Displays 87 (2025) 102958

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Beyond visual cues: Emotion recognition in images with text-aware fusion ☆

Kerim Serdar Sungur, Gokhan Bakal *

Department of Computer Engineering, Abdullah Gul University, Barbaros, Erkilet Blvd. Sumer Campus, Kayseri, 38080, Turkey

ARTICLE INFO

Dataset link: https://t.ly/1AWwu

Keywords:
Sentiment analysis
Hybrid model
Image & text processing
Deep learning

ABSTRACT

Sentiment analysis is a widely studied problem for understanding human emotions and potential outcomes. As it can be performed over textual data, working on visual data elements is also critically substantial to examining the current emotional status. In this effort, the aim is to investigate any potential enhancements in sentiment analysis predictions through visual instances by integrating textual data as additional knowledge reflecting the contextual information of the images. Thus, two separate models have been developed as image-processing and text-processing models in which both models were trained on distinct datasets comprising the same five human emotions. Following, the outputs of the individual models' last dense layers are combined to construct the hybrid multimodel empowered by visual and textual components. The fundamental focus is to evaluate the performance of the hybrid model in which the textual knowledge is concatenated with visual data. Essentially, the hybrid model achieved nearly a 3% F1-score improvement compared to the plain image classification model utilizing convolutional neural network architecture. In essence, this research underscores the potency of fusing textual context with visual information to refine sentiment analysis predictions. The findings not only emphasize the potential of a multi-modal approach but also spotlight a promising avenue for future advancements in emotion analysis and understanding.

1. Introduction

Sentiment analysis, a computational effort dedicated to interpreting the intricate emotional nuances embedded mostly within textual representations, whether a sentence, paragraph, or document, has garnered significant attention within distinct research communities [1–3]. Its applicability spans diverse domains, mostly containing customer services, marketing, legal authority services, and the vigilant monitoring of social media landscapes [4,5]. While the traditional scope of sentiment analysis has historically centered on textual data, the contemporary research landscape is witnessing a notable surge in interest in integrating visual data into this analytical framework [6]. This increased attention is grounded in the belief that visual data incorporating elements such as facial expressions and body language, holds the potential to enhance our understanding of an individual's emotional state. This paradigm shift underscores the evolving nature of sentiment analysis and its adaptation to incorporate multi-modal data sources [7].

Within the framework of this paper, our conceptualization unfolds within the confines of a dual-model architecture, an intricately designed construct where the inaugural pillar is anchored in an image-processing model. This model, ingeniously fashioned upon the robust foundations of a convolutional neural network (CNN), harnesses the potency of deep learning to meticulously extract nuanced features latent within visual stimuli [8]. In tandem, the second pillar takes shape as a text-processing model intricately woven into the fabric of a fully-connected neural network (ANN) [9]. This textual counterpart assumes the responsibility of deciphering the descriptive features intricately interwoven within the textual content. Notably, the outputs stemming from these seemingly disparate yet symbiotic models converge harmoniously to craft a unified prediction, emblematic of the symbiosis of both visual and textual cues.

The validation of our proposed model unfolds across a meticulously curated dataset, a rich tapestry interwoven with textual and visual data elements, each adorned with five discrete emotional labels (classes): *happiness, sadness, anger, surprise, and fear*. The empirical validation resonates with resounding endorsement by underscoring the robustness of our hybrid model. A notable accuracy surge of **3%**, when juxtaposed against the standalone image-processing model, stands testament to the efficacy of our proposed framework. This work stands not only as a testament to the efficacy of our hybrid model but also serves as a clarion call by articulating the potential encapsulated within the confluence of visual and textual data as a steadfast approach for reinforcing the foundations of traditional sentiment analysis.

☆ This paper was recommended for publication by Guangtao Zhai.

* Corresponding author.
E-mail address: gokhan.bakal@agu.edu.tr (G. Bakal).

https://doi.org/10.1016/j.displa.2024.102958
Received 20 February 2024; Received in revised form 23 October 2024; Accepted 27 December 2024
Available online 2 January 2025

The implications of our research reverberate across various applications, ranging from enhancing the finesse of customer service interactions to introducing new dimensions to marketing strategies. Furthermore, it extends to deepening the insights derived from the ever-dynamic landscape of social media. The notable contributions of this study, outlined with specificity, are listed below.

- **Multi-modal Fusion for Enhanced Sentiment Analysis**: *The introduced hybrid model deftly fused visual and textual data for sentiment analysis and obtained relatively better accuracy.*
- **Architectural Innovation**: *The dual-pillar design, utilizing a well-known approach, a convolutional neural network (CNN) for image processing and a fully-connected artificial neural network (ANN) for text processing, offers a strategic alignment with the strengths of each modality*.
- **Performance Improvement and Practical Applicability**: *A commendable 3% accuracy improvement against our base image-processing model highlights the model' s prowess in gleaning deeper insights by leveraging both textual and visual data.*

These contributions collectively mark a significant advancement in sentiment analysis methodologies, presenting a compelling case for the integration of visual and textual data. This research paves the way for an enriched understanding of human sentiment and emotional expressions, thereby enhancing the effectiveness of applications across industries that rely on sentiment analysis for informed decision-making.

The rest of the paper is organized as follows. Section 2 introduces general background knowledge about the relevant topics, while Section 3 explains the dataset details used in the experiments. Section 4 describes the hybrid model-building details. Section 5 presents experimental results and comparative evaluations. Finally, Section 6 concludes the study with an overall summary and Section 7 briefly mentions limitations and potential future directions.

2. Background

The exploration of sentiment analysis has, over the years, captivated the research community, primarily honing in on either text or visual data as discrete entities to construct predictive models [10,11]. An illuminating example of harnessing text data for emotion analysis is evident in the work of Wang et al. [12]. Their study introduces a technique that amalgamates lexical and syntactic features, offering a comprehensive model for affect in text. This method has proven effective in detecting emotions within a dataset of tweets, showcasing the versatility of textual data in capturing nuanced emotional states.

In a parallel trajectory of research, there exists a distinct yet interconnected domain that exclusively delves into the utilization of image data for emotion analysis. A notable illustration is found in Levi and Hassner' s [13] study, which introduces a methodology anchored in a convolutional neural network (CNN) and binary patterns. This approach is tailored for recognizing emotions in facial images, highlighting its efficacy in discerning emotions such as happiness, sadness, anger, and disgust. The authors conducted extensive experiments on a sizable dataset of facial images and achieved commendable accuracy in the identification of diverse emotional states encapsulated within the visual modality.

These two distinct research strands, one pivoting around the textual landscape and the other navigating the visual realm, represent complementary approaches to sentiment analysis. By delving into the intricacies of both text and image data, researchers have sought to broaden the scope of emotion analysis, acknowledging the unique strengths each modality brings to the fore. As we delve into the convergence of these two domains, an emergent theme is poised to redefine the landscape of sentiment analysis, presenting novel avenues for constructing robust predictive models that seamlessly integrate both textual and visual cues.

Since the inception of deep learning methods, a revolutionary paradigm has emerged, placing multi-modal approaches at the forefront and advocating for the fusion of multiple models to enhance performance [14–16]. A striking illustration of such a multi-modal initiative is evident in the research conducted by Miller et al. [17]. Their study represents a noteworthy instance wherein the integration of various modalities aims to enrich image classification. The crux of their approach lies in seamlessly integrating natural language understanding and leveraging associated metadata effectively. This integration culminates in the creation of a multi-modal image classification model, seamlessly combining convolutional methodologies with linguistic comprehension drawn from descriptions, titles, and tags.

The amalgamation of ResNet-50 image features and Universal Sentence Encoder embeddings stands out, yielding impressive results with an enhanced Top 5 error rate of 73.05% and a Top 1 error rate of 54.65%, surpassing benchmark results by 1.56%. These outcomes underscore the substantial augmentation that external textual features can bring to image classification. In tandem with this experiment, our study delves into the exploration of multi-modal techniques in the domain of emotion detection. Envisaging a landscape where information fusion from diverse sources, spanning both image and text data, gives rise to more resilient models, our approach proves particularly beneficial in scenarios where images are accompanied by supplementary information such as captions, tags, and descriptions.

Recognizing the limitations of single-modality approaches in sentiment analysis, especially in social media with its inherent variability, You et al. [18] proposed a robust visual–textual sentiment analysis framework. Their approach emphasizes the importance of treating visual and textual information jointly in a structured manner. By leveraging a semantic tree structure derived from sentence parsing, their model aligns textual words with corresponding image regions to achieve a more accurate analysis. Furthermore, they incorporate an attention mechanism with LSTM and an auxiliary semantic learning task to learn a robust joint visual–textual representation. Their findings demonstrate that this structured approach, combined with attention mechanisms, significantly improves sentiment analysis performance compared to existing joint models. Addressing the need for more sophisticated multi-modal sentiment analysis approaches, Huang et al. [19] introduced the Deep Multimodal Attentive Fusion (DMAF) model. This model utilizes a mixed fusion framework to effectively capture the discriminative features and internal correlations between visual and textual content. DMAF employs two unimodal attention models to learn emotion classifiers for image and text modalities independently. Furthermore, an intermediate fusion-based multimodal attention model exploits the inter-modal relationships between visual and textual features. Finally, a late fusion scheme combines the outputs of all three attention models to predict sentiment. Their experimental results highlight the effectiveness of this attentive fusion approach in improving sentiment classification accuracy on both weakly labeled and manually labeled datasets.

Diverging from prior works, our investigation takes a distinctive path by seamlessly integrating textual data in tandem with visual information. Our study stands apart not only in terms of model design and dataset selection but also in its contribution of fresh insights, achieved through the implementation of a deep learning-driven multi-modal architecture. This divergence in methodology and perspective contributes to the evolution of the field, offering novel insights that advance our understanding of the synergies between textual and visual information in deep learning applications.

In addition, recent studies delve into sentiment analysis of video data by amalgamating audio cues [20]. These investigations illuminate that incorporating audio information augments sentiment analysis on video data. However, considerable room for refinement remains. Central challenges encompass the absence of suitable datasets and complexities surrounding testing methodologies for assessing the impact of these markers. As ongoing research showcases promising outcomes, integrating these aspects into future studies holds the potential to amplify and refine the accuracy of sentiment analysis. Further advancement lies in analyzing diverse markers for emotions, extending to the auditory realm.

3. Dataset details

In this experimental hybrid model, we used well-known and publicly available datasets from the Kaggle platform for training, while we manually collected and annotated 100 emotion-reflecting images from the Twitter platform for testing. For the textual data submodel, we employed a valuable resource for Natural Language Processing (NLP) tasks focused on emotion analysis available via the link: https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp. This dataset comprises a diverse collection of textual data, annotated with corresponding emotional labels, making it an ideal asset for training and evaluating machine learning models aimed at understanding and predicting emotions from text [21]. In the dataset, we have 20K instances labeled uniquely by one of the six emotions, including *joy, love, sadness, surprise, anger*, and *fear*. After applying the cleaning step to the dataset, we obtained 14,696 instances to utilize in the experiments. The dataset covers a wide range of emotions which allows researchers, data scientists, and practitioners to develop and fine-tune intelligent algorithms to detect and classify emotions expressed in various contexts, such as social media posts, customer reviews, surveys for services, or other textual forms. This resource holds substantial potential for advancing sentiment analysis and emotion recognition tasks, that contribute to applications ranging from sentiment-aware chatbots to more nuanced customer feedback analysis.

For the image-based submodel, we utilized another substantial image resource for facial expression analysis tasks available via the link: https://www.kaggle.com/datasets/msambare/fer2013. The FER2013 dataset is a collection of 48 × 48 pixel grayscale images of faces with emotion labels. The faces have been automatically registered so that they are centered and occupy approximately the same amount of space in each image. The dataset consists of two sets: a training set of 23,308 examples and a public test set of 5834 examples. The task is to categorize each face based on the emotion shown in the facial expression into one of seven categories: *angry, disgusted, fearful, happy, sad, surprised*, or *neutral*.

Regarding the test set for the proposed model' s evaluation, we collected 100 text and image examples from Twitter containing *anger, fear, happy, sad*, and *surprised* emotion classes. The curated test dataset was subjected to image-based, textual-based, and hybrid multimodal-based models. Following, the obtained performance scores were compared and reported for assessment purposes. Here, the selection of this compact size of the test set is a deliberate choice since the carefully curated nature of the Twitter dataset, with associated textual descriptions, allows for rigorous testing of the model's ability to fuse visual and textual information. Here, employing a compact test set is a deliberate choice since the carefully curated nature of the Twitter dataset, with associated textual descriptions, allows for rigorous testing of the model's ability to fuse visual and textual information.

4. Methods

In this section, we describe the methods used to demonstrate the performance improvement of the multimodel infrastructure. We built two distinct model configurations: **image model** and **multimodel**. The image model was trained on an image dataset, while the multimodel was trained on a dataset containing images and text instances. We evaluated the performance of each model on a held-out test set. The overall proposed method is represented in Fig. 1.

4.1. Preprocessing

Data preprocessing serves as a critical phase in the machine learning pipeline, which holds significant importance in ensuring that the data is clean, consistent, and formatted in a specific way that is intelligible to the model. In light of this fact, we have implemented meticulous preprocessing procedures for both image and text datasets. The primary objective is to cultivate the creation of more accurate models by refining the input data. This comprehensive approach guarantees that the data is thoroughly well-prepared and optimized, setting the stage for subsequent phases of the machine-learning process. The ultimate aim here is to enhance the overall efficiency and accuracy of the models utilized in our study by contributing to the robustness of the analytical framework and the reliability of the outcomes.

4.1.1. Image data preparation

The first step in working with image datasets is often to separate the visual data into different categories. This is typically done by organizing the images into different folders based on their labels. Once the data is separated, a variety of preprocessing steps are applied to ensure consistency and remove any irrelevant or low-quality images. These steps include:

Image resizing: This step involves resizing the images to a consistent size. This is important for ensuring that the images are all processed in the same way and that the model does not overfit to any particular size.

Normalization: This step involves adjusting the values of the image pixels so that they have a mean of 0 and a standard deviation of 1. This is done to ensure that the data is evenly distributed and that the model does not overfit to any particular range of values.

Cleaning: This step involves removing any irrelevant or low-quality images from the dataset. This can include removing images that are corrupted, blurry, or do not contain the target object.

4.1.2. Text data preprocessing

Text data preprocessing is a process of cleaning and transforming unstructured text data into a format that can be understood by machine learning algorithms. The specific steps involved in text data preprocessing vary depending on the specific dataset and the machine learning algorithm being used, but some common steps include:

Removing stop words: Stop words are words that are commonly used in a language but do not carry much meaning, such as "the", "and", and "but". Removing stop words can help to reduce the noise in the data and improve the accuracy of the machine learning model.

Tokenization: Tokenization is the process of breaking down text into individual words or phrases. This can be done using a variety of techniques, such as word boundaries or regular expressions.

Stemming: Stemming is the process of reducing a word to its root form. This can help to improve the accuracy of the machine learning model by grouping together words that have similar meanings.

Lemmatization: Lemmatization is a more sophisticated form of stemming that takes into account the grammatical context of the word. This can help to further improve the accuracy of the machine learning model.

Vectorization: Vectorization is the process of representing text data as a numerical vector. This can be done using a variety of techniques, such as word2vec models.

4.1.3. Data preparation for the multi-odal approach

The same preprocessing steps are applied to the dataset created to test the multi-modal model as the ones applied to the training dataset. This is important to ensure that the test dataset is preprocessed in the same way as the training dataset, so that the model can generalize well to unseen data. For image data, the preprocessing steps applied to the test dataset typically include:

- Resizing the images to a consistent size, such as 48 × 48 dimensions. This ensures that all images in the dataset are of the same size and can be easily processed by the model.

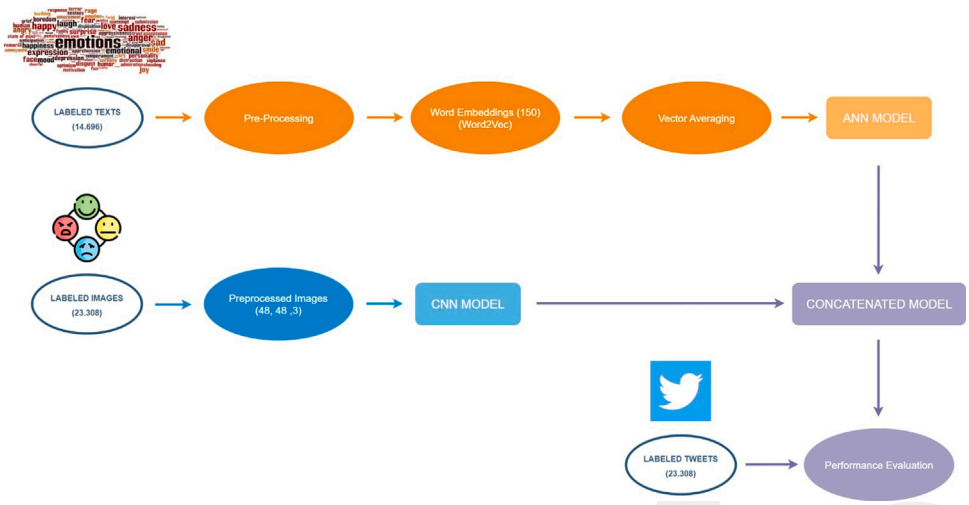

Fig. 1. Overall proposed model architecture representation.

- Rescaling the pixel values to a specific range, such as 0–255. This ensures that the pixel values are all in the same range and can be easily compared by the model.

For text data, the preprocessing steps applied to the test dataset typically include:

- Cleaning the data by removing any unnecessary words, such as stop words. This step is vital because it helps to reduce the amount of noise in the data and makes it easier for the model to extract meaningful information.
- Using a pre-created word embedding space, such as a 150-dimensional word space, to determine the position of the words in the text data. This allows the text to be represented in a numerical format that can be used as input for the multi-modal model.

By applying the same preprocessing steps to both the training and test datasets, we can ensure that the model is trained on data that is representative of the data it will encounter during testing. This helps to improve the accuracy and performance of the model.

4.2. Model construction details

Modeling involves two primary components: crafting the model' s architecture and subsequently training or fitting it. An alternative approach is to construct a conventional model. While traditional methods prove effective for specific tasks, they exhibit several limitations. They typically yield inferior results when applied to various image processing tasks, especially those entailing intricate patterns or image variations. Furthermore, traditional methods struggle with processing extensive datasets and are ill-suited for images with high resolutions and complex features. To address this constraint, rather than opting for a conventional model, a deep learning model was devised employing the TensorFlow framework [22]. TensorFlow offers an array of pre-built functions and tools tailored for constructing deep learning models, particularly well-suited for image processing tasks. In crafting the model' s architecture, the sequential model was employed—a linear stack of layers [23]. Depending on the image complexity and the task, layers such as convolutional layers, pooling layers, and fully connected layers were incorporated into the model. Furthermore, the model' s performance and stability were enhanced through the batch normalization process. After formulating the architectural framework, the consequent phase entailed the training or fitting of the model. This procedure encompassed the systematic input of images or textual data into the model, subsequently fine-tuning the model' s parameters to mitigate the disparity between the predicted and actual output. This optimization process was executed through the well-established method known as backpropagation [24].

In this specific experimental work, a series of distinct models were meticulously devised and utilized to enhance the accuracy and effectiveness of emotion analysis from human facial expressions. These models were meticulously crafted and refined, each with unique architectural features and parameters, to yield improved prediction rates and robust performance in facial emotion analysis. The subsequent sections will delve into the intricacies of these individual models, elucidating their design principles with strategies and their respective contributions to advancing the state of emotion analysis from human faces.

4.2.1. Image data model

The image model operates on a meticulously preprocessed dataset, which undergoes a series of essential preprocessing steps, including extraneous data removal and uniform resizing to a standardized resolution of 48 × 48 pixels. This process also categorizes the images, resulting in a dataset comprising 23,308 photographs earmarked for the training phase. For instance, a randomly picked example of resized images in the dataset is shown in Fig. 2.

To further elaborate on the architecture of the image model, we employed a sequential CNN model comprising three convolutional layers, each followed by a max-pooling layer for downsampling. The convolutional layers utilized ReLU activation functions and employed filters with sizes of 32, 64, and 128, respectively. Subsequently, a flattened layer converted the multi-dimensional feature maps into a one-dimensional vector, fed into a dense layer with 128 neurons and ReLU activation. The final layer of the image model was a dense layer with five neurons and a softmax activation function to predict the five emotion categories. During training, the Adam optimizer with a learning rate of 0.001 was employed, and the model was trained for five epochs with a batch size of 32. Plus, an early stopping mechanism was implemented to prevent overfitting.

4.2.2. Textual data model

As demonstrated in the provided illustration in Fig. 1, a textual dataset has been meticulously curated, undergoing prior preparation and division into distinct test, train, and validation subsets, mirroring the structure of the image dataset. Similarly, this textual dataset encompasses data associated with five distinct emotions.

The initial phase of data processing comprises pre-processing steps, where all words within the textual content are systematically converted to lowercase, and extraneous articles such as "a", "an", and "the"

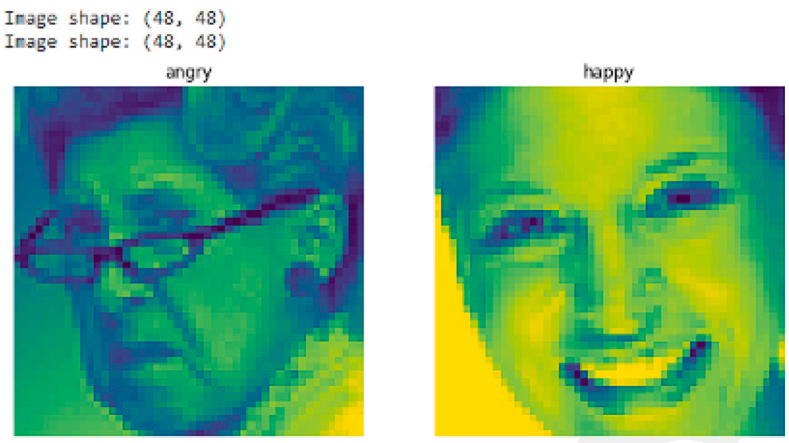

Fig. 2. n example picture drawn from the preprocessed dataset.

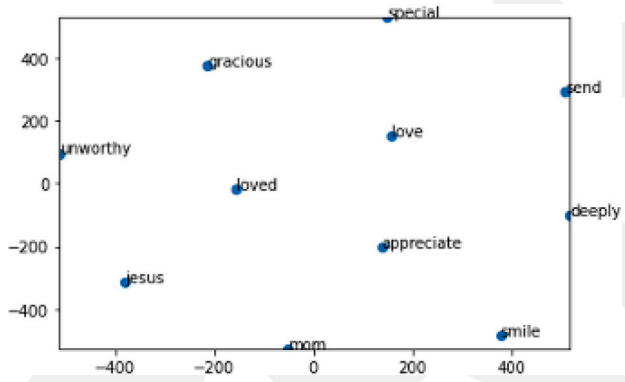

Fig. 3. A sample embedding space for the word "love".

$$\begin{bmatrix} W_{11} \\ W_{12} \\ \\ W_{1n} \end{bmatrix} + \begin{bmatrix} W_{21} \\ W_{22} \\ \\ W_{2n} \end{bmatrix} + \cdots + \begin{bmatrix} W_{n1} \\ W_{n2} \\ \\ W_{nn} \end{bmatrix} = \begin{bmatrix} \frac{W_{11}+W_{21}+\ldots+W_{n1}}{n} \\ \\ \frac{W_{1n}+W_{2n}+\ldots+W_{nn}}{n} \end{bmatrix}$$

Fig. 4. The representation of the average word embeddings calculation formula.

are removed. Plus, significant efforts were devoted to constructing a word embedding space, involving the configuration of parameters including a window size of 5, continuous bag-of-words algorithm, and an embedding dimension of 150 via the utilization of the Word2Vec model architecture [25]. The Word2Vec model represents a family of shallow neural network models that efficiently learn distributed vector representations of words by capturing their semantic relationships and context within the text. To provide a solid example, the spatial representation of the word "love" in this embedding space is shown in Fig. 3.

The primary aim underlying the creation of this embedding space is to furnish numerical representations for individual words. This approach facilitates the computation of a sentence' s position within the expansive 150-dimensional embedding space by calculating the average of the word embedding representations contained by the sentence. This mathematical calculation is presented in Fig. 4.

Once the averages have been computed, they are subsequently input into an Artificial Neural Network (ANN) model exclusively composed of fully connected dense layers [26,27]. Furthermore, the categories have been numerically encoded through the `LabelEncoder` method to represent them across five distinct classes.

4.2.3. Multi-modal fusion of image and text data

As visually depicted in Fig. 1, our proposed multi-model incorporates two distinct input pathways—one for handling image data and the other for processing textual data. Both of these input channels undergo a series of pre-processing stages to ensure proper formatting, preparing the image and text data for subsequent feature extraction. Notably, the pre-trained image and text models are configured as non-trainable, meaning that their weights and parameters remain static throughout the model's subsequent training phase. The image and text data are then individually processed through their respective models by extracting relevant features unique to each modality. Following this step, the resulting feature vectors are concatenated into a unified vector representation.

The concluding step involves the utilization of a dense layer with a softmax activation function, facilitating predictions across five possible categories. This design renders the model suitable for multi-class classification tasks. The amalgamation of both image and text modalities within this composite architecture epitomizes a flexible strategy for harnessing diverse data types in the realm of machine learning applications. This approach holds the promise of enhancing predictive performance, showcasing the adaptability of our model in addressing a variety of classification tasks involving multiple categories.

5. Results & discussion

Our research delved into a comprehensive examination of the performance enhancements achieved through the implementation of our

Table 1
Performance scores of the image-only and multimodel fusion models.

		Class	Performance Metrics		
			Precision	Recall	F1-score
Image-only Model Results	Individual Class Metrics	angry	0.15	0.23	0.18
		fear	0.25	0.10	0.14
		happy	0.73	0.73	0.73
		sad	0.39	0.58	0.47
		surprise	0.67	0.20	0.31
	Overall Average Metrics	macro avg	0.44	0.37	0.37
		weighted avg	0.56	0.53	0.52
	Accuracy	0.53			
Multimodel Fusion Results	Individual Class Metrics	angry	0.22	0.15	0.18
		fear	0.25	0.20	0.22
		happy	0.74	0.82	**0.78**
		sad	0.32	0.50	0.39
		surprise	0.67	0.20	0.31
	Overall Average Metrics	macro avg	0.44	0.37	0.37
		weighted avg	0.56	0.57	**0.55**
	Accuracy	**0.57**			

proposed multi-modal fusion architecture. To observe the model's effectiveness, we meticulously subjected the test dataset, sourced from Twitter, to rigorous evaluations using both the image-only (base) model and the proposed multi-modal architecture. Distinctly, the base image-only model underwent training solely on image data, while the multi-modal counterpart underwent training on both image and text data. The discerned outcomes of this evaluation, encapsulated essential performance metrics, such as precision, recall, F1-score, and accuracy, have been systematically tabulated in Table 1 for elucidation and in-depth analysis.

The image-only model exhibits varied precision scores across different emotion categories. It performs relatively well in classifying "happy" emotions with a precision of 0.73, indicating a solid ability to correctly identify happy expressions. However, it struggles to distinguish "angry" and "fear" emotions, with precision scores of 0.15 and 0.25, respectively, suggesting a higher rate of false positives in these categories. However, the multi-modal fusion model demonstrates improved precision scores compared to the image-only model. It achieves a precision of 0.74 for "happy" emotions, showcasing its ability to accurately classify happy classes. Additionally, it maintains precision scores above 0.20 for "angry", "fear", and "surprise" emotions.

Regarding the recall metric, the model effectively captures "happy" emotions with a score of 0.73, reflecting its ability to detect the most happy expressions. Nevertheless, it faces challenges in identifying "fear" and "surprise" emotions, with recall scores of 0.10 and 0.20, respectively, indicating a notable number of false negatives. Unlike the image-only model, the proposed multimodel exhibits favorable recall rates for most emotion categories. Notably, it attains a recall of 0.82 for "happy" emotions, indicating its proficiency in correctly identifying instances of happiness. However, it still faces challenges in "angry" and "fear" emotion detection, with recall scores of 0.15 and 0.20, respectively.

Concerning the F1-score comparison, the F1-scores for the image-only model display a balance between precision and recall. It achieves the highest F1-score of 0.73 for "happy" emotions, signifying a substantial harmonic mean of precision and recall. However, the F1 scores are 0.18 and 0.14 for "angry" and "fear" emotions, respectively, indicating much room for improvement. On the other hand, the F1 scores for the multi-modal fusion model demonstrate a well-balanced performance. It achieves an F1-score of 0.78 for "happy" emotions, indicating a robust trade-off between precision and recall. While it performs better than the image-only model in most categories, there needs to be the same improvements in "angry" and "fear" emotion classification.

In addition to evaluating the performance of individual emotion categories, macro, and weighted averages can provide comprehensive insights into the overall classification effectiveness of both the image-only and multi-modal fusion of image and text data models. The point is that macro averaging calculates the average of the precision and recall scores for each emotion category, regardless of their prevalence in the dataset, while weighted averaging calculates the average of the precision and recall scores for each emotion category, weighted by their prevalence in the dataset. This definition means that more prevalent classes have a superior impact on the overall average, which can be meaningful for evaluating the model on the most common categories.

The macro average for precision, recall, and F1-score hovers around 0.44, 0.37, and 0.37, respectively. This outcome indicates a balanced evaluation of model performances across all emotion categories, with neither positive nor negative biases. The consistent macro average values suggest that the models exhibit a relatively similar level of performance across different emotions. Unlike the macro average scores, the weighted average accounts for class imbalances by considering the number of instances in each emotion category. For both models, the weighted average for precision, recall, and F1-score is higher than the macro average, hovering around 0.56, 0.57, and 0.55, respectively. These results signify that the models perform better when evaluated with consideration of class frequencies. The higher weighted average scores indicate that the models excel in classifying the more prevalent emotions in the dataset, such as "happy". However, they still encounter challenges in classifying less frequent emotions, such as "angry" and "fear". While the multi-modal approach demonstrated a modest overall improvement, the relatively low accuracy in classifying "angry" and "fear" highlights challenges in distinguishing subtle emotional expressions. The reason is likely due to a combination of factors, including the inherent ambiguity of these emotions and potential limitations in the dataset's representation of these specific affective states.

Considering the overall accuracy performance metric, it becomes evident that the multi-modal Fusion model outperforms the image-only model. The image-only model achieves an overall accuracy of 0.53, indicating its ability to correctly classify emotions in 53% of the instances within the test dataset. In contrast, the multi-modal fusion model demonstrates a slightly higher overall accuracy of 0.57, signifying a modest yet discernible improvement over its single-modal counterpart. This enhanced accuracy proves the advantages of leveraging image and text data modalities by reaffirming the potential for more robust and accurate multi-class emotion classification results.

When we evaluate the model performance using the confusion matrix as illustrated in Fig. 5, the model exhibited high precision and recall for the "happy" emotion (74% and 82%, respectively, F1-score: 78%), indicating strong performance in this category. Conversely, performance was significantly lower for "angry", "fear", "sad", and "surprise", which suggests challenges in distinguishing between these emotions. The uneven performance across emotion categories indicates a

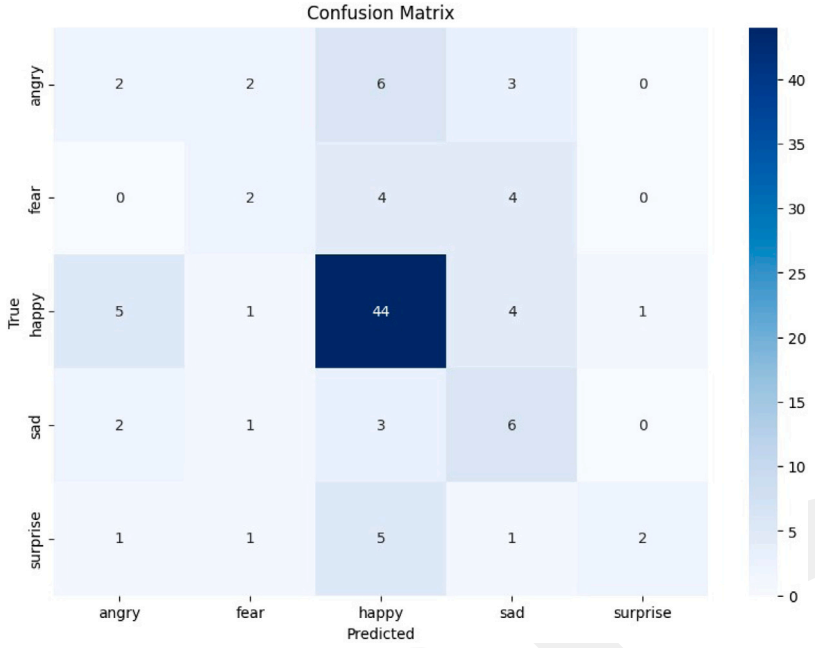

Fig. 5. The confusion matrix of the multi-model configuration.

need for targeted improvements to address the model's limitations in distinguishing between these more ambiguous emotional states.

It is also important to note that our evaluation primarily focused on assessing the performance of our proposed method on a self-built dataset curated from Twitter. This dataset was specifically designed to reflect the real-world scenario of images with associated textual descriptions, commonly found on social media platforms. While comparisons with existing methods on publicly available datasets are valuable, such comparisons might not be directly applicable in this context due to potential differences in emotion labels, modalities, and dataset characteristics. Our controlled comparison with a baseline image-only model provides clear evidence for the effectiveness of incorporating textual information for sentiment analysis in this specific scenario.

6. Conclusion

In this experimental investigation, our focus was directed toward evaluating the effectiveness of two distinct models employed in multi-class emotion analysis: the image-only model and the multi-modal fusion model, which seamlessly integrates both image and text data. Our thoroughgoing examination encompassed a spectrum of performance metrics, including precision, recall, F1-score, accuracy, and their macro and weighted averages. The findings brought to light the clear superiority of the multi-modal fusion model, manifesting significant improvements over its image-only counterpart across various evaluation criteria, including precision, recall, F1 score, and overall accuracy. Remarkably, the multi-modal fusion model demonstrated exceptional proficiency in precisely classifying "happy" emotions, underscoring its ability to achieve a well-balanced performance across diverse emotion categories. This in-depth analysis contributes valuable insights into the efficacy of multi-modal approaches in enhancing emotion analysis, which emphasizes the multifaceted advantages of integrating both image and text data for more robust and accurate outcomes.

In conclusion, the proposed multi-modal fusion model, empowered by the collective strengths of both image and text data modalities, stands out as the most suitable option for conducting multiclass emotion analysis. Nevertheless, it is imperative to recognize the persistent endeavors aimed at achieving further enhancements in performance, especially when less prevalent emotions are concerned. This study not only highlights the considerable potential of multi-modal approaches in propelling the field of emotion analysis forward but also indicates promising directions for future research initiatives and optimization endeavors. The outcomes underscore the dynamic nature of emotion analysis which rests on the foundation for ongoing exploration and refinement in this ever-evolving domain.

7. Limitations & future directions

This study's findings are subject to certain limitations stemming primarily from the characteristics of the datasets employed. While publicly available, the image dataset may not fully represent the diversity of real-world scenarios. Similarly, the textual data, though comprehensive, might not perfectly capture the nuanced complexities of human emotional expression in all contexts. These limitations warrant consideration when interpreting the results. The potential future directions for further improvements are listed below.

- **Exploration of alternative model architectures:** One future research could investigate the performance of alternative deep learning architectures, such as transformers or more sophisticated hybrid models incorporating attention mechanisms, for enhanced multi-modal fusion and improved emotion classification accuracy. This approach would involve exploring architectures capable of better handling the complexities of visual and textual data integration.
- **Investigation of larger and more diverse datasets:** The current study utilized a specific dataset with certain limitations. Another promising future direction should focus on expanding the scope of the dataset to encompass an increased sample size and greater diversity in terms of image and text modalities. Integrating data from various sources and representing a broader spectrum of emotions would contribute to a more robust and generalizable model.
- **Incorporation of additional modalities:** This study focused on a bimodal approach combining visual and textual information. Further research could investigate the potential benefits of incorporating additional modalities, such as audio data (e.g., tone of voice, background sounds) or physiological signals (e.g., heart rate, skin conductance). This multi-modal fusion approach could capture a richer representation of emotional states and lead to more accurate emotion classification.

CRediT authorship contribution statement

Kerim Serdar Sungur: Formal analysis, Data curation, Investigation, Methodology, Software, Visualization, Writing – original draft. **Gokhan Bakal:** Conceptualization, Formal analysis, Investigation, Project administration, Software, Supervision, Writing – review & editing.

Funding

No funds, grants, or other types of support were received.

Ethical approval

Not Applicable.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The access links used in training are mentioned in Section 3. In addition, the test dataset collected is available via https://t.ly/1AWwu.

References

[1] Ali Akkaya, Gokhan Bakal, A computational drug repositioning effort using patients' reviews dataset, in: 2023 International Conference on Smart Applications, Communications and Networking, SmartNets, IEEE, 2023, pp. 1–6.
[2] Betul Erkantarci, Gokhan Bakal, An empirical study of sentiment analysis utilizing machine learning and deep learning algorithms, J. Comput. Soc. Sci. (2023) 1–17.
[3] Walaa Medhat, Ahmed Hassan, Hoda Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Eng. J. 5 (4) (2014) 1093–1113.
[4] Abdullah Hussein Alamoodi, Bilal Bahaa Zaidan, Aws Alaa Zaidan, Osamah Shihab Albahri, Khalid Ibrahim Mohammed, Rami Qays Malik, Esam Motashar Almahdi, Mohammed A. Chyad, Ziadoon Tareq, Ahmed Shihab Albahri, et al., Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review, Expert Syst. Appl. 167 (2021) 114155.
[5] Hao-Chiang Koong Lin, Tao-Hua Wang, Guo-Chung Lin, Shu-Chen Cheng, Hong-Ren Chen, Yueh-Min Huang, Applying sentiment analysis to automatically classify consumer comments concerning marketing 4Cs aspects, Appl. Soft Comput. 97 (2020) 106755.
[6] Yilin Wang, Baoxin Li, Sentiment analysis for social media images, in: 2015 IEEE International Conference on Data Mining Workshop, ICDMW, IEEE, 2015, pp. 1584–1591.
[7] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, Amir Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion 91 (2023) 424–444.
[8] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al., Recent advances in convolutional neural networks, Pattern Recognit. 77 (2018) 354–377.
[9] Razvan Pascanu, Tomas Mikolov, Yoshua Bengio, On the difficulty of training recurrent neural networks, in: International Conference on Machine Learning, Pmlr, 2013, pp. 1310–1318.
[10] Gokhan Bakal, Orhan Abar, On comparative classification of relevant COVID-19 tweets, in: 2021 6th International Conference on Computer Science and Engineering, UBMK, IEEE, 2021, pp. 287–291.
[11] Lifang Wu, Mingchao Qi, Meng Jian, Heng Zhang, Visual sentiment analysis by combining global and local information, Neural Process. Lett. 51 (2020) 2063–2075.
[12] Bo Wang, Maria Liakata, Arkaitz Zubiaga, Rob Procter, Eric Jensen, Smile: Twitter emotion classification using domain adaptation, in: CEUR Workshop Proceedings, vol. 1619, Sun SITE Central Europe, 2016, pp. 15–21.
[13] Gil Levi, Tal Hassner, Emotion recognition in the wild via convolutional neural networks and mapped binary patterns, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 2015, pp. 503–510.
[14] Li Deng, Dong Yu, et al., Deep learning: methods and applications, Found. Trends® Signal Process. 7 (3–4) (2014) 197–387.
[15] Bo Li, Jiansheng Zhu, Linlin Dai, Hui Jing, Zhizheng Huang, The impact of introducing textual semantics on item instance retrieval with highly similar appearance: An empirical study, Image Vis. Comput. (2024) 104925.
[16] Xin Pan, Hao Zhai, You Yang, Lianhua Chen, Anyu Li, Improving multi-focus image fusion through noisy image and feature difference network, Image Vis. Comput. 142 (2024) 104891.
[17] Stuart J. Miller, Justin Howard, Paul Adams, Mel Schwan, Robert Slater, Multimodal classification using images and text, SMU Data Sci. Rev. 3 (3) (2020) 6.
[18] Quanzeng You, Liangliang Cao, Hailin Jin, Jiebo Luo, Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp. 1008–1017.
[19] Feiran Huang, Xiaoming Zhang, Zhonghua Zhao, Jie Xu, Zhoujun Li, Image–text sentiment analysis via deep multimodal attentive fusion, Knowl.-Based Syst. 167 (2019) 26–37.
[20] Verónica Pérez Rosas, Rada Mihalcea, Louis-Philippe Morency, Multimodal sentiment analysis of Spanish online videos, IEEE Intell. Syst. 28 (3) (2013) 38–45.
[21] Cecilia Ovesdotter Alm, Dan Roth, Richard Sproat, Emotions from text: machine learning for text-based emotion prediction, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005, pp. 579–586.
[22] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, 2016, arXiv preprint arXiv:1603.04467.
[23] Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, MIT Press, 2016.
[24] Yann LeCun, Yoshua Bengio, Geoffrey Hinton, Deep learning, Nat. 521 (7553) (2015) 436–444.
[25] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, Armand Joulin, Advances in pre-training distributed word representations, 2017, arXiv preprint arXiv:1712.09405.
[26] Berat Bozkurt, Kerem Coskun, Gokhan Bakal, Building a challenging medical dataset for comparative evaluation of classifier capabilities, Comput. Biol. Med. 178 (2024) 108721.
[27] Anil K. Jain, Jianchang Mao, K. Moidin Mohiuddin, Artificial neural networks: A tutorial, Computer 29 (3) (1996) 31–44.