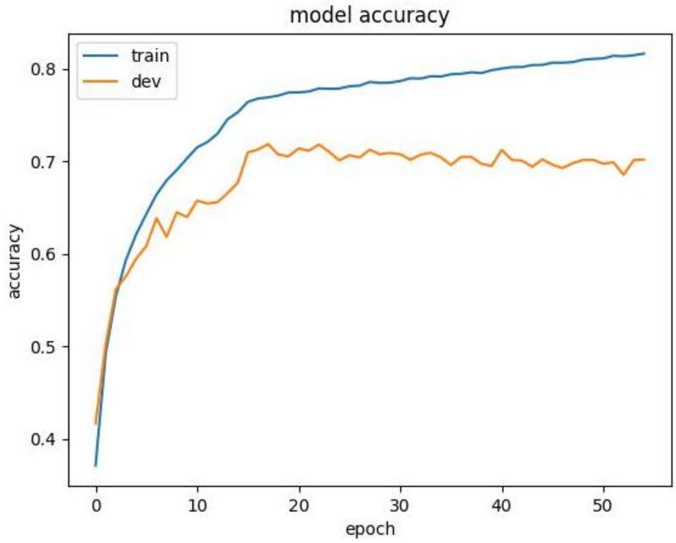

Fig. 4. Accuracy of proposed model

The optimizers considered were Adam, RMSprop, Adagrad, and Stochastic Gradient Descent (SGD). Among these, Adam consistently delivered the best performance in conjunction with other parameters. The model, trained with both image and text data, showed slight overfitting beginning at 15 epochs, which became pronounced by 25 epochs. The optimal performance was observed at around 25 epochs.

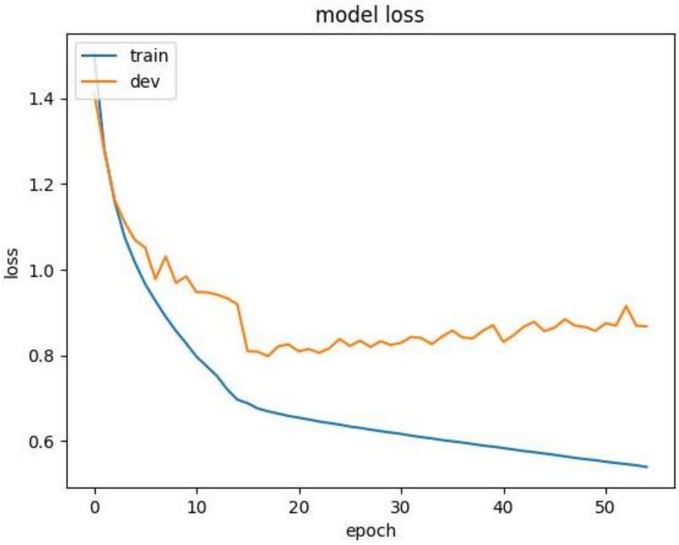

Fig. 5. Loss of proposed model

An interesting observation was noted in the precision, recall, and F1 metrics, as shown in Table 2. Due to class imbalance, the model disregarded the "fear" and "disgust" emotions.

Table 2

Emotion Recognition Multimodal Metrics

Emotion	Precision	Recall	F1	Support
angry	0.55	0.71	0.62	973
disgust	0.50	0.01	0.02	111
fear	0.20	0.00	0.01	251
happy	0.88	0.80	0.84	1782
sad	0.63	0.65	0.64	1238
surprise	0.70	0.87	0.77	801
accuracy			0.70	5156
macro-average	0.58	0.51	0.48	5156
weighted average	0.69	0.70	0.68	5156

As those emotions consist less than 5% of all the values, the model performed as expected by ignoring them. The better approach to improve the recognition of those emotions is to provide better augmentation of these two specific classes or combining the dataset with a third one to balance the classes with fewer entries. The weighted average is 69% for precision and 70% for recall, and the F1 score is 68%.

The confusion matrix displayed in Fig. 6, reveals several more insights into the performance of the multimodal emotion recognition model.

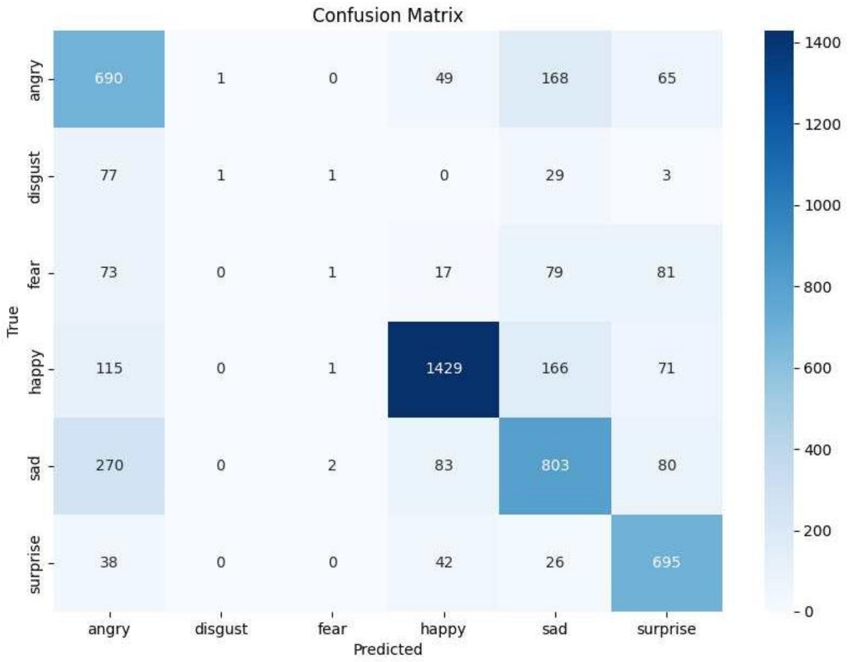

Fig. 6. Multimodal confusion matrix

The model exhibits strong accuracy in predicting "happy" and "surprise" emotions, as evidenced by the high number of correct predictions (1429 for "happy" and 695 for "surprise"). However, the model struggles with "fear" and "disgust," often misclassifying these emotions or not recognizing them at all, which aligns with the earlier observation of class imbalance. For instance, "fear" is frequently misclassified as "happy" or "sad," and "disgust" is often misidentified as "angry" or "sad." To address this issue, future efforts might focus on augmenting data for these underrepresented classes or incorporating additional datasets to balance the entries, thereby improving overall model performance and ensuring a more accurate and robust emotion classification.

Multimodal models offer a comprehensive understanding of content by integrating diverse data types, enabling them to clarify ambiguous contexts through multiple sources. Leveraging complementary information from various data types, multimodals often achieve higher accuracy and demonstrate resilience to noise and missing data.

In this paper, the researchers propose a new dataset and a novel multimodal model. In the created dataset the text information can be enhanced by image features and thus even if there are scenarios where one modality is missing, corrupted or ambiguous, the model can still make reliable predictions by drawing on other modality.

The state of the art of the DailyDialog classification models hovers around 59% [21] while the FER2013 database is around 70% +/- 5% [20,22].

The proposed model achieved 70.19% on the newly created dataset being a value which shows potential to the multimodal approach. Insights gained from one modality can enhance learning in another, facilitating improvements in tasks such as sentiment classification through the incorporation of textual information, and vice versa. Multimodal learning enables the model to leverage knowledge acquired from different sources, leading to enhanced performance across various tasks.

6. Conclusions

This paper has succeeded in implementing a new created dataset obtained by merging DailyDialog and FER2013 alongside with a multimodal model which is able to obtain 70.19% accuracy on the unique ImaText database.

The ImaText dataset consists of merging text with the images. The number of entries obtained are 25,780 entries across six emotions. The final distributions is 4,865 “angry”, 555 “disgust”, 1,255 “fear”, 8,910 “happy”, 6,190 “sad”, and 4,005 “surprise”.

By using this newly created dataset, the researchers have designed the architecture, implemented and fine-tunned the multimodal model to accommodate the novel database. These various experiments (using different layer configuration, optimizers, parameters, etc.) have resulted in an accuracy of 70.19%. To the best of the authors' knowledge, ImaText is the first multimodal dataset designed specifically for sentiment analysis that includes only text and images.

The obtained results are novel and open the way for future research in multimodal sentiment analysis using images and text alike in one dataset.

This model can be also applied to other various domains, including medical diagnosis, psychology, etc. For instance, analysing a person's written responses along with their facial expressions can offer valuable insights into their mental state. Another potential application is improving human-computer interactions, enabling virtual assistants to adapt to users' moods and, in critical scenarios, suggest seeking medical assistance.

This paper presents a fresh perspective on classifying the FER2013 dataset by incorporating novel insights. Leveraging information from a different dataset introduces a novel approach aimed at enhancing performance.

Addressing the scarcity of multimodal datasets, this paper proposes leveraging existing information to augment model accuracy, offering a novel solution to the existing challenge.

REFERENCES

[1] *P. Ekman and W. V. Friesen*, “Constants across cultures in the face and emotion”, in Journal of personality and social psychology, **vol. 17**, no. 2, pp. 124–129, 1971.

[2] *S. Nawaz, A. Calefati, M. Caraffini, N. Landro, and I. Gallo*, “Are these birds similar: Learning branched networks for fine-grained representations”, in 2019 International Conference on Image and Vision Computing, pp. 1-5, 2019.

[3] *Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao and L. Wang*, “Learning to navigate for fine-grained classification”, in Proceedings of the European Conference on Computer Vision (ECCV), pp. 420-435, 2018.

[4] *C. Alberti, K. Lee, and M. Collins*, “A bert baseline for the natural questions”, arXiv preprint arXiv:1901.08634, 2019.

[5] *C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie*, “The Caltech-UCSD Birds-200-2011 Dataset”, California Institute of Technology, CNS-TR-2011-001, 2011.

[6] *I. Gallo, G. Ria, N. Landro, and R. L. Grassa*, “Image and Text fusion for UPMC Food-101 using BERT and CNNs”, in 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), doi: 10.1109/IVCNZ51579.2020.9290622., pp. 1-6, 2020.

[7] *X. Wang, D. Kumar, N. Thome, M. Cord, and F. Precioso*, “Recipe recognition with large multimodal food dataset”, 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1-6, 2015.

[8] *Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova*, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, **vol. 1**, pp. 4171–4186, doi: 10.18653/v1/N19-1423, 2019.

[9] *C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna*, "Rethinking the inception architecture for computer vision", in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826, 2016.

[10] *Z. Liu, A. Braytee, A. Anaissi, L. Qin G. Zhang, and J. Akram*, "Ensemble Pretrained Models for Multimodal Sentiment Analysis using Textual and Video Data Fusion", in Companion Proceedings of the ACM on Web Conference 2024 (WWW '24), pp. 1841–1848, doi: 10.1145/3589335.3651971, 2024.

[11] *OpenAI*, ChatGPT. https://openai.com/blog/chatgpt. [Accessed on March 2024]

[12] *Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian* Sun, "Deep Residual Learning for Image Recognition", *arXiv preprint; DOI: 10.48550/arXiv.1512.03385*, 2015.

[13] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", *arXiv preprint, DOI: 10.48550/arXiv.1409.1556*, 2014.

[14] *R. Pereira, C. Mendes, N. Costa, L. Frazão, A. Fernández-Caballero and A. Pereira*, "Human-Computer Interaction Approach with Empathic Conversational Agent and Computer Vision", in Artificial Intelligence for Neuroscience and Emotional Systems (IWINAC 2024*)*, pp. 431-440, doi:10.1007/978-3-031-61140-7_41, 2024.

[15] *P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou*, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks", in IEEE Journal of Selected Topics in Signal Processing, **vol. 11**, no. 8, pp. 1301-1309, doi: 10.1109/JSTSP.2017.2764438, 2017.

[16] *F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne*, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions", in Proc. of IEEE Face & Gestures 2013, pp. 22-26, 2013.

[17] *Y. Li, H. Su, X. Shen, W. Li, Z. Cao and S. Niu*, "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset", in Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017), pp. 986-995, 2017.

[18] *I. J. Goodfellow et al.*, "Challenges in Representation Learning: A Report on Three Machine Learning Contests", in Neural Information Processing *(*ICONIP 2013), pp. 117-124, doi: 10.1007/978-3-642-42051-1_16, 2013.

[19] *Tensorflow*. tf.keras.layers.InputLayer, https://www.tensorflow.org/api_docs/python/tf/keras/layers/InputLayer [Accessed in May 2024].

[20] *Y. Khaireddin and Z. Chen*, "Facial Emotion Recognition: State of the Art Performance on FER2013", *To be published in Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.2105.03588 [Accessed on March 2024].*

[21] *Shen Weizhou, Siyue Wu, Yunyi Yang, and Xiaojun Quan*, "Directed Acyclic Graph Network for Conversational Emotion Recognition", in Annual Meeting of the Association for Computational Linguistics, 2021.

[22] *S. Vignesh, M. Savithadevi, M. Sridevi, and R. Sridhar*, "A novel facial emotion recognition model using segmentation VGG-19 architecture", in International Journal of Information Technology, **vol. 15**, pp. 1777–1787, doi: 10.1007/s41870-023-01184-z, 2023.