INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

# Context Aware Image Sentiment Classification using Deep Learning Techniques

**Sohit Agarwal**[1]*, **Mukesh Kumar Gupta**[2]

**1** Research Scholar, Department of Computer Engineering and Information Technology, Suresh Gyan Vihar University, Jaipur, Rajasthan, India
**2** Professor, Department of Electrical Engineering, Suresh Gyan Vihar University, Jaipur, Rajasthan, India

OPEN ACCESS
**Received:** 20-09-2022
**Accepted:** 08-11-2022
**Published:** 21-12-2022

**Citation:** Agarwal S, Gupta MK (2022) Context Aware Image Sentiment Classification using Deep Learning Techniques . Indian Journal of Science and Technology 15(47): 2619-2627. https://doi.org/10.17485/IJST/v15i47.1907

* **Corresponding author**.

sohit.agarwal@gmail.com

**Funding:** None

**Competing Interests:** None

Published By Indian Society for Education and Environment (iSee)

**ISSN**
Print: 0974-6846
Electronic: 0974-5645

## Abstract

**Objectives**: To propose context aware sentiment classification using deep learning techniques. **Methods:** We used EfficientNetB-7 deep learning framework for caption generation for the input image and to classify the sentiment of generated caption using machine learning techniques. First, we employ several real-time and synthetic image datasets, then apply pre-processing and normalization for data balancing. Then efficient module implementation for feature extraction and selection using convolutional and pooling layers were done. Despite this proceeding, it generates the caption for respective images. The various feature extraction and selection Natural Language Processing (NLP) techniques such as TF-IDF, lemmas, dependency and correlational features have been used and classify the sentiment label using attention model and greedy approach. Finally, generating the blue score for the entire testing dataset and show the effectiveness of the proposed system. **Findings**: Our model gives higher accuracy with different deep learning techniques which is demonstrated in result section. The proposed model archives 73.80% average accuracy for EMOTIC dataset. The module has evaluated with different features and deep learning classification algorithms proposed earlier. **Novelty**: This research is the collaboration of Deep learning and machine learning classification techniques. We first extract the visual features from the input image using deep learning and classify with machine learning with the collaboration of NLP processes. We also carried out various feature extraction techniques such as N-gram, dependency features, co-relational features and determined the sentiment of generated captions.

**Keywords:** Image Sentiment analysis; Emotic dataset; CNN; EfficientNetB7; Attention based LSTM; GRU

## 1 Introduction

Context-based sentiment analysis presents several challenges, the most significant of which is the inclusion of multilingual data. Many social networking websites, including

Facebook and Instagram, place a greater emphasis on the posting of images than they do on the posting of text. We require thoughtful methods, and a significant amount of work will be involved to evaluate the emotions conveyed by images. The various authors have proposed similar methods for the detection of the sentiment using numerous deep learning and machine learning methods on various social media sources[1–4]. Numerous systems face problems, such as high error rates, low accuracy, overfitting problems etc.

Visual sentiment analysis for social media using deep learning has been developed by Ganesh Chandrasekaran et al. in[1]. The pretrained module has been used to detect image sentiment using VGGNET, RESNET and DENSENET. This module works for specific pretrained images only, not for real-time images. It also generates an overfitting problem when it deals with heterogeneous image datasets. The context-based emotion detection techniques have been developed by Manh Hung Hoang et al. in[2]. The EMOTIC dataset has used sentiment prediction according to internal and external features. In significant issues of this system, it can't provide accurate sentiment when it deals with complex data such as valence and arousal images. Various sentiment analysis models have been developed by Mayur Wankhede et al. in[3]. The process of sentiment analysis faces many difficulties. These difficulties make it harder to evaluate emotions correctly and choose the correct sentiment identification. With text extraction and NLP, sentiment analysis locates and extracts emotional information from the text. This approach also covers a thorough explanation of the procedure for carrying out this work as well as sentiment analysis applications. Then, to fully comprehend the benefits and drawbacks of each method, it assesses, contrasts, and researches them. This approach is only evolutionary based according to theoretical evaluation, and no implementation has been done.

According to[4], the Transfer learning techniques using the VGG19 model have been used for detecting and predicting image sentiment using visual features. VGG19 model has been used with different image datasets such as CK+, JAFEE and FER2013. The major problem of this approach is very low accuracy for the CK+ dataset, and it is around 65% accurate detection of sentiment. It also produces a higher error rate on a heterogeneous image dataset. Borth et al.[5] published a dataset that includes more than 3000 associate noun pairs to assist researchers in contributing to the field. Their research also includes a collection of baseline models, which are typically applied in the process of benchmarking approaches that are based on associate noun pairs.

The utilization of deep neural networks and transfer learning from architectures that have been trained on huge-scale categorization datasets, like ImageNet, is another commonly utilized method and gets around the requirement for massive sentiment datasets. For example, Chandrasekaran et al.[6] used a Twitter dataset to fine-tune a current pre-trained model called VggNet to extract and categorize the generated emotions. Additionally, certain projects use domain datasets to train image sentiment analysis methods[7]. Al-Halah et al.[8] developed a system for forecasting emoticons (also known as emojis) based on a certain image. Emojis are used to express one's feelings in response to a picture. They gathered a dataset from Twitter that included over 4 million photos and emoticon combinations, which they then utilized to train a unique convolutional neural network model called SimleyNet[8]. The dataset came from Twitter. Image sentiment analysis was also implemented in some of the works by using the accompanying text of the photographs. For example, in[9], the concept of an attention-based network known as Attention-based Modality-Gated Networks is presented to utilize the association among textual and visual data to do sentiment analysis. In more recent times, some attempts have been made to learn objective properties of photographs, like emotions, through auxiliary sources of data[10]. The feature extraction techniques used in this method cannot extract complex features such as structural dependency features and correlational features due to the system's high error rate issue.

Several different efforts for multi-level and multi-scale representations to extract visual cues for sentiment analysis have recently been produced. Ou et al.[11] developed a cross-layer feature fusion strategy that used a multi-level context pyramid network framework to merge local and global information. The utilization of both the local characteristics and the global context is the primary goal of the multi-level representation. This is necessary because the size and location of significant visual cues (i.e., items and background knowledge) in images might vary greatly. In addition to differences in size and location, visual clues of moods and emotions may include several items. Therefore, the relationship or interaction between the various items in an image should be considered while performing visual sentiment analysis. To include the interactive qualities of items, Wu et al.[12] offered a method utilizing Graph Convolutional Networks to retrieve interaction features. This approach was developed to obtain interaction features.

According to[13] this theory, the classifier can more successfully relate a user's prior behaviour to the content of a tweet by using past feelings. Consequently, enabling learning algorithms to link previous actions in figuring out the present feelings. To do this, we suggest three sliding window features for gathering historical emotion from the time series data. In this study, we offer seven versions of context-aware sliding window (CSW) features using machine learning and deep learning techniques.

According to[14] categories, expressing emotions into good or negative emotions and employing deep learning approaches in this case. Anger, boredom, emptiness, hatred, sorrow, and concern are negative emotions, whereas excitement, fun, happiness, love, neutrality, and relief are further subdivided into upbeat categories. We tried and assessed the technique of employing recurrent neural networks and long short-term memory on three distinct datasets to demonstrate high emotion classification accuracy. According to [15], the method uses increased feature weighting by attention layers and an LSTM-RNN-based network as its foundation. The system generates over fitting problem when it deals with unstructured and imbalanced data.

### 1.1 Gap analysis of existing methods

- Low accuracy with various CNN-GAN and RNN-GAN for detection of sentiment for real time image dataset.
- The ensemble framework can't detect all aspects or object consists in images, it effects on training, these experiments illustrate this module can detect only large objects.
- Ensemble can't detect new more features than RNN-GAN and CNN-GAN, it only collects unique features and detect as ensemble results.
- Some color object can't detect by both algorithms as well as ensemble model in image.

## 2 Methodology

We proposed a context aware sentiment detection using real time images dataset named Emotic Dataset. Our proposed system is designed to predict the sentiment of real time image. Initially, the real time synthetic dataset is subjected to some pre-processing and normalization procedures. In order to generate the image description for each distinct image, an EfficientNetB7 model is utilized. The classification of sentiment labels is accomplished through the utilization of Attention-based LSTM and GRU greedy method. At last, the BLEU score is utilized to assess the accuracy of the predictions generated by autonomous machine translation systems. Consider taking a look at the following illustration of the proposed system framework, Figure 1.

The following is an in-depth discussion of the detailed descriptions of each level of the proposed system:

The data was gathered using a synthetic real time dataset, and then it was subjected to pre-processing and normalization. Before feeding it as an input to the convolutional neural network, the data will be balanced here in preparation for that stage. We define the fixed image size as 300 × 300 pixels and get rid of any elements in the image that cannot be read. The processing of the data is performed on the host computer because the dataset is so enormous, and the findings are then stored as pickle files before being used for the model training process.

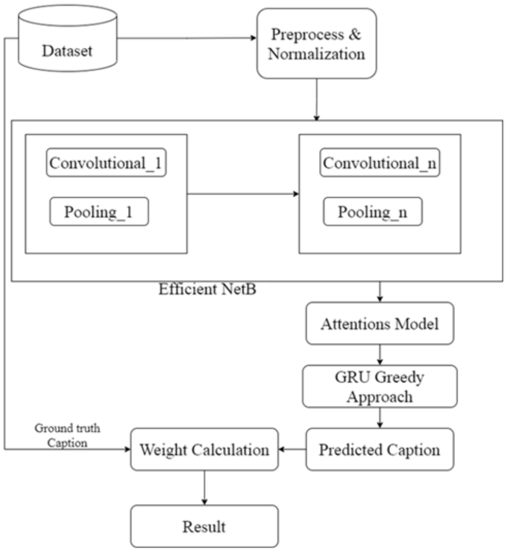


**Fig 1.** Proposed System Framework

### 2.1 Image Caption Generation

Image caption generation is the method of identifying the background of an image and labeling it with appropriate captions utilizing deep learning and machine vision. This technique is known as image caption generation. Image Caption Generation It entails the process of classifying a picture with English keyword phrases using the datasets that are presented during the model's training phase.

Image captioning seeks to accomplish the transformation of an image's specified input into a description written in human language. For the development of image captions, we relied on a model known as EfficientNetB7. This model combines the principles of machine vision and the Natural Language Processing system in order to understand the context of pictures and explain them using a natural language such as English.

The process of writing captions for images can be logically divided into two distinct parts:

• **Image-based modeling:** This method extracts features from the image we provide.

• **Language-based model:** This model takes the traits and objects gathered from our image-based model and convert them into a natural-sounding sentence.

For image-based and language-based models we make use of convolutional neural networks and attention-based LSTMs.

Our input image has its features extracted by a CNN that has been trained previously. In order for the feature vector to possess the same dimension as the input dimension of the LSTM model, a linear transformation is performed on it. This network has been trained as a language model using our feature vector as the training data. Consider the following example of image captioning in Figure 2.

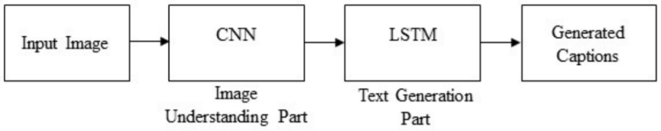


**Fig 2.** Example of image captioning

We predefine both the label and the target text before beginning the training process for our LSTM model. If the caption reads ”A man wears a hat,” then our caption and target would read as follows:

Description — [<start>, A, man, wears, a, hat. ]

Target — [ A man wears a hat .,<End> ]

This is done to ensure that our system is aware of both the beginning and the conclusion of the labelled series.

When developing the image's vocabulary, the first step is to cleanse the text by separating it into words, then addressing any punctuation or capitalization concerns that may arise. Because computer systems do not know English words, we symbolize them with digits and map every word of the vocabulary with a distinctive index value. Next, we encapsulate every word into a vector with a predefined length and portray every word as a number. After these steps, the text is comprehensible by the machine, and it produces the annotations for the images. In the following manner, we clean up the text to accomplish the desired size of the vocabulary:

- A string must be created from the contents of the file after the document file has been loaded and read.
- Create a description vocabulary that can map photos with all of the captions.
- Make the data more accurate by using every description as an input. When working with textual data, we undertake numerous forms of cleaning, including converting uppercase to lowercase, removing punctuation, and getting rid of words that contain numbers.
- Compile a vocabulary by using all of the one-of-a-kind terms gleaned from the various descriptions.
- Create a descriptions.txt file to keep all the captions.

## 2.2 Tokenizing Vocabulary

Keras makes available the Tokenizer class, which has the capability of learning the mapping based on the data that has been loaded.

Given the text that is entered into the photo caption, this will pass through a tokenizer. We assign a one-of-a-kind index value to every vocabulary term that we map.

A function is available within the Keras package that can be used to generate tokens based on vocabulary and then pass them on to a tokenizer.pkl pickle file.

## 2.3 Classification

A recurrent GRU can gather a significant amount of data from an image, both in terms of its temporal and spatial components. Furthermore, in to generate sentences that are accurate from a semantic standpoint, semantic information is necessary in order to improve the characteristics that are retrieved from the encoder. In our technique, the generation of semantic vectors is accomplished with the help of the Greedy technique, and the decoding process is handled by a semantic compositional network. Moreover, a dual learning approach is used to store the semantic vector's data during training, enabling the semantic data in the forward flow to be efficiently exploited.

## 2.4 Evaluation

The BLEU is a measurement that compares the quality of a produced sentence to the quality of a reference sentence. In order to evaluate the accuracy of the predictions provided by autonomous machine translation systems, the score was devised. A score of 1.0 is offered when there is a perfect match, whereas a score of 0.0 is assigned when there is a perfect mismatch.

The gathered actual and anticipated descriptions are assessed with the help of the corpus BLEU score, which provides an overall summary of how closely the created text matches the expected text.

## 2.5 Testing

Following the completion of the training phase, the model is evaluated using a variety of random inputs. Because the predictions already include the maximum length of the index values, we continue to use the previous tokenizer. pkl to retrieve the words based on the index values they have.

## 2.6 Emotic Dataset

The EMOTIC database is a large-scale annotated image database that focuses on individuals in their natural environments. People are categorized in this database according to the emotions that appear to be influencing their behavior, with a comprehensive list of 26 psychological classifications and also continual dimensions (arousal, valence and dominance). The photographs, which depict the surroundings of the individual being studied, include a diverse assortment of settings and activities. This makes it possible to expand the research of emotion recognition far beyond examination of facial expressions.

At the moment, there are 18316 images in the EMOTIC database, and 23788 people have been annotated on those photographs. The EMOTIC dataset is divided into three sets: the training set, validation set and testing set of (70%, 10% and the 20% respectively. We compiled a list of affective states as a lexicon in order to characterize the 26 different emotion types. We began the process of generating word-groupings by making use of word interconnection (such as synonyms, associations, and significance) as well as the interconnectedness of a collection of words (such as psychology studies and affective computing). The ultimate 26 categories were determined by going through a variety of iterations, consulting several dictionaries, and conducting studies in the area of emotional computing describes Figure 3.

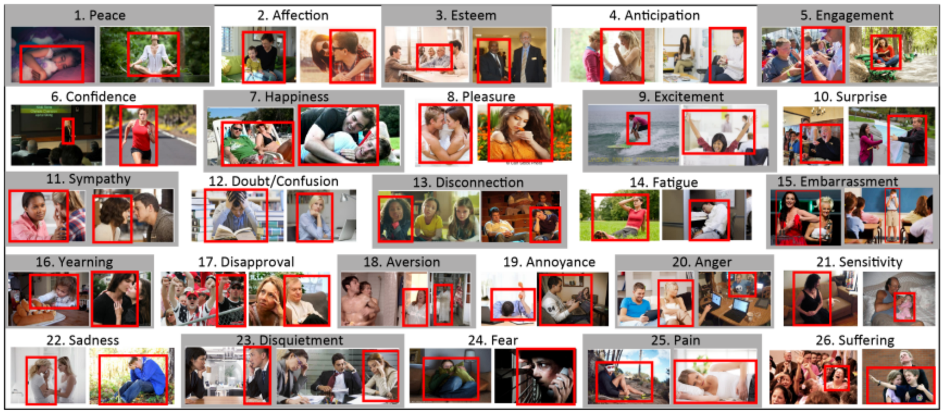


**Fig 3.** In the EMOTIC dataset, there are examples of people labeled with each ofthe 26 different emotion categories

The Figure 3 illustrates the example of individuals captioned with each of the 26 different emotion categories as mentioned in above figures.

## 3 Results and Discussion

In the experimental evaluation of model, we evaluate system in two different ways. First context extraction from image dataset using deep learning module and sentiment classification using machine learning techniques. The EMOTIC dataset has used of 6000 image dataset with 10-fold cross validation. Transfer learning can be used to implement these models. Models are trained on high-configuration GPUs that can be found online as well as those on the system we have used GTX 1650. The training period varies depending on the GPU; for such applications, systems with larger Gigabyte GPUs, such as 4-16 GB, are desirable, and software such as Python is required for the tensorflow libraries, with IDEs for python such as Jupyter Notebook or PyCharm or Visual Studio Code with libraries such as the tensor flow, numpy, keras, pandas and few others are required.

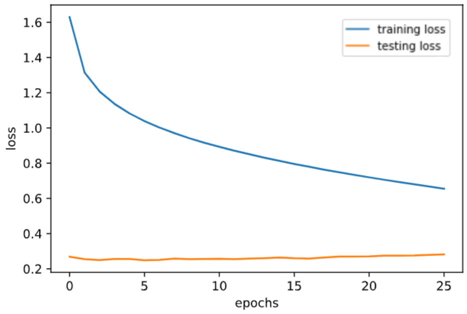
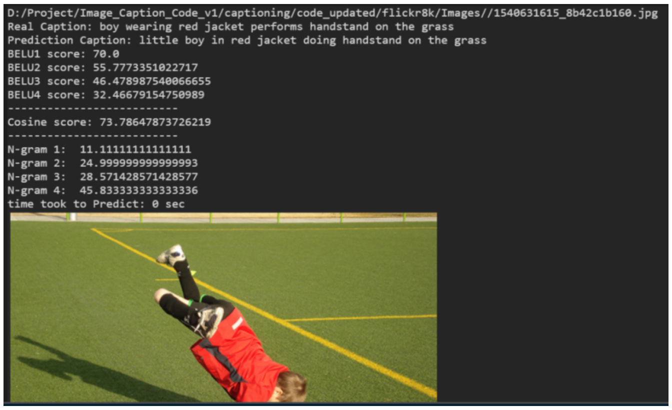


**Fig 4.** Model performance on training

The Figure 4 describes training and testing model performance with proposed EfficientNetB3 on Flickr8k dataset. With a count of 6000 pictures validated in the training and a 1000 for testing set. When system deals with high epoch values it reduces the overall loss in training while it slightly increases or constant for validation.

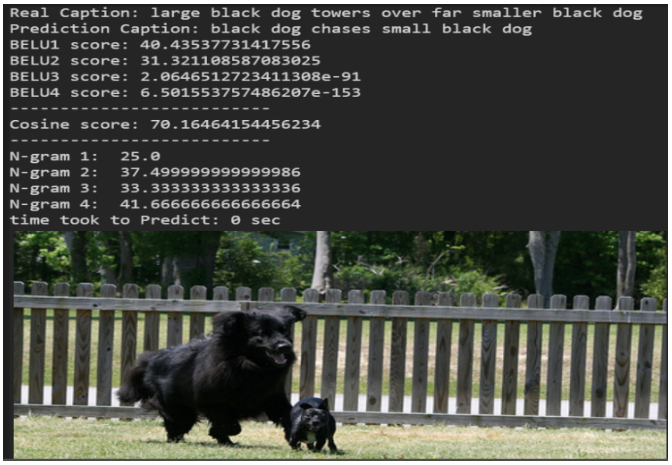


```
D:/Project/Image_Caption_Code_v1/captioning/code_updated/flickr8k/Images//1540631615_8b42c1b160.jpg
Real Caption: boy wearing red jacket performs handstand on the grass
Prediction Caption: little boy in red jacket doing handstand on the grass
BELU1 score: 70.0
BELU2 score: 55.7773351022717
BELU3 score: 46.478987540066655
BELU4 score: 32.46679154750989
-------------------------
Cosine score: 73.78647873726219
-------------------------
N-gram 1:  11.11111111111111
N-gram 2:  24.999999999999993
N-gram 3:  28.571428571428577
N-gram 4:  45.833333333333336
time took to Predict: 0 sec
```

**Fig 5.** Model performance testing with various feature selection techniques on test image 1

The Figures 5 and 6 describes the caption generation in validation phase for two different images. The generated caption has evaluated with both ground truth sentence and evaluated with 4 BELU functions and Bigrams methods with cosine similarity algorithm. According to both results we conclude the default cosine similarity produces higher accuracy over the other methods.

The Figures 7 and 8 all weighted accuracy for both objects evaluated in Figures 3 and 4. BELU-3 and N-Gram generates lower results than other techniques thus cosine similarity produces high caption generation accuracy for system. It produces around 73.80% accuracy for cosine similarity algorithm. In another experiment we have done comparative analysis with CNN-LSTM[4] and CNN-Bi-LSTM[5] as existing system for caption generation. The below Table 1 demonstrates similarly.



```
Real Caption: large black dog towers over far smaller black dog
Prediction Caption: black dog chases small black dog
BELU1 score: 40.43537731417556
BELU2 score: 31.321108587083025
BELU3 score: 2.0646512723411308e-91
BELU4 score: 6.501553757486207e-153
--------------------------
Cosine score: 70.16464154456234
--------------------------
N-gram 1:  25.0
N-gram 2:  37.499999999999986
N-gram 3:  33.333333333333336
N-gram 4:  41.666666666666664
time took to Predict: 0 sec
```

**Fig 6.** Model performance testing with various feature selection techniques on test image 2

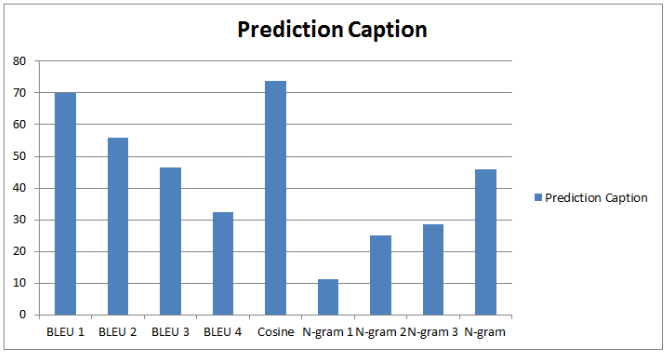


**Fig 7.** Prediction accuracy for caption generate with different techniques ontest image

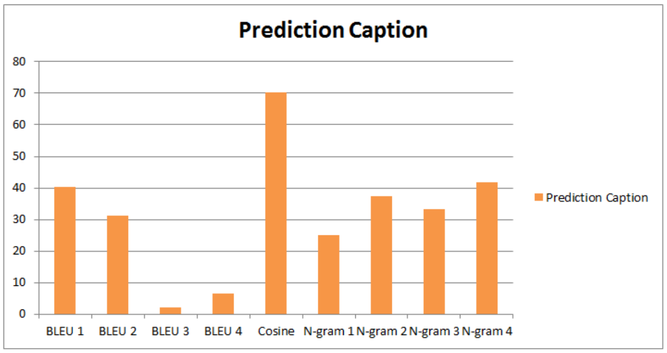


**Fig 8.** Prediction accuracy for caption generate with different techniques on test image

**Table 1.** Comparative analysis of proposed model

| Method | BLEU-1 | BLEU-2 | BLEU-2 | BLEU-2 | ROUGE-L | Cosine |
|---|---|---|---|---|---|---|
| CNN-LSTM [(4)] | 46.7 | 31.32 | 19.4 | 9 | 25.7 | NA |
| CNN-Bi-LSTM [(5)] | 55.00 | 34.9 | 24.8 | 13.1 | 27.9 | NA |
| Proposed | 70.00 | 55.77 | 46.47 | 32.46 | NA | 73.80 |

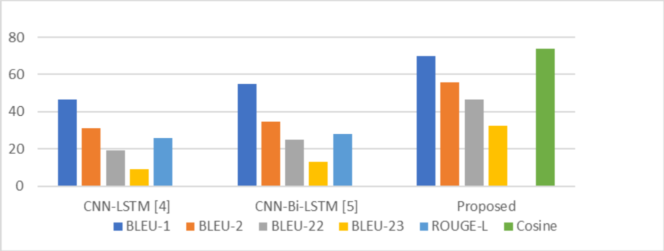


**Fig 9.** Experiment analysis and comparative analysis of system

According to this Figure 9 we conclude our system predicts superior results in terms of all performance parameters. The proposed algorithm is compared with the deep CNN[4] and CNN-Ni-LSTM[5] algorithm. We also evaluate some machine learning algorithm such as such as Navie Bayes, Random Forest & Support Vector Machine Here we conclude the performance of the system is better than compared to the existing system. The proposed system is also compared with ReLu, TanH & Sigmoid activation functions. Out of these three activations functions, the Sigmoid gives better accuracy compared to ReLu & Tanh.

## 4 Conclusion

This system describes context-aware sentiment classification using a deep learning model. Initially, we extracted the metadata and generated captions from the image using the EfficientNetB-7 model, and generated text has evaluated using the NLP model. The NLP process performs various feature extraction and selection methods, and then the classifier trains using machine learning and deep learning. Our model gives higher accuracy with cosine similarity techniques at 73.80% for the EMOTIC dataset by using EfficientNetB-7 model for text generation and NLP process for text-based sentiment classification. The pure image sentiment classification methods have been validated with a calculation of confusion matrix on similar datasets. The major limitation of this work is that it generates class ambiguity for sentiment prediction when it handles complex text generated by the EfficientNetB model. To determination of visual sentiment based on heterogeneous features such as audio, video, NLP and emotional-based visual features using deep CNN techniques will the future work of this system.

## References

1) Chandrasekaran G, Antoanela N, Andrei G, Monica C, Hemanth J. Visual Sentiment Analysis Using Deep Learning Models with Social Media Data. *Applied Sciences*. 2022;12(3):1030. Available from: https://doi.org/10.3390/app12031030.
2) Hoang MH, Kim SH, Yang HJ, Lee GS. Context-Aware Emotion Recognition Based on Visual Relationship Detection. *IEEE Access*. 2021;9:90465. Available from: https://doi.org/10.1109/ACCESS.2021.3091169.
3) Wankhade M, Rao ACS, Kulkarni C. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*. 2022;55(7):5731–5780. Available from: https://doi.org/10.1007/s10462-022-10144-1.
4) Meena G, Mohbey KK, Indian A, Kumar S. Sentiment Analysis from Images using VGG19 based Transfer Learning Approach. *Procedia Computer Science*. 2022;204:411. Available from: https://doi.org/10.1016/j.procs.2022.08.050.
5) Agughalam D, Pathak P, Stynes P. Bidirectional LSTM approach to image captioning with scene features. *Thirteenth International Conference on Digital Image Processing (ICDIP 2021)*. 2021;11878. Available from: https://doi.org/10.1117/12.2600465.
6) Chandrasekaran G, Hemanth DJ. Efficient Visual Sentiment Prediction Approaches Using Deep Learning Models. In: Knowledge Graphs and Semantic Web. Springer International Publishing. 2021;p. 260–272. Available from: https://doi.org/10.1007/978-3-030-91305-2_20.
7) Pournaras A, Gkalelis N, Galanopoulos D, Mezaris V. Exploiting Out-of-Domain Datasets and Visual Representations for Image Sentiment Classification. In: 2021 16th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP). IEEE. 2021;p. 1–6. Available from: https://doi.org/10.1109/SMAP53521.2021.9610801.
8) Al-Halah Z, Aitken AP, Shi W, Caballero J, Smile. Be Happy :) Emoji Embedding for Visual Sentiment Analysis. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). 2019;p. 4491–4500. Available from: https://doi.org/10.48550/arXiv.1907.06160.
9) Huang F, Wei K, Weng J, Li Z. Attention-Based Modality-Gated Networks for Image-Text Sentiment Analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications*. 2020;16(3):1–19. Available from: https://doi.org/10.1145/3388861.
10) Gelli F, Uricchio T, He X, Bimbo AD, Chua TSS. Learning Subjective Attributes of Images from Auxiliary Sources. In: Proceedings of the 27th ACM International Conference on Multimedia. ACM. 2019;p. 2263–2271. Available from: https://doi.org/10.1145/3343031.3350574.
11) Ou H, Qing C, Xu X, Jin J. Multi-Level Context Pyramid Network for Visual Sentiment Analysis. *Sensors*. 2021;21(6):2136. Available from: https://doi.org/10.3390/s21062136.
12) Zhao T, Hu Y, Valsdottir LR, Zang T, Peng J. Identifying drug–target interactions based on graph convolutional network and deep neural network. *Briefings in Bioinformatics*. 2021;22(2):2141–2150. Available from: https://doi.org/10.1093/bib/bbaa044.
13) Masood MA, Abbasi RA, Keong NW. Context-Aware Sliding Window for Sentiment Classification. *IEEE Access*. 2020;8:4870–4884. Available from: https://doi.org/10.1109/ACCESS.2019.2963586.

14) Shilpa PC, Shereen R, Jacob S, Vinod P. Sentiment Analysis Using Deep Learning. *Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. 2021. Available from: https://doi.org/10.1109/ICICV50876.2021.9388382.

15) Singh C, Imam T, Wibowo S, Grandhi S. A Deep Learning Approach for Sentiment Analysis of COVID-19 Reviews. *Applied Sciences*. 2022;12(8):3709. Available from: https://doi.org/10.3390/app12083709.