International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(14s) 2026 pp. 374-385
https://doi.org/10.70917/ijcisim-2026-4237
Article

Multimodality Deep Feature-Based Stress Classification Using Mita Algorithm

V. Roseline[1], M.Vijayalakshmi[2]

[1]PG Department of Computer Science, Sadakathullah Appa College (Autonomous),
Affiliated to Manonmaniam Sundaranar University, Rahmath Nagar, Tirunelveli – 627 011, Tamil Nadu, India.
[2]Department of Computer Science, Govt. Arts and Science College,
Affiliated to Manonmaniam Sundaranar University, Manur, Tirunelveli – 627 201, Tamil Nadu, India.
[1]rose.vasee2014@sadakath.ac.in,
[2]vksk2979@gmail.com

Abstract: Clients generate enormous volumes of data massively and dynamically these days, thanks to the growing variety of Web technologies. In this context, emotional analysis seemed to be a key technique for automating the extraction of understanding from user-generated material. So far, sentiment classification on data from social media has been limited to a single modality, such as text or picture. However, easily available multimodal information, such as images and other types of texts, as a group can assist in more precisely anticipating thoughts. Deep learning has recently demonstrated exceptional performance in the field of emotion and is regarded as the advanced approach. In this research, we present a Multimodal Image and Text with Attention (MITA) algorithm for classifying the sentiment of social media tweets by combining the Backpropagation Long Short-Term Memory with Conditional Random Field (BiLSTM-CRF) and Fusion awareness models for both text and image. Each character's context text is modeled by the BiLSTM layer. The BiLSTM layer's hidden states are handled at the CRF layer to improve sequential labeling with the help of neighboring labels. On the other hand, the image is addressed with a basic structure of Convolutional neural Networks (CNNs) including three perspectives such as spatial, a channel with squeezing, and excitation at various levels. Our suggested effort yielded encouraging results and enhanced sentiment classification accuracy.
Keywords: BiLSTM, Channel Attention, CRF, Pooling, Spatial attention, Squeeze, and Excitation.

1. INTRODUCTION

With the increased use of web-based media, emotional analysis has gained appeal among a diverse range of people with varying interests and motivations. Aside from the information on customers' frequented locations, purchasing habits, and so on, understanding their emotions as they are transmitted by their messages at various phases has become an essential part of assessing people's perspectives on a certain issue. An extremely common strategy is to identify the extreme of a paragraph in terms of client satisfaction, disappointment, or impartiality.

The sentimental analysis involves the study of breaking down users' feelings using standard language handling, message research, and insights. The finest organizations understand what their users are speaking about, how they're expressing it, as well as what they mean. User feedback can be discovered in tweets, comments, audits, and other places where people see our representation. Emotional Analysis is the field of comprehending these sentiments through computing, and it is an indisputable prerequisite for programmers and business professionals to grasp in today's working environment.

Similarly to various other domains, advances in transfer learning have moved emotive analysis closer to the forefront of cutting-edge computations. Today, we use normal language handling, measures, and message inspection to separate and discriminate between positive, negative, and neutral categories of words.

1.1 The motivation for the Research work

Emotions are difficult to quantify. Their magnitude may be measured slightly by seeing your actual reaction (brainwaves, for example), but as a general rule, it is difficult to distinguish one sensation from another. Considering the valence and excitement ratings, provided by Feldman in 1995, is one way of checking feelings. The amount or measure of physical reaction (force) is associated with excitement, but valence evaluates the emotional orientations (good or bad feelings). This should be obvious in Fig. 1, in which the two axes dealing with positivity and arousal are shown for specific sensations within it. Valence ratings, for example, may distinguish between angry (feeling gloomy) and satisfied (excellent). Arousal can distinguish between furious or left (serious sentiments) and loose or weary sensations (powerless feelings).

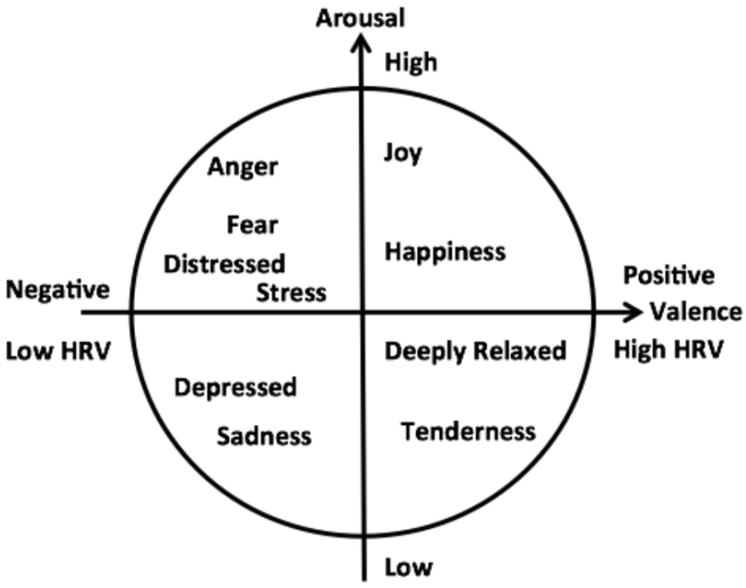

Fig. 1: Sentimental classification of Arousal and Valence

We concentrated on the polarity proportions of sensations in this work. The architecture that authorizes and categorizes the outcome utilizes three fine-grained opinion classifications, namely 'positive,' 'negative,' and 'neutral.' One of the approaches for emotional analysis on Twitter is machine learning, which is described below. Deep learning models have lately achieved tremendous results in Computer vision and speech recognition. To address NLP (Natural Language Processing) challenges, machine learning may be used by combining an effective learning calculation with a large sample of data to build expertise with the categorization rules. A few tactics use algorithms such as SVM or Naive Bayes, and a big share of such techniques evaluate communication word by word, grouping a phrase to either negative or positive by separating down the term in the message, occasionally losing data by eliminating keywords without another word; CNN (Convolutional Neural Networks) is a machine learning model that has documented amazing results in image recognition quite some time ago and has recently documented amazing results in natural language handling, there is a convolutional layer to create a portion of sayings can be started to think of as together.

2. BACKGROUND

Using user evaluations within this enormous pool of user-produced data has led to the discovery of several functional applications in the commercial and government insight sectors. The literature has taken into consideration Sentiment Research [4] on all mediums (text, picture, video, and sounds) of social information. Essential research using vocabulary, machine learning, and evolutionary algorithms is widely available. The literature on multi-modal [5-7] and heterogeneous emotional analysis [14-18] is remarkable, with audits and overviews. The majority of sentiment analysis research has relied upon sentiment classification with simple words Jaiswal and Kumar [20] used supervised SC approaches to explore 2 different holding capacities (Facebook as well as YouTube) combined sentiment classification. In another paper, Kumar and Sebastian[21] suggest but also analyze a methodology for harvesting emotion through Instagram account, along with a composite technique that determines the translational orientation of evaluative phrases in retweets using the other annotation and English oxford methods. An auxiliary [21] summarizes the major research in the field of textual sentiment classification. Young [22] claimed that continuous patterns and Natural language processing were allowed in pattern recognition. To determine sentiments on Twitter, Felbo [23] developed a mix of consideration-based BiLSTM and CNN. Concurrently, Furthermore, image data

augmentation was already considered. relevant literature research on visually sentimental analysis. Zhao [24] suggested a model predict individualized image emotions assessments one per unique spectator, taking into account characteristics such as visualizations, social context, transitory development, and area effect, and tested it on the Flickr dataset. Kumar [17] suggested a visual emotional system based on a convolutional neural network as well as tested their model using photos from Flickr and Twitter. On picture emotions, many probability circulation models were presented. Zhao [25] suggested a methodology to forecast the continuous confidence interval of photograph sentiments discussed in the electro negativity dimension and developed an Appearance collection to show the transmission of sentiments and use a stochastic mixture model. The author furthermore presented an AI method for representing the objective visual sentiment as a disconnected possibility delivery (DPD) [24]. An additional review [26] addressed the issue of regional variation in picture emotion detection they talked about ways to unsupervised modify discontinuous possibility distribution function such as figure sensations beginning such a foundation environment route for an objective space. Their Emotion GAN model streamlines the Generative Adversarial Network (GAN) loss, linguistic compatibility loss, and extrapolation loss. Morency [27] addressed the multimodality problem as an emerging field of sentimental analysis study by using features-level textual integration, voice, as well as live coverage within the Statistics from YouTube. De Silva [28] five sentiments were distinguished among characteristics using analysis measures as well as deep Belief networks. (video) and impassioned speech (furious, scorn, fear, happy, gloomy, and astonishment) (sound). Sebe [29] suggested a unique audible feeling identification method that is tested taking place 38 people amid 11 Computer influence states. An additional evaluation agreed that the usage of something like the Invisible Dynamical System Chains on behalf of the identification of auditory emotions is appropriate [30]. Sun [31] created a library of facial expressions and tested a few computer vision computations to express feelings recognition during authentic films. Wöllmer [32] examined opinion in the ICT-MMMO dataset using auditory, visual, and literary approaches highlighted with cross hybrid fusion. Rozic [33] developed a composition of SVM trees containing voice, text, and video highlights. Metalline [34] used an early fusion technique with sound and text-based methodological highlighting for feeling recognition. Wu [37] chose to blend sound and linguistic signals. Nicolaou [38] used the results of sound and face LSTMs to anticipate emotion. Poria [39] used a convolutional neural network (CNN) that segregates items first from modality (message, sound, as well as record) along with then used multiple-kernel knowledge (MKK) to analyze opinions. The author goes on to expand on the CNN and MKL groups [40]. In another review, the authors [41] used OpenSMILE [36] to eliminate audio components and text2vec [42] as well as the grammatical form to extract literary highlights. Pérez Rosas analyzed the Delivery methods Multilingual Sentiment Interjections Database, including includes Netflix product auditing, includes [43]. Tensor combination is used by Zadehy [44]. McDuff [45] demonstrates the use of face inspection to assess the preferences of American voters. Siddiquie[46] suggested a very effective multimodal technique for automatically classifying politically persuasive online recordings by extracting the sound, visual, and written highlights. Author Poria [47] provides a thorough overview of one of the key phases of such a methodology for diverse influence characterization, Zhao [48] suggested a multi-modular microblog categorization approach as in several co-educational settings systems toward categorizing online press content hooked on distinct substances like as organizations, goods, as happenings to analyze potential commercial viability, significance, overall impact. Kumar [49] suggested a multimodal emotional analysis approach for determining sentiment intensity and scoring for literary, image, and typographic Twitter data. Kumar [50] Sarc-M was offered also as mimicking design identification within typographic memes employing classification algorithm phonological, behavioral, and ontological considerations in another focus. Recently, the use of data models in multimodal emotional analysis has been discussed. Paridhi[51] used the astounding Bayesian categorization approach on behalf of multidimensional emotion classification as well as classified emotions as joyful, miserable, or impartial. Majumder [52] developed a situationally layered fusion paradigm for print, sound, and video. Another effort by Poria [53] uses lengthy short-term memory to segregate logical data from encompassing phrases (LSTM). Senti Bank and at least three reasons ratings were employed for regional utilizing convolution neural network (CNN), while for creative sentiments, a situation combination (terminology as well as Intelligent systems) method was applied.

3. MATERIAL AND METHODS

3.1 Frame Work And Design Goal

The recommended Multimodal Image and Text with Attention (MITA) method is made up of three modules: text representation, image representation, and classification (Fig. 2). The Text summarization component is working on the creation of message extraction utilizing a PS-POS word-based layer connected to emotional analysis using BiLSTM-CRF [2].

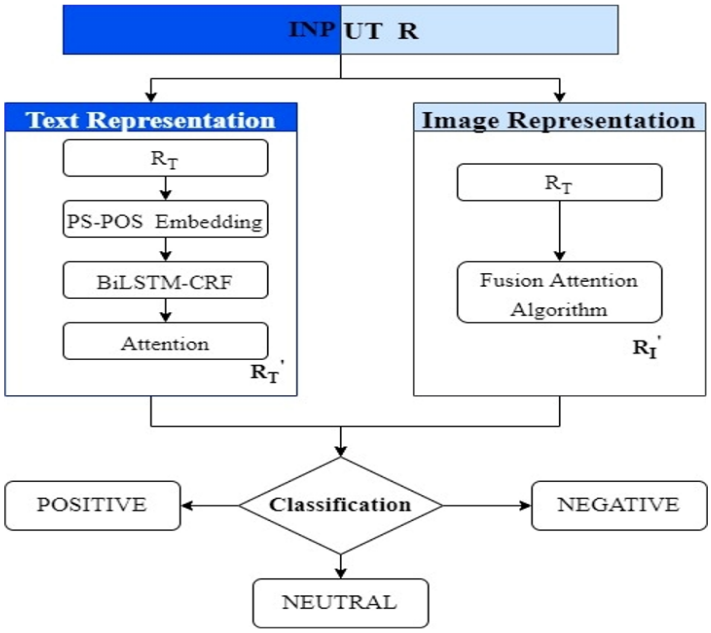

Fig. 2: Flow Diagram of Proposed Methodology

The spatial arrangement module does image analysis using the Fusion Attention method [57]. The classification module is responsible for categorizing the valence level in the emotional view of something like the Twitter15 dataset. The extra sub-areas go through the specifics of each module.

3.2 Input

A little amount of data is frequently seen as an absolute requirement to gain greater accuracy. The information data R for this suggested model was gathered from several internet-based media platforms such as Facebook, Insta, Flicker, and Tweets, in which the photographs were used with the typical purpose of sharing and passing the data as a replacement for textual information. A few photographs within those databases have also been produced from the Twitter15 Dataset, which includes a larger collection of images that may be worked on. In this work, unique datasets were created that combined photos from online stages.

3.3 Text Representation Module

This is the primary task for evaluating the targeted representation modeling utilizing PS-POS (Predecessor Successor - Parts of Speech) word embeddings combined with character embeddings. The textual portion of the entry R is taken also as RT. The POS provides a plethora of details about a term and its neighbors, as well as morphological classifications of words and similarities and differences between them. The provided POS tag is transformed into a good vector subsequently harmonized with PS-POS embedded vectors. As a result, PS-POS embedded vectors will contain word-synthesized information. A concise depiction of the BiLSTM-based sequence of organization labeling frameworks with possible CRF layers is provided in the following sections [1]. CNN has outperformed PC vision processing, speech confirmation, and normal language planning because of its capacity to erase neighborhood interconnections of geographical or transitory structures.

3.3.1 **Bidirectional Long Short Term Memory**

U and Z are our two emblems. Our mission, or unary scores, are denoted by U(x, y). It's essentially a rating for the mark y there at kth previous time step given our X vector. We may think of it as the kth outcome of a BiLSTM. In principle, our X vector may be anything. Like word representations from a moving window, our X vector is often the connection of enclosing constituents. In our approach, each unary constituent is evaluated by a memorization load. Provided we view them as LSTM outputs, this is simple.

The partition function is denoted by Z(x). We may think of it as a standardizing element because we may require probability in the future. This is similar to the minimum of the Softmax. To derive probabilities, we used a

typical clustering model with the "latest Softmax activation." We are currently adding additional learnable loads to demonstrate the gunshot at a labeling yk being tracked by yk+l. By revealing this, we are establishing trust between progressive names. In this vein, the moniker linear chain CRF! To do so, we double our previous probability by P(yk+l |yk), which we may update using exponential properties as integer arithmetic scores U(x, y) and teachable transitional values T(y, y):

$$P(\mathbf{y} \mid \mathbf{X}) = \frac{\exp\left(\sum_{k=1}^{\ell} U(\mathbf{x}_k, y_k) + \sum_{k=1}^{\ell-1} T(y_k, y_{k+1})\right)}{Z(\mathbf{X})} \quad (1)$$

T(y, y) should be seen in code as a foundation with form (nb labels, nb labels), with each passage discussing the progress of traveling from the i-th surname to the j-th name. We should go through all of our new factors:

- Emissions or unary scores (U): scores addressing how probably is Uk given the info Xk.
- Transition scores (T): scores addressing how probably is yk trailed by yk+l
- Partition function (Z): standardization factor to get likelihood dispersion over successions.

$$Z(\mathbf{X}) = \sum_{y_1'} \sum_{y_2'} \cdots \sum_{y_k'} \cdots \sum_{y_\ell'} \exp\left(\sum_{k=1}^{\ell} U(\mathbf{x}_k, y_k') + \sum_{k=1}^{\ell-1} T(y_k', y_{k+1}')\right) \quad (2)$$

3.3.2 **Conditional Random Field**

A level advancement grid serves as the border of a Conditional Random Field (CRF). CRF effectively uses past and future labels to predict the present label, the same as the BiLSTM layer uses past future present knowledge. The CRF equation is as follows, where Y represents the hidden state (for example, a grammar characteristic) when X is observable within this approach, parameters are the material or language variety that surrounds it.

The CRF equation is divided into two parts:

$$p(y|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} \exp\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, X_t)\} \quad (3)$$

1. Normalization: As you can see, there is no possibility that the right component of the equation, contains the weights and characteristics. In any instance, the outcome is assumed to be likely, and standardization is therefore required. The constant of normalization Z(x) is the sum of all possible state groupings such that the fundamental becomes 1.

2. Loads and Highlights: This section provides the estimated relapsing equations with pressures and the comparative highlights. The greatest likelihood assessment performs the weighted assessment k, and we characterize the components.

After the process of BiLSTM the CRF is performed we arrive at RT'.

3.4 Image Representation Module

The input picture is separated and presented as RI. The picture is then sent through the Merge Attention technique after shrinking preparation, resulting in RI'. We use a Fusion Attention algorithm [57], which consists of three attention layers. We create a channel attention map by utilizing the feature between channel relationships. Because each channel of such a feature map is thought of as a feature indication, channel focus is on 'what' is important given an input picture. The inter-spatial connection of highlights is used to generate a map using spatial attention. In contrast to channel attention, spatial attention focuses on where there are informative portions; this is complementary to channel attention. The Compression block modeling and simulation recalibrates and improves the quality of a network's depictions by explicitly presenting the interdependencies between its convolutional highlights channels. When we combined all three great attentions, we got an approach with a lot of variances. Convolution is performed using the method using degree increments. In each convolution, the various attentions are switched, as well as the maximum and average pooling. The picture is initially run through three levels of 64-bit convolution, followed by pooling, and then every level has three attentions. Later, the procedure is repeated for 128-bit and 256-bit convoluted by varying the intensity of attention. This exchange and recurrence are done to provide the viewer with a clear picture of the imagery. The following is a basic description of the attention layers.

Algorithm

Step 1: Input Image RI

Step 2: Three partition PA, PB, PC

Step 3: Conv64(with k size of 5,7,7 for three partitions respectively)

Step 4:Conv64(with k size of 3,5,7 for three partitions respectively)

Step 5: αPA= MP(FSE(MP(PA)))

Step 6: αPB = AP(FS(MP(PB)))

Step 7: αPC= MP(FC(MP(PC)))

Step 8: Conv128(with k size of 5,7,3 for three partitions respectively)

Step 9: βPA= MP(FS(αPA))

Step 10: βPB = AP(FC(αPB))

Step 11: βPC= MP(FSE(αPC))

Step 12: Conv256(with k size of 7,5,3 for three partitions respectively)

Step 13: γPA= FC(βPA)

Step 14: γPB = FSE(βPB)

Step 15: γPC= FS(βPC)

Step 16: RV'= Softmax ((γPA+ γPB)+(γPB+γPC))

Where MP-Max Pooling, AP-Average Pooling, Fs-Function of spatial attention, FSE-Function of squeeze and excitation attention, FC-Function of channel attention.

3.5 Classification Module

In terms of Sentiment classification, the two presented components give a reliable process for both image and text. Based on two inputs of both Photos RI' and textual RT' We group the user's sentiment classification into three broad categories: positive, negative and neutral.

4. EXPERIMENTAL RESULTS ANALYSIS AND PERFORMANCE EVALUATION

4.1 Evaluation Metrics

We have used the conventional classification performance and Macroeconomic F1 as assessment measures in several earlier works for entity-level sentiment categorization [9]. One criterion for evaluating classification models is accuracy. Essentially, Authenticity is the fulfillment of our actress's assumptions. Accuracy has the following accepted definition:

$$\text{Accuracy} = \frac{\text{Number of correct prediction}}{\text{Total number of prediction}} (4)$$

Efficiency in the classification model may also be measured in terms of either positives or negatives, as seen below:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} (5)$$

Where TP denotes True Positives, TN denotes True Negatives, FP denotes False Positives, whereas FN denotes False Negatives.

The Macro F1 is referred to as the average of the F1 scores for each class/label. Macro F1 is defined as follows:

$$\text{Macro F1} = \frac{\text{TP}}{\text{TP}+\frac{1}{2}(\text{FP}+\text{FN})} \quad (6)$$

4.2 Teenagers Sample

In this work, the survey approach was applied. A sample of 100 students ranging in age from 16 to 19 years was used. Participants in this sample mostly utilized social media via mobile phones and PCs. Their consistent updates were recognized on Facebook, Twitter, and Instagram. Their data was sent to Psychiatrists. We assessed the good and negative statistics with their assistance. Later, psychiatrists gave a database of kids who were stressed based on their social Facebook posts. The dataset was generated using the psychological report. This information was put through the suggested method, which resulted in effectiveness of 72.43%. Table 1 shows the results of the dataset constructed using classmates' posts.

Table 1: Classification Result of Teen Sample dataset

	Positive	Negative	Neutral	Total
Facebook	581	301	121	1003
Instagram	689	427	270	1386

Table 2: Classification Result of the Twitter15 dataset

	Positive	**Negative**	**Neutral**	**Total**
Train	928	368	1883	3179
Dev	303	149	670	1122

Table 2 shows the outcome of the data execution. This table displays the most significant difference between the preceding works. The suggested system is compared to current algorithms in Table 3, and a chart is given in Figs. 2 and 3. The comparable methods are listed below for Text, Image, and Text + Image.

Text

- The BiLSTM-CRF model focuses target views from input weights of the connections and clusters each phrase according to the amount of target unequivocally stated in it. Then, we present the 1d-CNN emotional classification model, which estimates opinion polarity separately for ou pas sentences, one-target phrases, and multi-target sentences [55].
- CNN was hired to quickly extract the text's nearby highlights. The BiLSTM was then used to extract global text highlights carrying important semantics. The components segregated by the three neural network models (NNs) were then concatenated and processed by the Softmax classifier for document-level sentiment categorization [56].

Image

- Res-Target is a fundamental combination of objective element depiction and visual context representation that essentially connects Ht and Qv, followed by the following vectors to a SoftMax layer providing order [9].
- Resnet50, but instead of learning unreferenced functions, learn residual functions based on layer inputs. Rather than assuming that every combination of stacked layers will suit an optimum encoding, residual nets allow these layers to fit the purpose of the survey. They build a network by stacking the remaining bricks on top of each other's [3].
- Alexnet is made up of alternating layers with learnable limits. The model has five layers, with a mix of convolution and pooling being followed by three entirely related layers, and they apply RELU activation in all of these levels except the outcome layer [3].

- VGG16 has 16 layers. In the last four completely linked layers, the evolution of VGGs is the same. The overall structure has 5 convolutional layer configurations, followed by a MaxPool. What important is that ever more entire convolutional layers are memorized for the five convolutional layer layouts [3].

Text + Image

- Res-RAM, Res-MGAN, and Res-ESTR, three variants of a straightforward multimodal fusion strategy [11], which first applies max-pooling over the visual elements to obtain g = MaxPool(R), and then links g and the text-based portrayal from RAM, MGAN, and ESTR, followed by a Softmax layer for characterization.
- Res-RAM-TFN and Res-MGAN-TFN, two variants of another type of multimodal convolution [12], which use a bilinear interaction operator to combine g and the textual recognition from RAM and MGAN via a complex fusion matrix and feed the resulting matrix to the Viewpoint Inference Network segment for final classification;
- MIMN, a new cutting-edge multidimensional approach for aspect-level sentiment opinion order to address the issue by Xu et al. [13], which employs some co recollection network to exemplify the interactive attention here between aspect word, the written text circumstances, and the visual context; ESAFN, based on entity-sensitive textual and visual representations, we employ another bilinear pooling superintendent to capture their changes have taken place communication [54].

Table 3: Performance analysis

TEXT	**%**		**IMAGE**	**%**		**TEXT + IMAGE**	**%**	
	Macro F1	**ACC**		**Macro F1**	**ACC**		**Macro F1**	**ACC**
BiLSTM-CRF + 1D-CNN	66.73	85.98	Res-Target	46.48	59.88	Res-RAM	64.68	71.55
BiLSTM + CNN	68.65	88.46	Resnet50	25.02	60.06	Res-RAM-TFN	61.49	69.91
BiLSTM-CRF + 2DCNN + Attention [PROPOSED]	**70.00**	**90.20**	Alexnet	47.41	59.28	Res-MGAN	63.88	71.65
			VGG16	25.74	62.89	Res-MGAN-TFN	64.14	70.3
			FUSION ATTENTION [PROPOSED]	**51.29**	**64.15**	MIMN	65.69	71.84
						Res-ESTR	63.98	72.03
						ESAFN	67.37	73.38
						MITA [TWITTER]	**69.14**	**75.67**
						MITA [TEENSAMPLE]	**69.04**	**73.51**

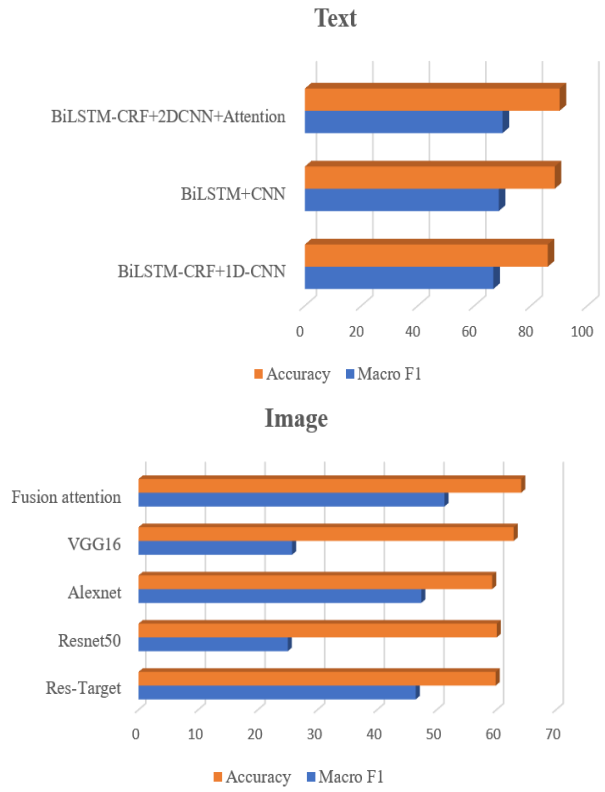

Fig.4: Performance of the Text and Image with Existing Algorithms

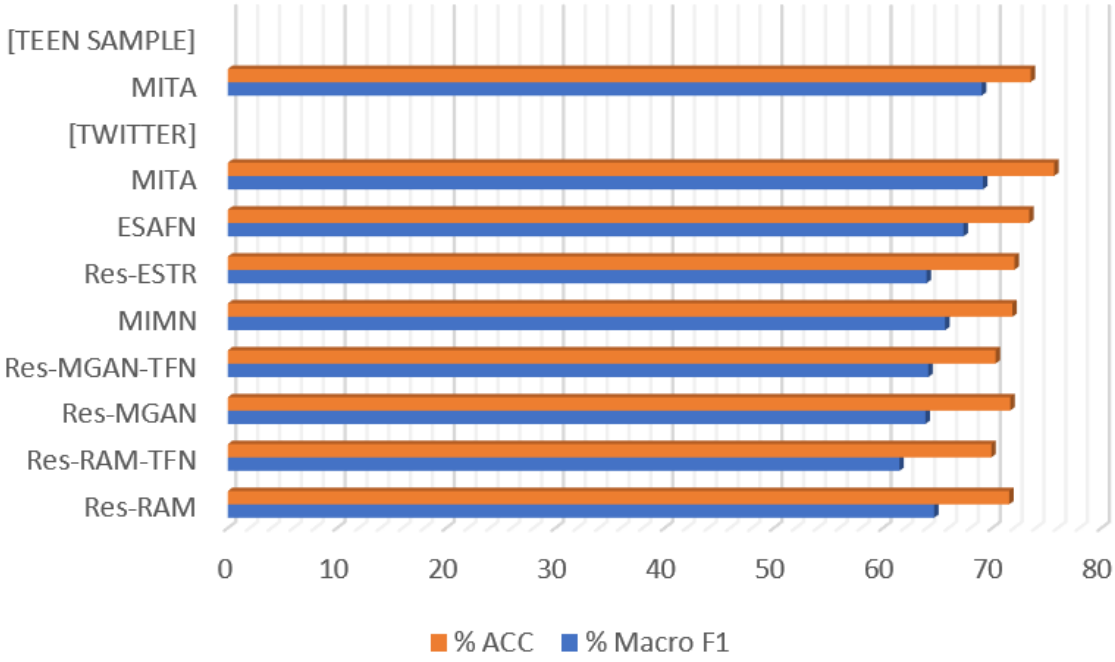

Fig.5: Performance of the MITA with Existing Algorithms

5. DISCUSSION

The text module has an accuracy level of 90.2%, while the image module has a level of 64.15%. When integrated with both categorization modules, the accuracy jumps to a staggering 75.67%. The experimental evaluation revealed that this technique is significantly more productive than previous methods. We also see that the F1 score for the Twitter dataset is at a high of 69.14%, a substantial increase above the picture module's 51.29% and close to the text module's 70.00%. The accuracy of the teen sampling dataset is 73.51%, and the macro F1 score is 69.04%.

6. CONCLUSION

We focused on multimodal sentiment categorization in this research and suggested a Multimodal Image and Text with Attention (MITA) technique to depict interpersonal and inter exchanges including text and image. To get a smooth result, the text examination was carried out using the BiLSTM-CRF. The picture analysis was then followed by the fusion attention method. The CNN network is a compact and deep network that is more resistant to overfitting for the Twitter15 dataset than traditional neural networks in the first phase. Furthermore, it has significant spatial-based classification capabilities, such that densely linked CNN designs may properly extract spatial data from different scales and merge them. Second, to improve the extraction of key highlights, we expanded the channel domain by employing a channel attention method to increase the hardness of significant data while decreasing the heaviest of low-need material. Third, we use the squeeze excitation respect to operate on the photos' notable components. The CNN network was shown to be dominant in text and picture classification tests using the Twitter15 dataset. In future studies, we will investigate new models and develop applications for emotional analysis.

References

1. Roseline, V., & Dr. Heren Chellam, G. (2020). “PS-POS embedding target extraction using CRF and BiLSTM”. Int. J. of Advanced Science and Technology, 29(3), 10984–95. http://sersc.org/journals/index.php/IJAST/article/view/27986.
2. V. Roseline, Dr. Heren Chellam G. (2020). “Sentiment Classification Using PS-POS Embedding with BiLSTM-CRF and Attention”. Int. J. of Future Generation Communication and Networking, ISSN: 2233-7857, Vol. 13, no. 4, Dec. 2020, pp. 3520-3526.https://sersc.org/journals/index.php/IJFGCN/article/view/34482.
3. Dr. Heren Chellam, G., & Roseline, V. (2021). “Analysis of sentimental images using deep learning approach”. J. Math. Comput. Sci. 11(5), 5474–5486. https://doi.org/10.28919/jmcs/5835.
4. Dave, K., Lawrence, S., & Pennock D, M (2003).” Mining the peanut gallery: Opinion extraction and semantic classification of product reviews”. Proceedings of the 12th international conference on World Wide Web (pp. 519–528). ACM.https://doi.org/10.1145/775152.775226.
5. Pang, B., Lee, L., &Vaithyanathan, S. (2002). “Thumbs up? sentiment classification using machine learning techniques”. In: Proc. Empirical Methods on Natural Language Processing (pp. 79–86).
6. Ravi, K., & Ravi, V. (2015). “A survey on opinion mining and sentiment analysis: Tasks, approaches, and applications”. Knowledge-Based Systems, 89, 14–46.
7. Appel, O., Chiclana, F., & Carter, J. (2015).” Main concepts, state of the art and future research questions in sentiment analysis”. Acta PolytechnicaHungarica, 12, 87–108.
8. Xin Li, Lidong Bing, Wai Lam, and Bei Shi. “Transformation networks for target-oriented sentiment classification”. In Proceedings of ACL, 2018.
9. Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. “Recurrent attention network on memory for aspect sentiment analysis”. In Proceedings of EMNLP, 2017.
10. Feifan Fan, Yansong Feng, and Dongyan Zhao. “Multi-grained attention network for aspect-level sentiment classification”. In Proceedings of EMNLP, 2018.
11. Devamanyu Hazarika, SoujanyaPoria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. “Conversational memory network for emotion recognition in dyadic dialogue videos”. In Proceedings of NAACL, 2018.
12. Amir Zadeh, Minghai Chen, et al. “Tensor fusion network for multimodal sentiment analysis”. In Proceedings of EMNLP, 2017.
13. Nan Xu, Wenji Mao, and Guandan Chen. “Multi-interactive memory network for aspect-based multimodal sentiment analysis”. In Proceedings of AAAI, 2019.
14. Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-. F., &Pantic, M. (2017). “A survey of multimodal sentiment analysis”. Image and Vision Computing, 65, 3–14.
15. Hussien, Intisar O., &Jazyah, Yahia Hasan (2018). “Multimodal sentiment analysis: A comparison study”. Journal of Computer Science, 14(6), 804–818.
16. Fulse, S., Sugandhi, R., & Mahajan, A. (2014). “A survey on multimodal sentiment analysis”. International Journal of Engineering Research and Technology, 3(11), 1233–1238.
17. Medhat, W., Hassan, A., &Korashy, H. (2014). “Sentiment analysis algorithms and applications: A survey”. Ain Shams Engineering Journal, 5, 1093–1113.
18. Marjan, S. (2014). “A survey for multimodal sentiment analysis methods”. Int.J.Computer Technology & Applications, 5, 1470–1476.

19. Kumar, A., & Jaiswal, A. (2017). December. ”Image sentiment analysis using convolutional neural network”. International Conference on Intelligent Systems Design and Applications (pp. 464–473). Springer.
20. Kumar, A., & Jaiswal, A. (2017). “Empirical study of Twitter and Tumblr for sentiment analysis using soft computing techniques”. Proceedings of the world congress on engineering and computer science. 1. Proceedings of the world congress on engineering and computer science (pp. 1–5).
21. Kumar, A., & Sebastian, T. M. (2012). “Sentiment analysis on Twitter”. International Journal of Computer Science Issues. 9. International Journal of Computer Science Issues (pp. p.372–). IJCSI.
22. Young, T., Hazarika, D., Poria, S., & Cambria, E (2018).” Recent trends in deep learning based natural language processing”. IEEE Computational Intelligence Magazine, 13(3), 55–75.
23. Felbo B., Mislove A., Søgaard A., Rahwan I., Lehmann S. (2017). “Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm”, arXiv preprint arXiv:1708.00524.
24. Zhao, S., Ding, G., Gao, Y., & Han, J. (2017). “Approximating discrete probability distribution of image emotions by multi-modal features fusion”. Transfer, 1000(1), 4669–4675.
25. Zhao, S., Yao, H., Gao, Y., Ding, G., & Chua, T. S. (2016). “Predicting personalized image emotion perceptions in social networks”. IEEE Transactions on Affective Computing, 9(4), 526–540.
26. Zhao, S., Zhao, X., Ding, G., &Keutzer, K. (2018). “EmotionGAN: Unsupervised domain adaptation for learning discrete probability distributions of image emotions”. 2018 ACM multimedia conference on the multimedia conference (pp. 1319–1327). ACM
27. Morency, L. P., Mihalcea, R., & Doshi, P. (2011). “Towards multimodal sentiment analysis: Harvesting opinions from the web”. Proceedings of the 13th International Conference on Multimodal Interfaces (pp. 169–176). ACM Nov. 14-18.
28. De Silva, L. C., & Ng, P. C. (2000). “Bimodal emotion recognition”. Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (pp. 332– 335). IEEE Cat. No. PR00580.
29. Sebe, N., Cohen, I., Gevers, T., & Huang, T. S. (2006). “Emotion recognition based on joint visual and audio cues”. 18th International Conference on Pattern Recognition. 1. 18th International Conference on Pattern Recognition (pp. 1136–1139). IEEE ICPR'06.
30. Song, M., Bu, J., Chen, C., & Li, N. (2004). “Audio-visual based emotion recognition-a new approach”. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2IEEE 2004CVPR 2004.
31. Sun, Y., Sebe, N., Lew, M. S., &Gevers, T. (2004). “Authentic emotion detection in the real-time video”. International Workshop on Computer Vision in Human-Computer Interaction (pp. 94–104). Springer.
32. Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., et al. (2013).” YouTube movie reviews: Sentiment analysis in an audio-visual context”. IEEE Intelligent Systems, 28(3), 46–53.
33. Logic, V., Ananthakrishnan, S., Saleem, S., Kumar, R., & Prasad, R... (2012). “Ensemble of SVM trees for multimodal emotion recognition”. Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 1–4). Asia-Pacific, IEEE 20122012.
34. Metallinou, A., Lee, S., & Narayanan, S. (2008). “Audio-visual emotion recognition using gaussian mixture models for face and voice”. Tenth IEEE International Symposium on ISM (pp. 250–257). IEEE 2008.
35. Eyben, F., W¨ollmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., & Cowie, R. (2010). “On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues”. Journal on Multimodal User Interfaces, 3(1–2), 7–19.
36. Eyben, F., Wöllmer, M., & Schuller, B. (2010). “open smile: The Munich versatile and fast open-source audio feature extractor”. ACM International Conference on Multimedia MM.
37. Wu, C.-. H., & Liang, W.-. B. (2011). “Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels”. IEEE Transactions on Affective Computing, 2(1), 10–21.
38. Nicolaou, M. A., Gunes, H., &Pantic, M. (2011). “Continuous prediction of spontaneous effect from multiple cues and modalities in valence – Arousal Space”. IEEE Trans. Affect. Comput. 2(2), 92–105.
39. Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016). “Convolutional mkl based multimodal emotion recognition and sentiment analysis”. ICDM, Barcelona (pp. 439–448).
40. Poria, S., Peng, H., Hussain, A., Howard, N., & Cambria, E. (2017). “Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis”. Neurocomputing, 261, 217–230.
41. Poria, S., Cambria, E., Howard, N., Huang, G. B, & Hussain, A (2016). “Fusing audio, visual and textual clues for sentiment analysis from multimodal content”. Neurocomputing, 174, 50–59.
42. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). “Distributed representations of words and phrases and their compositionality”. Advances in Neural Information Processing Systems, 23, 3111–3119.
43. Rosas, V. P., Mihalcea, R., &Morency, L.-. P. (2013). “Multimodal sentiment analysis of Spanish online videos”. IEEE Intelligent Systems, 28(3), 38–45 2013.
44. Zadehy, A., Chen, M., Poria, S., Cambria, E., &Morency, L.-. P. (2017). “Tensor Fusion Network for Multimodal Sentiment Analysis”. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017. Copenhagen, Denmark: EMNLP, Association for Computational Linguistics1103–1114.
45. McDuff, D., Kaliouby, R., Kodra, E., & Picard, R. (2013). “Measuring voter's candidate preference based on affective responses to election debates”. Conference on Affective Computing and Intelligent Interaction (ASCII).
46. Siddiquie, B., Chisholm, D., &Divakaran, A. (2015). “Exploiting multimodal effect and semantics to identify politically persuasive web videos”. ACM on International Conference on Multimodal Interaction 2015.

47. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). “A review of affective computing: From unimodal analysis to multimodal fusion”. Information Fusion, 37, 98–125 2017.
48. Zhao, S., Gao, Y., Ding, G., & Chua, T. S. (2017). “Real-time multimedia social event detection in microblog”. IEEE Transactions on Cybernetics, 48(11), pp.3218–pp. 3231.
49. Kumar, A., & Garg, G. (2019). “Sentiment analysis of multimodal Twitter data”. Multimedia Tools and Applications, pp.1–pp17.
50. Kumar, A., & Garg, G. (2019). “Sarc-M: Sarcasm Detection in Typo-graphic Memes”. International Conference on Advances in Engineering Science Management & Technology (ICAESMT) 2019 Available at SSRN 3384025.
51. Paridhi, Gupta, Tanu, Mehrotra, Ashita, Bansal, & Bhawna, Kumari (2017). “Multimodal sentiment analysis and context determination: Using perplexed Bayes classification”. Proceedings of the 23rd International Conference on Automation & Computing University of Huddersfield.
52. Majumder, N., Gelbukh, A., Hazarika, D., & Cambria, E. (2018). “Multimodal sentiment analysis using hierarchical fusion with context modeling”. Knowl-Based Syst, 161, 124–133.
53. Poria, S, Cambria, E, Hazarika, D., Majumder, N., Zadeh, A., &Morency, L. P. (2017). “Multi-level multiple attentions for contextual multimodal sentiment analysis”. ICDM 2017 (pp. 1033–1038). IEEE.
54. J Yu, J Jiang, R Xia “Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification”, IEEE/ACM Transactions on Audio, Speech, and Language Processing 28, 429-439.
55. TaoChena, RuifengXuab, YulanHec, XuanWanga, “Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN”. Elsevier Expert Systems with ApplicationsVolume 72, 15 April 2017, Pages 221-230.
56. C. Zhang and M. J. Zhou, “A text sentiment classification method based on the fusion of CNN and two-way LSTM,” Computer Times, ed. Yunhe Pan, Zhejiang: Department of Science and Technology of Zhejiang Province, vol. 12, pp. 38- 41, 2019.
57. V. Roseline and Dr. G. HerenChellam(2022), “A Novel Fusion Attention Algorithm for Sentimental Image Analysis“, Indian Journal of Science and Technology 15(9):386-394.