International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Multimodal Sentiment and Emotion Classification Using BiLSTM Model with Bird Intelligence and Bald Eagle Optimization

Prashant K. Adakane[1], Dr. Amit K. Gaikwad[2]

Submitted: 28/01/2024 **Revised**: 06/03/2024 **Accepted**: 14/03/2024

Abstract: In the current digital era, when social media, customer reviews, and other online platforms generate massive amounts of text data every day, the ability to automatically identify and classify attitudes and emotions has become more and more important. Models of sentiment and emotion classification are vital tools for studying human emotional expression. Nevertheless, these models face many challenges, such as the need for effective multimodal data integration, handling of data imbalance, and the ability to capture nuanced emotional aspects. With the use of a white headed bird (WHB-) optimization and a bidirectional long short-term memory (BiLSTM) classifier, this work offers a strong approach for multimodal emotion classification. Furthermore, in order to improve overall efficiency, the paper presents a unique hybrid optimization technique that optimizes weights and biases in classifier parameters. The scientific novelty is in the combination of the cutting-edge WHB-optimization approach and the highly effective BiLSTM model, which together push the boundaries of multimodal emotional classification. By doing so, it makes a unique and beneficial addition to the field in the process. The created model achieves 94.90%, 95.27%, 94.79% and 95.78% accuracy performance for image, text, audio, and multimodal data, respectively.

Keywords: *BiLSTM, Multimodal, White Headed Bird Optimization*

1. Introduction

Sentiment and emotion classification is a subfield within natural language processing and machine learning that focuses on recognizing and categorizing human emotions and sentiments conveyed in text data. In today's digital age, the capacity to automatically assess and categorize emotions and attitudes has become increasingly crucial [1,3]. In the wide areas of human communication, emotions and feelings lie at the center of our utterances [4]. Sentiment and emotion categorization, often known as sentiment analysis or emotion recognition, is the technologically driven quest to decode this emotional materials hidden in text [5, 6]. The field of sentiment and emotion categorization aims to educate machines to recognize and interpret subjective parts of human communication, such as positive, negative, neutral, joyful, sad, and furious, and provides valuable insights about public opinion, consumer feedback, and societal trends [7, 8].

Sentiment analysis has applications across numerous fields. In marketing, sentiment analysis may help organizations assess client happiness and change their strategy accordingl y [9, 11]. These systems have developed to manage the complexity of human language, such as sarcasm, context, and cultural differences. The area continues to improve, spurred by the increasing quantity of text data accessible and the need for mining valuable insights from it [12-14]. This field has practical applications across various fields, such as marketing, social media monitoring, and customer support. Machine learning models in this field examine text input by preprocessing and translating it into numerical characteristics, and are trained on labeled data to learn patterns and correlations between text and sentiment or emotion labels [15-18]. Once trained, they can predict sentiment or emotion labels for fresh text inputs, making them effective for analyzing human emotions and views in diverse applications such as social media analysis and customer feedback analysis [19]. However, sentiment and emotion categorization faces several challenges due to the complexity of human language and emotional expression [20]. Intrinsic ambiguities, context-dependence, cultural variations, imbalance in sentiment category distribution within datasets, fine-grained categorization, and ethical issues regarding privacy and data usage contribute to the complexity [21-23].

The research's unique method for multimodal emotion and sentiment classification uses a BiLSTM classifier to efficiently assess input from multiple sources, including visual, linguistic, and auditory components. The combination of embeddings-based dimensionality reduction, feature extraction, and a BiLSTM network improves classification accuracy and context capture. This hybrid optimization method advances the field of

[1] *Research Scholar, Computer Science and Engineering Department,
G H Raisoni University, Amravati (Maharashtra-India),
ORCID ID : 0009-0002-5575-4611*
[2] *Associate Professor, Computer Science and Engineering Department,
G H Raisoni University, Amravati (Maharashtra-India),
ORCID ID : 0000-0002-0760-2165*
* *Corresponding Author Email: prashant.adakane@ghru.edu.in*

multimodal analysis by providing a thorough solution for precise and effective mood and sentiment categorization in various data modalities.

White-headed bird optimization combines bird swarm intelligence with the Bald Eagle's unique traits, balancing exploitation and exhaustive investigation. The algorithm uses the swarm's agility for local searches and the eagle's vision for global exploration. Their cooperative synergy allows them to make decisions based on memorized experiences and encourage convergence towards global information. This method encourages group exploration and utilization of solution spaces, demonstrating the potential of combining computational optimization approaches with biological concepts. WHB based BiLSTM for sentiment and emotion classification offers precise classification by using finely tuned features and the WHB method to find suitable classifier parameters.

2. Literature review:

Mahesh G. Huddar et al. [1] increase the efficacy of multimodal fusion in detection by providing a unique attention-based approach targeted at gathering contextual information among utterances. The model's performance was observed to be better than state-of-the-art methods for text-audio and text-video combinations in sentiment analysis and emotion detection.

Sarah A. Abdu et al. [2] A long detailed survey was presented by the authors in which presented a comprehensive overview of recent deep learning models and algorithms in multimodal sentiment analysis, categorizing thirty-five state-of-the-art models into eight categories. Different architectures like Majority voting, Hidden Markov Models, and Early Fusion showed varying levels of performance on the datasets, with Recurrent Memory Fusion Network demonstrating improved performance in modeling cross-modal interactions.

Maria Teresa Garcia-Ordas et al. [3] this paper proposes a sentiment analysis technique that is not a priori fixed and can handle audio of any duration. A Fully Convolutional Neural Network architecture is proposed as a classifier, enabling the prompt analysis of sentiment in many domains such as contact centers, medical consultations, and financial brokers. Mel spectrogram and Mel Frequency Cepstral Coefficients are used as audio description methods.

Dong Zhang et al. [4] lessen the time-consuming and labor-intensive annotation effort needed for multi-modal sentiment classification. This model maintained or even improved classification performance while significantly reducing the annotation burden, but it required a sizable amount of unlabeled data in order to mine useful information.

Harnain Kour and Manoj K. Gupta [5] commenced on the attempt of user's mental state prediction by differentiating between those with depressed and non-depressive inclinations, utilizing data gathered from Twitter. Their technique exhibited amazing accuracy when applied to a benchmark depression dataset. The model's possible shortcoming is its dependence on textual Twitter data, possibly omitting individuals who convey their mental condition via non-textual information like photographs or videos on social media.

Hu Zhu et al. [6] introduced a low-rank tensor multimodal fusion technique that lowers computing complexity and increases efficiency. three multimodal fusion tasks—CMU-MOSI, IEMOCAP, and POM—that are based on a public data set were used to evaluate the model. The model that is being presented performs well and captures both local and global linkages with flexibility. Experiments indicate that the model can consistently obtain better outcomes under a range of attention processes when compared to previous multimodal fusions represented by tensors.

Xiaocui Yang et al. [7] attempted to enhance multimodal sentiment analysis by constructing a large-scale emotion dataset and presenting a new multimodal emotion analysis model (MVAN) that incorporates correlations across modalities and addresses particular emotions stated by users. The possible downside might be the computational complexity of the MVAN model owing to its multi-stage process and memory network utilization, which may need large resources.

Ruo-Hong Huan et al. [8] intended to increase video multimodal emotion identification by using gated unit to improve temporal context learning. The considerable benefit of this methodology is obvious in the heightened accuracy gained in emotion recognition for both single modalities and video multimodal emotion recognition, exceeding the previous approaches. However, it is worth mentioning a possible downside the accompanying computational complexity, which might need large computer resources for real-time processing.

Peng Wu et al. [9] the study presents a bidirectional long short-term memory network (SC-ABiLSTM) and attention mechanism-based sentiment categorization technique for large-scale microblog material. The effectiveness of the suggested method is shown by comparing it with baseline techniques utilizing large-scale microblog data from real-world applications. The research centers on the innovative application of the attention mechanism in a deep learning network for the purpose of examining extensive social media data. The difficulties of sentiment classification in the setting of microblog content and the significance of effective feature extraction are also covered in the study.

Tian Chen et. al [10] created a multimodal fusion emotion identification algorithm combining EEG and ECG data. The noticeable benefit resides in the higher performance displayed by the suggested multimodal fusion model as compared to single-modal models. This leads in better accuracy in both the Arousal and Valance dimensions. A possible downside might be the necessity for specific equipment (EEG and ECG) for data collection, which may restrict the applicability of the procedure in particular circumstances.

The primary research issues include:

Insufficient or noisy training data might lead to poor model accuracy. Emotion classification generally needs labeled emotional data, which may be challenging to gather in big amounts [4].

Selecting acceptable elements from diverse modalities (text, audio, visual, etc.) and combining them successfully is hard and may need subject knowledge [4].

Imbalanced class distribution, when particular emotions are underrepresented in training data, may lead to biased models [5].

Integrating data from many modalities while keeping its relevance and limiting information loss is a problem [8].

Handling sensitive emotional data creates privacy and ethical considerations, particularly in applications using personal user data [10].

3. Proposed Methodology:

This study uses a BiLSTM classifier to categorize multimodal emotions from two datasets [29] and [30]. The inputs include visual, linguistic, and auditory components. The data undergoes preprocessing to improve image quality, and feature extraction is performed using a unimodal learning system, where the facial characteristics, including eyes, nose, and mouth are obtained from the visual images, acoustic data yields characteristics related to clarity and intelligibility, and the language features are extracted using BERT. Visual, verbal, and audio embeddings are used to reduce feature dimensionality. The BiLSTM network is integrated via joint embedding, resulting in decreased processing time and efficient sequential data categorization. The model also incorporates two sub-networks for forward and backward sequence processing, increasing its ability to collect contextual information. The key novelty is a hybrid optimization using standard features, which improves classifier efficiency by improving weights and biases.

3.1. Input:

Two separate datasets are used as a multimodal dataset, specifically the [29] and [30] datasets, which are used for sentiment analysis and classification. M represents multimodal vector whereas M_I represents image Unimodal vector, M_T represents text Unimodal vector, M_S represents sound or audio Unimodal vector,

$$M = \{[M_I],[M_T],[M_S]\} \tag{1}$$

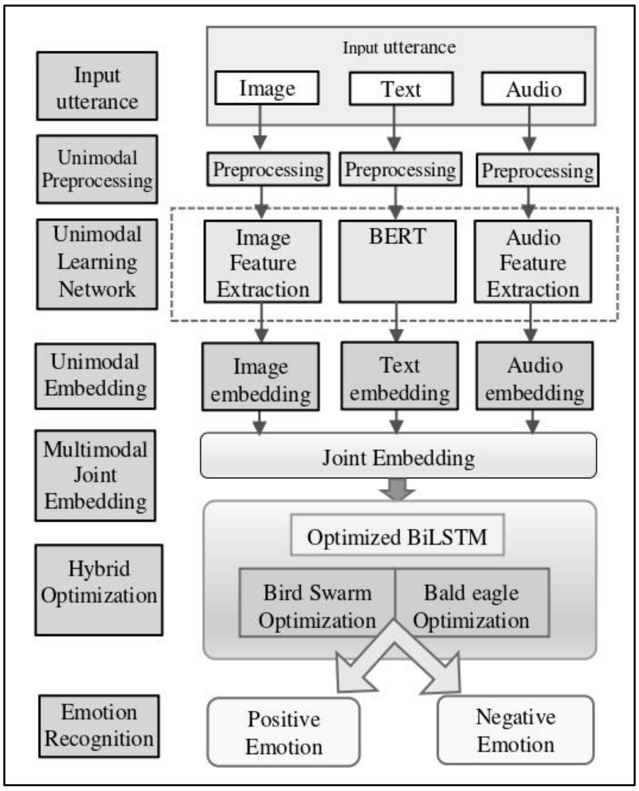

Fig. 1. Proposed model for sentiment and emotion classification

3.2. Preprocessing:

Denoising mechanisms improve the quality of input from feature extraction and classification by preprocessing image, text, and acoustic data, thereby enhancing multimodal emotion and sentiment classification accuracy.

3.2.1. Image Preprocessing:

Image preprocessing in facial analysis aims to remove noise and undesired artifacts, which can affect the accuracy of future analysis. Ambiguity can lead to mistakes in the analysis process. Denoising algorithms are used to address this issue. This study uses a sophisticated denoising strategy to limit noise effects on face photographs, detecting and reducing undesired components while preserving the inherent structure of the data. This approach aims to improve future face analysis accuracy and yield more reliable conclusions. The output that has undergone visual preprocessing is shown as M_I^*.

3.2.2. Text Preprocessing:

Text cleaning enhances the quality of language data by removing irrelevant or distracting terms, making it more receptive to feature extraction. The linguistic pre-processed output is represented as M_T^*

3.2.3. Sound Preprocessing:

Acoustic data refers to audio recordings containing speech or other auditory stimuli. It begins with noise reduction, which eliminates background noises, interference, or recording defects. Clean audio data is crucial for effective feature extraction, as it ensures the retrieved acoustic characteristics accurately reflect real emotional emotions. The pre-processed data is represented as M_S^* .

3.3. Unimodal Learning network:

The unimodal learning network is a crucial element in embedding multimodal information, facilitating the extraction of data from visual, linguistic, and auditory components. It plays a crucial role in feature extraction, recording modality-specific information for discriminative and meaningful representations. The study aims to improve the model's capacity to identify and use important characteristics from multiple sources, laying the groundwork for more efficient multimodal activities. The unimodal learning network is essential for capturing the significance of each modality in a meaningful way.

3.4. Visual Embedding:

Visual embedding is a transformational process that extracts visual qualities from pictures to capture information, making it crucial for converting unprocessed visual input into a concise, useful representation for multimodal analysis.

3.4.1. Facial feature extraction using Viola Jones and VGG-16 model:

The study uses a combination of the VGG-16 model and Viola Jones to extract face features, a crucial aspect of computer vision applications. The Viola-Jones approach, which uses a cascade of classifiers to identify areas of interest in pictures, is effective in face identification due to its integration of Haar-like features. Important facial features like the mouth, nose, and eyes are isolated using a face masking approach. The VGG-16 model, a Convolutional neural network, extracts high-level features and builds embeddings to improve face analysis depth. This method ensures a comprehensive understanding of face characteristics.

3.4.2. Embedding using VGG-16:

Visual embedding is the process of converting extracted visual characteristics, such as face landmarks, into a common representation space using Visual Geometry Group (VGG). VGG, or CNN architecture, is used to extract relevant characteristics from pictures by converting them into abstract visual representations that encapsulate crucial information. These representations represent the visual data in an abstracted and standardized manner. The VGG-16 network has a small receptive field size of [3×3] and is made up of 16 convolutional layers. A total of five integrated Max pooling layers, each with a size of [2×2], are included in the design. The network design consists of three fully connected layers, a soft max classifier, and ReLU activation for all hidden layers. RELU introduces non-linearity, allowing the model to learn complex representations from retrieved characteristics. This architecture ensures effective information propagation. The output layer is the final soft max classifier, which categorizes input data using learned features and offers probability distributions across anticipated classes. The forward pass through the network layers converts a vector (v) into a numerical representation, representing VGG-16 mathematically.

$$M_I^{\#} = VGG - 16(v) \tag{2}$$

Here $M_I^{\#}$ symbolize the extracted features, while VGG signifies the output of the $VGG-16(v)$ network when the vector v undergoes processing. Collectively, these operations learn and encapsulate crucial image features.

3.5. Language embedding:

The BERT model is utilized for language embedding, converting natural language text into semantic vectors and utilizing contextual awareness to interpret ambiguous language and capture subtle emotional content. This enhances analysis and provides a more comprehensive understanding of emotions within the BiLSTM classifier architecture.

3.5.1. Feature extraction and embedding using BERT:

The study suggests that the constant size of pre-trained vectors in natural language processing (NLP) can hinder the learning of token semantic properties in diverse settings. To address this, the study recommends using the pre-trained BERT model, which understands the complex meaning of text by considering the context in which words, phrases, or sentences occur. BERT is trained using multi-layer bidirectional transformer encoders to create character-level context over grammar characteristics within a given sequence. The three main parameters of the BERT model are the number of layers in the transformer block, concealed size, and the number of self-attention heads in a transformer block. The model uses two phrases as training examples to capture the contextual and semantic properties of tokens. After tokenizing these phrases, the model creates fixed-length vectors by combining token, position, and segment embeddings. BERT feature extraction examines both the surrounding context and the words themselves, allowing the classifier to better understand the meaning and nuances of the language, which is crucial for tasks like mood and emotion categorization.

$$M_T^{\#} = BERT_{transformer}\left(t_{[H]}, t_1, t_2, ...t_n, t_{[S]}\right) \tag{3}$$

Where, H and S delineate the commencement and conclusion of sentences, respectively.

3.6. Sound or Audio embedding:

Acoustic embedding using Mel-Frequency Cepstral Coefficients (MFCC) is a popular method for encoding audio data. MFCCs extract temporal patterns and spectral characteristics from audio signals, aligning them with visual and linguistic embeddings. This ensures a coherent multimodal analysis, ensuring auditory information harmonizes with visual and textual data.

3.6.1. Audio feature extraction:

Acoustic preprocessed data is used for audio feature extraction to capture sound qualities, such as intelligibility and clarity. This process provides insights into the quality and comprehensibility of audio data. The retrieved acoustic elements provide valuable insights into the data's quality and comprehensibility

$$M_S^{\#} = \sum_{L=1}^{L} \log(Y[l]) \cos\left(\frac{\pi\, n\,(L-0.5)}{L}\right) \tag{4}$$

In this scenario, the length of each frame in MFCC is represented by $n = 1, \ldots.. N$. The output energy of the L^{th} band is denoted as $Y[l]$, and the resulting acoustic embedding output is represented as $M_S^{\#}$.

3.7. Joint embedding:

The vector $M_T^{\#}, M_I^{\#}, M_S^{\#}$ forms input for joint embedding is a

representation space that merges features from visual, verbal, and audio modalities into a cohesive structure.

3.8. Classifying the emotions from multimodal data using Optimized BiLSTM model:

The BiLSTM model is designed for sentiment classification and emotion detection using embedded multimodality characteristics from text, visual, and audio data. It differentiates between positive and negative emotions using learned qualities. The model uses bidirectional processing capabilities to capture temporal dynamics in emotions, enabling more complex and context-aware categorization. This method helps identify minute patterns in multimodal data, improving comprehension of emotional tone. The BiLSTM model includes both past and future context, enhancing its understanding of sequential data. It has two sub-networks: one for forward sequence processing and the other for backward sequence processing. Each sub-network functions autonomously, handling incoming data according to its own direction. The outputs from both sub-networks are concatenated to provide a single representation encoding bidirectional context.

3.9. Proposed White headed bird optimization:

The WHB Optimization refines a BiLSTM model's weights and biases by combining bird swarm optimization with Bald Eagle characteristics. This results in a highly adaptive strategy, with the Bald Eagle's flexibility directing global exploration and collective intelligence, while the bird swarm's agility and collective intelligence are used for effective local searches. The eagle's decision-making based on global information enhances the optimization algorithm [27] [28], encouraging cooperative exploration and exploitation of the solution space. The white-headed bird optimizer chooses the best hunting region to avoid energy waste, showcasing the cautious nature of bird swarms. The strongest hunter memorizes food availability, with weakest birds used as scapegoats.

3.9.1. Initialization:

The BiLSTM model's parameter values have been enhanced through random initialization, promoting exploration, avoiding bias, and enhancing optimization dependability.

Let's write the answer as $P \in [w\_i, b\_i]$.

$$P_1^{t+1} = P_1^t + \left(P_{g1rand} * R_1 - R_2 * P_{g1rand}\right) \tag{5}$$

3.9.2. Evaluating the solution fitness:

When evaluating a solution, its objective function for every solution when $(t < t_{\max})$ is taken into consideration. The BiLSTM model that offers the greatest classification accuracy is the best candidate for tweaking.

$$fit[F(t)] = \max(Accuracy) \tag{6}$$

3.9.3. Search space selection:

The optimization approach, inspired by bird behavior, uses a dynamic and adaptive search space selection to determine the optimal location for local and global search efforts. This approach guides the algorithm's route, allowing it to explore and exploit areas in the solution space to identify the best solutions for the BiLSTM model,

$$s^{t+1} = s^*(M) + a_1 * r_1 * \left(s_{mean} - s^t\right) + a_2 * r_2 * \left(s_g - s_t\right) \tag{7}$$

$$a_1 = \frac{a_{\max} - a_{\min}}{\sum a} \tag{8}$$

$$a_2 = \frac{fit_{\max}\left(s^{t+1}\right)}{s_{mean}} \tag{9}$$

Here, $s^*(M)$ denotes the best search space in lifetime (memorized), a_1 *and* a_2 denotes the selection vector, s_g denotes the global best solution in the previous round, $r1$ *and* r_2 denotes the random number [0,1]. s_{mean} denotes the overall search space, s^t denotes the search space in current iteration, $rand$ denotes the random number, P_{per} represents the best previous solution, P_g denotes the global solution in previous iteration and P_g^{t-1} denotes the global solution in $t-1$ iteration. The algorithm's significance lies in selecting a search space based on available food to ensure faster global convergence, and then searching for food within this chosen space.

3.9.4. Search phase:

The bird initiates a random search in a designated space when the search random number is less than 0.5, resembling how birds explore their environment, exploring new places or returning to familiar ones. The model for the solution/birds in search space is,

$$P^{t+1} = P^t + y^t * \left(P^{t-1} - P^t\right) + r_3 P^t + \left(S^t - S^{mean}\right) + P^{t-2}\left(f * r_5\right) \tag{10}$$

$$y^{t+1} = \frac{y^{t-1} * r_3}{\max\left(y^t\right)} \tag{11}$$

Here, $r3$ denotes the random number from [0,1], f denotes the flight length, y^{t-1} represent the distance of i^{th} bird from food in previous iteration and y^t represent the i^{th} bird from food community. Since the search space is selected, the experience of the bird in the particular search space during the previous searches P^{t-1} & P^{t-2} are recalled for the decision over the position.

$$r_3 = A.a_4 + r_4 \tag{12}$$

Where, A denotes the coefficient vector, and the constant a_4 is modeled as,

$$a_4 = \max\ fit\left(x^t\right) \tag{13}$$

3.9.5. Global phase:

The algorithm uses global data and shared insights throughout the crucial global phase. It emphasizes exploration and the search for new, better answers while using the best solutions from the past to direct group behavior.

$$\begin{aligned} P^{t+1} = P^t &+ \left(P_{per} - P^t\right) \times c_1 * rand(0,1) + \left(P_g - P_t\right) \\ &\times c_2 * rand(0,1) + rand(0,1) * P_g^{t-1} \\ &+ \alpha . P_g^{t-2} + \frac{\alpha}{2}(1-\alpha) P_g^{\ t-3} \\ &+ \frac{1}{6}\alpha(1-\alpha)\,(2-\alpha) P_g^{t-4} \end{aligned} \tag{14}$$

3.9.6. Local phase:

The local optimization phase optimizes solutions in promising locations by adjusting them based on proximity to potential optima. It considers both worst experiences and individual's best locations for position updates to prevent search performance convergence.

$$\begin{aligned} P^{t+1} = P^t &+ \left(P_{worst} - P^t\right) rand(0,1) \\ &+ c_3 * rand(0,1) * \left(P^t - P_{per}\right) \end{aligned} \tag{15}$$

3.9.7. Attack and clear phase:

The bird's vigilant conduct in the assault and clear phase involves improving solutions close to possible optima while observing their alignment with intended parameters. Adaptive techniques are used to improve adjacent solutions or eliminate those below predetermined fitness criteria, enhancing the robustness and effectiveness of the optimization process and advancing the search for ideal solutions in the BiLSTM model optimization.

$$\begin{aligned} P^{t+1} = P^t &+ B_1 . \left(S_{mean} - P^t\right) \times rand(0,1) \\ &+ B_2 \left(P_{per} - P^t\right) \times rand(-1,1) \end{aligned} \tag{16}$$

$$B_1 = b_1 \times \exp\left(- \frac{fit\left(P_{per}\right)_{\max}}{\sum fit\left(P^t + \eta\right)} \times N_{rand} \right) \tag{17}$$

$$B_2 = \left(\frac{F_{\max} - F_{\min}}{F_{\max}} \right) \tag{18}$$

Here, η denotes the zero division error, N_{rand} is from the range

$[-1,1]$ and $\sum fit\left(P^t + \eta\right)$ denotes the summation of fitness of P_t

3.9.8. Termination:

The optimization procedure checks in this termination situation. The global solution is stated and the model's fitness is re-evaluated if this criterion is satisfied $if\left(t > t_{\max}\right)$.

4. Experimental setup:

The experiment on sentiment and emotion classification is conducted using Python on a Windows 10 system with an 8GB internal memory capacity.

Datasets Used:

Multi Modal Emotion Recognition [29]: The dataset enhances emotion recognition accuracy by analyzing visual and speech cues in human interactions, focusing on facial expressions and speech patterns for applications like human-computer interaction and sentiment analysis.

Multi Modal Sentiment Classification [30]: The dataset enhances sentiment classification accuracy by analyzing visual and speech data, focusing on crucial features, making it useful for emotion-aware technology and content analysis.

5. Result and discussion:

The WHB-BiLSTM model's efficacy in sentiment and emotion classification is systematically evaluated against various alternative methods to assess its performance and capabilities. In a comparative evaluation, the WHB-BiLSTM model effectiveness is assessed in comparison to other models, Deep CNN, BiLSTM and BSO-BiLSTM.

5.1. Comparative discussion:

5.1.1. Accuracy, Sensitivity & Specificity comparative analysis for image:

According to data presented in figure 2, the WHB-BiLSTM model outperforms other methods in sentiment and emotion classification. It surpasses the BSO-BiLSTM model by 5.14%, with an impressive 94.09% accuracy rate on a 90% training set. The proposed model shows 4.14% enhancement in sensitivity with 94.51% and 5.07% enhancement in specificity with 94.66% over the BSO-BiLSTM methodology.

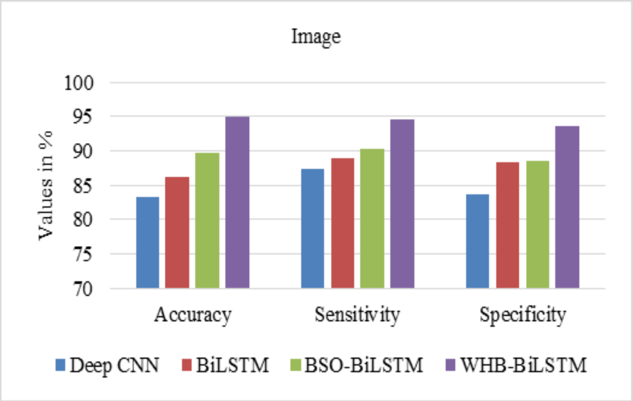

Fig. 2. Comparative analysis based on accuracy, sensitivity and specificity for image when TP = 90

5.1.2. Accuracy, Sensitivity & Specificity comparative analysis for text:

According to data presented in figure 3, the WHB-BiLSTM model achieved an accuracy of 95.27% in classifying sentiment and emotion, outperforming the BSO-BiLSTM model by 4.07%. It's sensitivity rate is 95.23%, a 4.53% enhancement compared to the BSO-BiLSTM method, and it's specificity exhibits 4.43% improvement reaching 95.15%. The model is trained using a 90% training percentage.

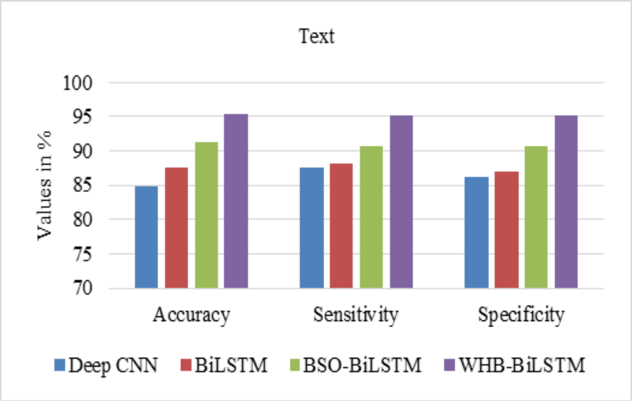

Fig. 3. Comparative analysis based on accuracy, sensitivity and specificity for text when TP = 90

5.1.3. Accuracy, Sensitivity & Specificity comparative analysis for audio:

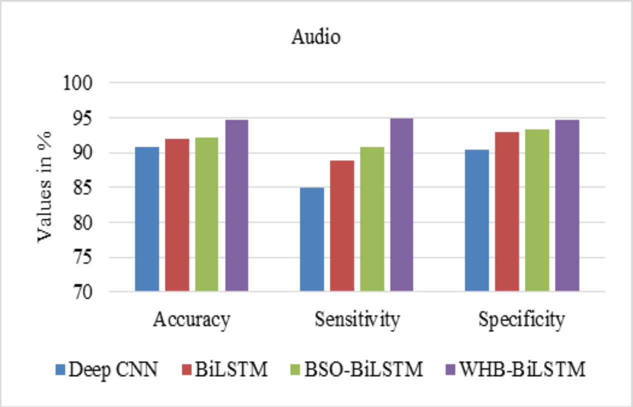

Fig. 4. Comparative analysis based on accuracy, sensitivity and specificity for audio when TP = 90

According to data presented in figure 4, the WHB-BiLSTM model demonstrates impressive accuracy of 94.79% with a training percentage of 90 in classifying sentiment and emotion. Notably, it surpasses the BSO-BiLSTM model by 2.62%. It outperforms with a sensitivity rate of 94.89%, representing a notable 4.04% increase. With 94.61%, the WHB-BiLSTM model exhibits a 1.29% improvement in specificity.

5.1.4. Accuracy, Sensitivity & Specificity comparative analysis for multimodal:

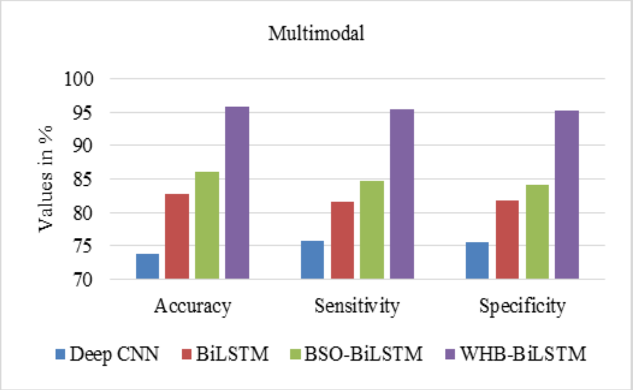

Fig. 5. Comparative analysis based on accuracy, sensitivity and specificity for multimodal when TP = 90

According to data presented in figure 5, the WHB-BiLSTM model demonstrates exceptional accuracy in classifying sentiment and emotion, achieving an impressive accuracy of 95.78% with a training percentage of 90. Notably, it surpasses the BSO-BiLSTM model by a substantial margin, outperforming it by 9.66%.

With a sensitivity of 95.04%, the WHB-BiLSTM model performs very well and offers a notable 10.67% increase.

WHB-BiLSTM model exhibits substantial 11.13% improvement in specificity with 95.3% when contrasted with the BSO-BiLSTM approach. These results highlight the exceptional performance of the WHB-BiLSTM model in this crucial field.

According to data presented in table 1, the 90% evaluation of TP across various data modalities demonstrates its comprehensive nature. The WHB-BiLSTM models have made significant advancements due to their ability to capture sequential dependencies and contextual information in multimodal data, considering both past and future context, enabling a more nuanced understanding of temporal patterns.

The incorporation of multimodal inputs further enhances the model's ability to discern complex relationships between different types of data, contributing to superior performance compared to the pre-existing model.

This rigorous comparative analysis underscores the efficacy of WHB-BiLSTM models in advancing the state-of-the-art in multimodal emotion classification. The achievement of the developed model using the image modality is 94.90%, text is 95.27%, audio is 94.79%, and multimodal data is 95.78% for accuracy.

Table 1. Comparative discussion table based on TP = 90

	Model	*Deep CNN*	*BiLSTM*	*BSO-BiLSTM*	*WHB-BiLSTM*
Image	Accuracy	83.25	86.27	89.76	94.90
	Sensitivity	87.34	88.93	90.37	94.51
	Specificity	83.78	88.42	88.54	93.61
Text	Accuracy	84.78	87.63	91.2	95.27
	Sensitivity	87.61	88.18	90.7	95.23
	Specificity	86.19	86.91	90.72	95.15
Audio	Accuracy	90.86	92.02	92.17	94.79
	Sensitivity	84.88	88.92	90.85	94.89
	Specificity	90.49	92.89	93.32	94.61
Multimodal	Accuracy	73.87	82.73	90.12	95.78
	Sensitivity	75.85	81.68	84.73	95.4
	Specificity	75.65	81.78	84.17	95.30

6. Conclusion:

In conclusion, this research presents a robust methodology for multimodal emotion classification, primarily driven by the utilization of a BiLSTM classifier and a WHB-optimization. Leveraging data from well-established datasets encompassing visual, language, and acoustic components, the research employs a systematic approach involving preprocessing, feature extraction, and dimensionality reduction. Notably, the research innovation lies in the introduction of a white headed optimization method that combines the bird swarm intelligence with the distinctive characteristics of a bald eagle which fine-tunes the classifier parameters effectively. This hybrid approach significantly enhances classifier efficiency by optimizing weights and biases. In essence, the contribution of this research lies in the integration of the powerful BiLSTM model and the novel hybrid optimization technique, collectively advancing the state-of-the-art in multimodal emotion classification. The seamless integration of these components not only showcases the effectiveness of the proposed methodology but also sets the stage for future advancements in the field of sentiment analysis model. The achievement of the developed model using the image modality is 94.90%, text is 95.27%, audio is 94.79%, and multimodal data is 95.78% for accuracy. Future work could explore the adaptation of the proposed BiLSTM and hybrid optimization methodology to real-world scenarios, assess its generalizability across diverse datasets, and investigate the potential integration of emerging technologies, such as attention mechanisms, for further enhancing multimodal emotion classification performance.

Author Contributions

Prashant K. Adakane: Conceptualization, Methodology, Field study, Software, Data curation Validation, Writing-Original draft reviewing and editing.

Amit K. Gaikwad: Field study, Data curation, Visualization.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Huddar, Mahesh G., Sanjeev S. Sannakki, and Vijay S. Rajpurohit, “Attention-based multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM,” Multimedia Tools and Applications, vol. 80, no. 9, pp.13059-13076, 2021.

[2] S. A. Abdu, A. H. Yousef, and A. Salem, “Multimodal video sentiment analysis using deep learning approaches, a survey,” Information Fusion (Print), vol. 76, pp. 204–226, Dec. 2021, doi: 10.1016/j.inffus.2021.06.003.

[3] Maria Teresa Garcia-Ordas, Hector Alaiz-Moreton, Jose Alberto Benítez-Andrades, Isaias Garcia-Rodríguez, Oscar Garcia-Olalla, Carmen Benavides, “Sentiment analysis in non-fixed length audios using a Fully Convolutional Neural Network,” Biomedical

Signal Processing and Control, Volume 69, p. 102946, Aug. 2021, doi: 10.1016/j.bspc.2021.102946.

[4] Zhang, Dong, Shoushan Li,Qiaoming Zhu, and Guodong Zhou, "Multi-modal sentiment classification with independent and interactive knowledge via semi-supervised learning," IEEE Access, vol. 8, pp. 22945-22954, 2020. doi: 10.1109/ACCESS.2020.2969205.

[5] H. Kour and M. K. Gupta, "An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM," Multimedia Tools and Applications, vol. 81, no. 17, pp. 23649–23685, Mar. 2022, doi: 10.1007/s11042-022-12648-y.

[6] Hu Zhu, Ze Wang, Yu Shi, Yingying Hua, Guoxia Xu, and Lizhen Deng, "Multimodal fusion method based on self-attention mechanism,|" Wireless Communications and Mobile Computing 2020 (2020), 1–8. doi:10.1155/2020/8843186

[7] Yang, Xiaocui, Shi Feng, Daling Wang, and Yifei Zhang, "Image-text multimodal emotion classification via multi-view attentional network," IEEE Transactions on Multimedia, vol. 23 pp.4014-4026, 2020. doi:10.1109/TMM.2020.3035277

[8] Huan Ruo-Hong, Jia Shu, Sheng-Lin Bao, Rong-Hua Liang, Peng Chen, and Kai-Kai Chi, "Video multimodal emotion recognition based on Bi-GRU and attention fusion," Multimedia Tools and Applications, vol. 80, no. 6, pp. 8213-8240, 2021. doi: 10.1007/s11042-020-10030-4

[9] Peng Wu, Xiaotong Li, Chen Ling, Shengchun Ding, Si Shen, "Sentiment classification using attention mechanism and bidirectional long short-term memory network," Applied Soft Computing, Volume 112, pp. 107792, 2021. doi:10.1016/j.asoc.2021.107792.

[10] Chen, Tian, Hongfang Yin, Xiaohui Yuan, Yu Gu, Fuji Ren, and Xiao Sun, "Emotion recognition based on fusion of long short-term memory networks and SVMs," Digital Signal Processing, vol. 117, pp.103153, 2021.

[11] Mai, Sijie, Haifeng Hu, Jia Xu, and Songlong Xing, "Multi-fusion residual memory network for multimodal human sentiment comprehension," IEEE Transactions on Affective Computing, 2020.

[12] Boateng, George, and Tobias Kowatsch, "Speech emotion recognition among elderly individuals using multimodal fusion and transfer learning," In Companion Publication of the 2020 International Conference on Multimodal Interaction, pp. 12-16, 2020.

[13] Lahat, Dana, TülayAdali, and Christian Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," Proceedings of the IEEE 103, no. 9, pp. 1449-1477, 2015.

[14] Celli F, Lepri B, Biel J-I, Gatica-Perez D, Riccardi G, Pianesi F, The workshop on computational personality recognition, In: Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, pp 1245–1246, 2014.

[15] Poria S, Cambria E, Bajpai R, Hussain A, "A review of affective computing: from unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp.:98–125, 2017.

[16] [Morency LP, Mihalcea R, Doshi P, "Towards multimodal sentiment analysis: harvesting opinions from the web," proceedings of the 13th international conference on multimodal interfaces, ICMI 2011, Alicante, Spain, pp. 14-18, 2011.

A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, ''Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages,'' IEEE Intell. Syst., vol. 31, no. 6, pp. 82–88, Nov. 2016.

[17] Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, ''Tensor fusion network for multimodal sentiment analysis,'' in Proc. EMNLP, Copenhagen, Denmark, pp. 1103–1114, 2017.

[18] S. Poria, E. Cambria, and A. Gelbukh, ''Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis,'' in Proc. Conf. Empirical Methods Natural Lang. Process., Lisbon, Portugal, pp. 2539–2544, 2015.

A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L. P. Morency, ''Memory fusion network for multi-view sequential learning,'' in Proc. AAAI, New Orleans, LA, USA, vol. 8, pp. 5634–5641,2020.

[19] Leiva V, Freire A, "Towards suicide prevention: early detection of depression on social media," International Conference on Internet Science. Springer, Cham, pp 428–436, 2017.

[20] Stephen JJ, Prabu P, "Detecting the magnitude of depression in Twitter users using sentiment analysis," International Journal of Electrical and Computer Engineering, vol. 9, no. 4, pp. 3247, 2019.

[21] Bahdanau D, Cho K, Bengio Y, "Neural machine translation by jointly learning to align and translate", pp. 1-15, 2014.

[22] Busso C, Bulut M, Lee C, Kazemzadeh A, Mower E, Kim S, Chang J, Lee S, Narayanan S IEMOCAP: interactive emotional dyadic motion capture database. J Language ResourEvaluat, vol. 42, no. 4, pp. 335–359, 2008.

[23] Cambria E, “Affective computing and sentiment analysis,” IEEE Intell Syst, vol. 31, no.2, pp.102–107, 2016.

[24] Huddar, Mahesh G., Sanjeev S. Sannakki, and Vijay S. Rajpurohit, “Attention-based word-level contextual feature extraction and cross-modality fusion for sentiment analysis and emotion classification,” International Journal of Intelligent Engineering Informatics, vol. 8, no. 1, pp.1-18, 2020.

[25] Zhang, Yazhou, Prayag Tiwari, Lu Rong, Rui Chen, Nojoom A.AlNajem, and M. Shamim Hossain, “Affective Interaction: Attentive Representation Learning for Multi-Modal Sentiment Classification,” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2021.

[26] Huddar, Mahesh G., Sanjeev S. Sannakki, and Vijay S. Rajpurohit, “Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification,” International Journal of Multimedia Information Retrieval, vol. 9, no. 2, pp. 103-112, 2020.

[27] Multimodal emotion recognition dataset, MVSA – Single from https://www.kaggle.com/datasets/vincemarcs/mvsasingle

[28] [30] Multimodal sentiment classification dataset, MVSA – Multiple from https://www.kaggle.com/datasets/vincemarcs/mvsamultiple

[29] Mishra, Kaushik, and Santosh Kumar Majhi, “A binary bird swarm optimization based load balancing algorithm for cloud computing environment,” Open Computer Science, vol. 11, no. 1 pp. 146-160, 2021.

[30] Sayed, Gehad Ismail, Mona M. Soliman, and Aboul Ella Hassanien, “A novel melanoma prediction model for imbalanced data using optimized SqueezeNet by bald eagle search optimization,” Computers in Biology and Medicine, vol. 136, pp. 104712, 2021.