International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-11 Issue-3, February 2022

Multimodal Offensive Meme Classification Using Transformers and BiLSTM

Roshan Nayak, B S Ullas Kannantha, Kruthi S, C. Gururaj

Abstract*: Nowadays memes have become a way in which people express their ideas on social media. These memes can convey various views including offensive ones. Memes can be intended for a personal attack, homophobic abuse, racial abuse, attack on minority etc. The memes are implicit and multi-modal in nature. Here we analyze the meme by categorizing them as offensive or not offensive and this becomes a binary classification problem. We propose a novel offensive meme classification using the transformer-based image encoder, BiLSTM for text with mean pooling as text encoder and a Feed-Forward Network as a classification head. The SwinT + BiLSTM has performed better when compared to the ViT + BiLSTM across all the dimensions. The performance of the models has improved significantly when the contextual embeddings from DistilBert replace the custom embeddings. We have achieved the highest recall of 0.631 by combining outputs of four models using the soft voting technique.

Keywords: Offensive Meme Classification; BilSTM; Transformer; Pooling; Confusion matrix;

I. INTRODUCTION

In recent times, the ease of access to the internet has fueled the copious amounts of social media users. People are more vocal and want their voices to be heard by a multitude of people. A social media post can have a huge impact on society [1] as well as on the individual [2]. There are cases where a social media post has caused riots and hate crimes in society. On an individual front, a social media meme can lead to depression and even suicide. In recent times people are using memes as a way to communicate their views on social media. A meme [3] can be an image or a video or just text which is commonly humorous by nature. It is shared rapidly among internet users, often without any changes. While social media is a vast platform for communication, memes are ways of representing the thoughts and feelings of a specific audience. If such a meme is highly irritating, angering or causing displeasure to other people, it is termed offensive.

Manuscript received on January 31, 2022.
Revised Manuscript received on February 05, 2022.
Manuscript published on February 28, 2022.
* Correspondence Author
Roshan Nayak*, Department of Electronics and Communication Engineering, B.M.S. College of Engineering, Bengaluru, India. Email: roshan.ec18@bmsce.ac.in
B S Ullas Kannantha, Department of Electronics and Instrumentation Engineering, B.M.S. College of Engineering, Bengaluru, India. Email: ksullas.ei18@bmsce.ac.in
Kruthi S, Department of Electronics and Instrumentation Engineering, B.M.S. College of Engineering, Bengaluru, India. Email: kruthis.ei18@bmsce.ac.in
C. Gururaj, Senior Member IEEE, Department of Electronics and Telecommunication, B.M.S. College of Engineering, Bengaluru, India. Email: gururaj.tce@bmsce.ac.in

Memes are offensive when they share pointed commentary on cultural symbols, current affairs, social ideas, and intend personal attack, homophobic abuse [4], racial abuse [5], attack on minorities etc. Memes have become a way to communicate political hate as well [6]. Clearly, there is a need for a filter before memes reach millions of people. One way to find out if a meme is offensive is to manually check each meme that is posted (or in the process of being posted) on social media, but verifying each meme manually is an impossible task. Therefore, effective automatic filtering of memes is required to avoid the spread of hateful content.

Memes are often images with a piece of text on them. In other words, memes are multimodal in nature and it is unpredictable whether the offensive content is associated with the embedded text or with the background image. They cannot be classified as offensive based on results obtained from a single modality. When we consider the image or text alone, we may not be able to perceive anything harmful, but when both the modalities are considered, it changes the context. Therefore, this project aims to construct a multimodal system that can process both image and text on the meme parallelly, to classify them as offensive or non-offensive with higher accuracy.

An Artificial Intelligence (AI) is a system that tries to mimic the human brain and behaviour [7]. Machine Learning is a concept in AI in which a machine can learn to do a task through training. Modelling the biological neuron [8] led to a computation system called Artificial Neural Networks (ANN). Deep Learning is made up of several layers of ANN, which is employed to solve complex tasks. The topics under AI like Computer vision and NLP are making great strides that no one could have predicted. Today's AI can detect faces [9], enhance cutomer experience [10], find gender or communal bias content in a sentence [11], retrieve [12] images accurately from the database [13], etc.

Image classification can be done using Decision trees [14], Support Vector Machines [15], Convolution Neural Networks (CNN)[16], etc. A new concept called attention mechanism is used to learn the importance of each part in the input sequence. Transformers are state of the art models that use attention mechanisms [17] to learn the importance of each part in the input sequence. Transformers is largely used in NLP however; Vision transformers can attain great results in image classification [18] tasks compared to CNN [19]. We have used several preprocessing techniques [20] and attention-based extract features from the image. These image features [21], along with the text features, are then used to classify the meme as offensive or not.

Text processing can be done using the NLP (Natural Language Processing) techniques. NLP is a field in which the analysis and understanding of the text are done. Sentiment analysis, question and answering, article summarization and many more can be done using NLP techniques. Lately, there has been a huge advancement in this field. One of them was the introduction of Attention in the "Attention is all you need"[17] which bought NLP to a whole new level. Several state-of-the-art models are based on this mechanism. Coming to the NLP part in this paper we have used several preprocessing techniques and attention-based models to clean and extract features from the text. These text features, along with the image features, are then used to classify the meme as offensive or not.

In section I, we gave an introduction to our work. We discuss the previous related work in section II. We discuss the data and its source in section III. Section IV explains the methods we have used to classify the memes as offensive or not. In section V, we explain the model architecture that we have used. Results are discussed in section VI, and finally, in section VII, we conclude our work along with the future scope.

II. RELATED WORK

A technique of early fusion is adopted to merge text with image modalities in order to compare them along with baseline models of image and text to look over its efficiency [22]. Here the author has shown the capability of retrieving offensive memes by using CNN (Convolutional Neural Network) on the text and a text classifier based on the Stacked LSTM abbreviated as Long Short-Term Memory, BiLSTM abbreviated as Bidirectional Long Short-Term Memory. The CNN based approach has better performance in terms of recall when text and image features are considered. Apeksha Aggarwal et al. [23] proposed two ways to classify hateful memes. The first one is the sequential approach which uses sentiment analysis, in which the image is converted to text based on sentiment and concatenated with image caption. The other approach is the multimodal approach which combines features from different modalities, to extract information from memes. The next one is solving the problem of the memes classification with a system using NLP (Natural Language Processing) and computer vision techniques [24]. Here the author used late fusion to classify text as well as the image before fusing the results

Douwe Kiela, Hamed Firooz et al. [25] show a way of identifying hate speech present in multimodal memes. Their results show that advanced fusion leads to better performance, with early fusion models outperforming middle and late fusion approaches. Riza Velioglu et al. [26] used pre-trained VisualBERT and applied majority voting on 27 best models to detect hate speech. The authors showed that majority voting boosts accuracy by 2.5%. The text-based classification methods to distinguish between hate speech and profanity [27]. The authors designed a system using word n-grams, word skip-grams and character n-grams, and showed that the best result was obtained by a character 4-gram model with an accuracy of 78%.

III. DATASET

We use the Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text [22]. Each type of data file has three columns. The name of the image with the file extension is given in the first column.

Table. 1. Summary of dataset

Data	Number of offensive data	Number of non-offensive data	Total
Training	187	258	445
Validation	58	91	149
Test	58	91	149

In the second column meme text is in English and finally, the third column is for annotation ("offensive" or "non-offensive"). The data is quite imbalanced with more offensive data than non-offensive. The data is divided into training data constituting 60%, testing data constituting 20% and the remaining 20% is validation data

IV. METHODOLOGY

In this section, we will give a brief explanation of the methods we have used to classify the memes as offensive or not. In the data preprocessing subsection, we make a walkthrough, through the steps we have employed. Then the subsequent sections will explain in brief the techniques we have used and the architectures we have tried on.

A. Data Preprocessing

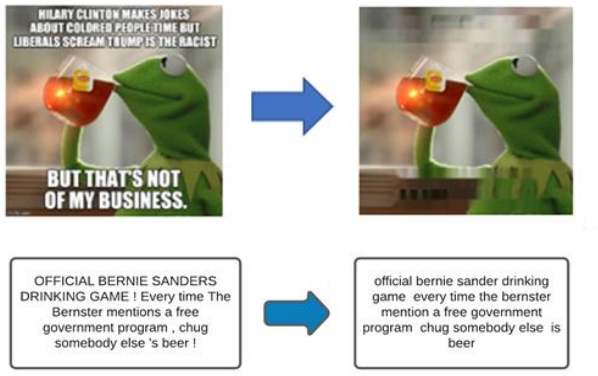

Figure 1: Preprocessed data

We have already seen in section III that the dataset being used for the purpose of the analysis is a multi-modal dataset consisting of image and text pairs. The images had text embedded on them as they do not help in extracting the features from the image. Data augmentation techniques like Flip and Blur were applied to the image while training. We first converted the meme captions to lowercase, decontracted the words (eg. don't -> do not), removed special characters, and then converted the words to their base form using the lemmatization technique. We can see the result of preprocessed image and text in Figure 1.

B. Word Embeddings

Computers cannot process data in the form of text. Hence words have to be represented in the form of numbers. In natural language processing, word embeddings are representations of the words as a vector of numbers. We have tried out

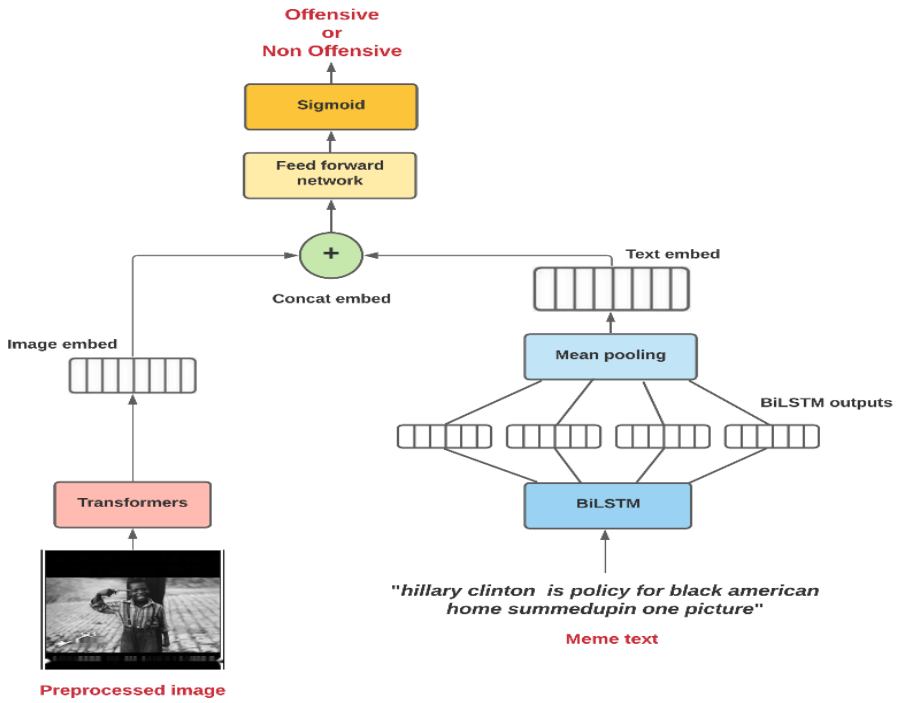

Figure 2: Proposed model architecture

Table II: Validation results

Architecture	Embeddings	Precision	Recall	F1 Score
SwinT + Stacked BiLSTM	Custom	0.573	0.595	0.575
ViT + Stacked BiLSTM	Custom	0.625	0.642	0.605
SwinT + Stacked BiLSTM	DistilBert	0.540	0.547	0.543
ViT + Stacked BiLSTM	DistilBert	0.575	0.554	0.560

two ways of embedding the words. In the first method, our model learns the embeddings during the training process. The embedding layer in PyTorch randomly initializes the embeddings of a word and then updates them during backpropagation. As these embeddings are learned during the training process, they are called task-specific embeddings. In a second way, we have used DistilBert [28]. DistilBert has an upper hand over some of the other models including Glove and Word2Vec because there is fixed representation in Word2Vec and Glove for each word, independent of the word's context, whereas in DistilBert, word embeddings are produced that are context-dependent and they are informed by other words which are present around them.

C. Visual Transformer

The architecture of ViT (Vision Transformer) can be compared with the transformer used for Natural Language Processing applications. Since the self-attention used in NLP tasks would be computationally inefficient for images due to a large number of pixels, the Global Attention technique was employed. Instead of a sequence of 1D token embeddings, ViT accepts flattened 2D patches in sequence as input. Considering (H, W) to be the original image's resolution, C to be the total channel number, and (P, P) to be each image patch's resolution, we have N = HW/P to be the Transformer's effective input sequence length [18]. Before passing, the flattened patches into the transformer they multiplied with an embedding matrix to get patched embeddings and then the positional embeddings are added as well. This eventually goes into the transformer. The transformer consists of an encoder block which consists of alternating layers of multi-headed self-attention blocks in order to understand the attention to be given for specific regions of the images. Layer Normalization and residual connection are applied before and after every block respectively.

D. Swin Transformer

The SwinT (Swin Transformer) is a type of Vision Transformer. SwinT is the first transformer-based backbone architecture for visual tasks. The contribution of the paper is that the authors are addressing a couple of problems that are faced while employing Transformers for visual tasks, like the large variations in the scale of objects in the image and also the pixel count in images compared to words in the text [29]. Calculating self-attention with every other vector is computationally inefficient. Hence the authors have proposed a

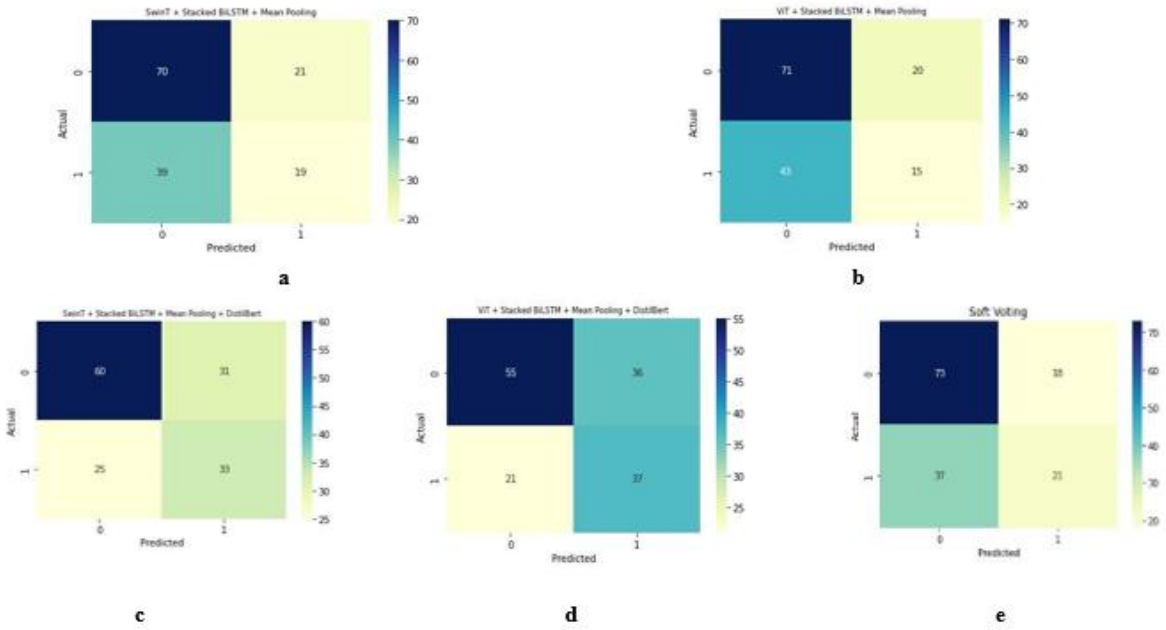

Figure 3: Confusion matrix of (a) **SwinT + Stacked BiLSTM** (b) **ViT + Stacked BiLSTM** (c) **SwinT + Stacked BiLSTM with DistilBERT Embeddings** (d) **ViT + Stacked BiLSTM with DistilBERT Embeddings** (e) **Soft Voting on all four models**

Table III: Result of test data

Architecture	Embeddings	Precision	Recall	F1 Score
SwinT + Stacked BiLSTM	Custom	0.577	0.597	0.578
ViT + Stacked BiLSTM	Custom	0.547	0.577	0.549
SwinT + Stacked BiLSTM	DistilBert	0.632	0.624	0.627
ViT + Stacked BiLSTM	DistilBert	0.639	0.617	0.622
Ensemble (Soft Voting)	—	0.615	0.631	0.613

hierarchical transformer with a shifted window mechanism for calculating self-attention. The hierarchical structure will enhance the flexibility of the model. Shifted window mechanism will limit the computation of self-attention to local windows and also allow for cross-window connection.

E. Bidirectional LSTM

LSTMs (Long Short Term Memory) are sequence models that are good at retaining information for a longer period of time. But since LSTMs learn only the forward sequence, the dependency of the current word on the future words in the sequence is not understood. Hence BiLSTMs (Bidirectional LSTM) [30] are used. In BiLSTM, two models are trained instead of just one. The first model learns the input in the sequence in which it is provided, and the second model learns the input in the reverse sequence. The outputs of the corresponding cell are then concatenated to generate the final output. We have used two BiLSTMs in our architecture placed on top of each other hence called Stacked BiLSTM.

F. Mean Pooling

Mean Pooling is a technique used to aggregate the outputs of the BiLSTM. The pooling layer converts variable-length hidden vector sequences into a single vector. In a BiLSTM network, each cell outputs a vector. As we have used dynamic LSTM the number of cells varies across each batch. Hence in the mean-pooling method, the outputs from the hidden state of the stacked BiLSTM are averaged on each vector [31], to obtain a single vector which is used as the text features.

Let v_1,v_2,...,v_n represent the output vectors of the hidden state of BiLSTM. The text features vector, the output is represented as,

$$output = \frac{1}{n}\sum_{i=1}^{n} v_i$$

Table IV: Predictions on Test Samples

Meme			
Text on meme	THIS IS A WALL INSIDE A NAZI GAS CHAMBER WHEN A DEMOCRAT MENTIONS GUN CONTROL REMEMBER THIS PICTURE	Donald Trump 's hair looks like someone tried to use the blur tool in photoshop to cover a bald spot	HEY GIRL EVERY TIME SOMEONE DECIDES TO VOTE FOR BERNIE SANDERS I SMILE
Correct labels	Offensive	Offensive	Non offensive
SwinT + BiLSTM	Offensive	Non offensive	Non offensive
ViT + BiLSTM	Non Offensive	Offensive	Non offensive
SwinT + BiLSTM (DistilBERT)	Offensive	Offensive	Non offensive
ViT + BiLSTM (DistilBERT)	Offensive	Offensive	Offensive
Soft Voting	Offensive	Offensive	Non offensive

G. Soft Voting

In soft voting, we take the average of the probabilities of each class obtained from multiple models to decide on the final Prediction [32]. A class is then considered to be the predicted class if it has the highest average probability. This is better when compared with the hard voting method because it considers the confidence level of each model. In Fig. 4 the predicted probabilities from the four models are passed on to the soft voting layer to make the final decision.

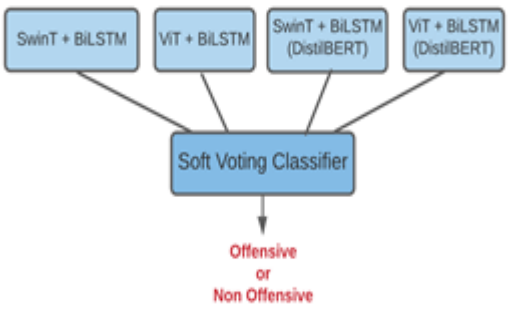

Figure 4: Soft voting

V. MODEL ARCHITECTURE

In this section, we will be proposing a generic architecture that we have used to predict if the image-text pair is offensive or not. Fig. 2 shows the generic architecture of our models. The texts and images are first preprocessed as mentioned in section VI A. Then the text is tokenized and converted to its corresponding word embeddings. Now the text and images are ready to be processed by the models. The stacked BiLSTM network accepts the text as its input. Mean pooling is applied to the hidden state outputs of the final BiLSTM in order to get the text embedding vectors of size 128. Images are transformed randomly using the argumentation library and then fed into the transformer-based models to get image embedding vectors of size 128. Finally, the image and text embedding vectors are concatenated and the resulting 256 sized vectors are fed into the classification head which makes use of two FC layers to generate the output. Sigmoid activation is applied to this output to get the probability. Four different models were trained to keep the proposed architecture common across all the models. The outputs of the models were then aggregated using the Soft Voting technique to generate the final output.

VI. RESULTS

During our analysis, we have used transformer-based models for images and BiLSTM for the text. The embeddings were either learnt during the training process from scratch or the contextual embeddings from the DistilBert model. The dataset split is shown in Table 1. Due to the imbalanced nature of the dataset, we have used weighted Recall, Precision, and F1-Score as parameters to evaluate model performance. That is the weighted average of the metrics of each class. The weight is calculated based on the number of true labels for each class. Every model was trained for 10 epochs on Google Colab.

Validation loss was the parameter being monitored to decide if the training should continue or stop. The validation results are shown in Table 2. ViT + Stacked BiLSTM with custom embeddings achieved the highest precision of 0.625, ViT + Stacked BiLSTM with Custom embeddings achieved the highest recall of 0.642, ViT + Stacked BiLSTM with Custom embeddings achieved the highest f1score of 0.605. Precision, Recall, and F1score range from 0.573 to 0.575, 0.547 to 0.642, and 0.543 to 0.605 respectively.

Our models were tested on the test set of 149 images. The results we obtained on the test data are shown in Table 3. Figure 3 shows the confusion matrix of each model, where 1 represents Offensive and 0 represents Non Offensive. From the table, it is evident that the SwinT + BiLSTM has performed better when compared to the ViT + BiLSTM across all the dimensions. It's also evident that the performance of the models has improved significantly when the contextual embeddings from DistilBert replace the custom embeddings. ViT + Stacked BiLSTM with DistilBert embeddings achieved the highest precision of 0.639, Soft Voting ensemble technique achieved the highest recall of 0.631, SwinT + Stacked BiLSTM with DistilBert embeddings achieved the highest f1score of 0.627. Precision, Recall, and F1score range from 0.547 to 0.639, 0.577 to 0.631, and 0.549 to 0.627 respectively.

Table 4 gives a sense of how the different models performed on the test samples. We see that on the first meme Vit + BiLSTM model with custom embeddings failed to predict the true label whereas all the other models could predict it correctly. Similarly, SwinT + BiLSTM with custom embeddings failed on the second meme and Vit + BiLSTM with DistilBert embeddings failed on the third meme. We see that Soft Voting is more accurate when compared to individual models. It's evident from the highest recall of 0.631.

VII. CONCLUSION

The proposed models perform better on the non-offensive class when compared to the offensive glass because of the imbalance in the dataset. The dataset used during the analysis is very small with just 445 images in the training set. Augmentation of the data did not improve the performance. The models were overfitting the training dataset. Usage of some regularizing techniques like dropout and early stopping did help to some extent. The soft voting of the different models trained was able to get the best-weighted recall score of 0.631 and SwinT + Stacked BiLSTM could get the best F1-score of 0.627 on the test dataset. In this paper, we have proposed a transformer-based image encoder, BiLSTM for text with mean pooling as text encoder and a Feed-Forward Network as a classification head. In the future, we would like to try out transformer-based models to encode text as well. Also, propose a more complex architecture that would make use of cross attention and region proposals [33] to understand more deeply how the text and image are related.

REFERENCES

1. W. Akram, R. Kumar, “A Study on Positive and Negative Effects of Social Media on Society,” International Journal of Computer Sciences and Engineering, Vol.5, Issue.10, pp.351-354, 2017.
2. Ostic D, Qalati SA, Barbosa B, Shah SMM, Galvan Vela E, Herzallah AM and Liu F (2021) “Effects of Social Media Use on Psychological Well-Being”: A Mediated Model. Front. Psychol. 12:678766. doi: https://doi.org/10.3389/fpsyg.2021.678766
3. Limor Shifman, Memes in a Digital World: “Memes in a Digital World: Reconciling with a Conceptual Troublemaker: Reconciling with a Conceptual Troublemaker”, Journal of Computer-Mediated Communication, Volume 18, Issue 3, 1 April 2013, Pages 362–377, https://doi.org/10.1111/jcc4.12013
4. Powell A, Scott AJ, Henry N. “Digital harassment and abuse: Experiences of sexuality and gender minority adults”. European Journal of Criminology. 2020;17(2):199-223. doi: https://doi.org/10.1177/1477370818788006
5. Matamoros-Fernández A, Farkas J. “Racism, Hate Speech, and Social Media: A Systematic Review and Critique”. Television & New Media. 2021;22(2):205-224. doi: https://doi.org/10.1177/1527476420982230
6. Anushka Kulkarn. Internet meme and Political Discourse: “A study on the impact of internet meme as a tool in communicating political satire”. Journal of Content, Community & Communication. Vol. 6 Year 3, June - 2017 [ISSN: 2395-7514]
7. Wang, Pei. "On Defining Artificial Intelligence" Journal of Artificial General Intelligence, vol.10, no.2, 2019, pp.1-37. https://doi.org/10.2478/jagi-2019-0002
8. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408. doi: https://doi.org/10.1037/h0042519
9. Maneesha K, Neha Shree, Pranav Datta R, Sindhu S K, C.Gururaj, “Real Time Face Detection Robot”, 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT-2017), ISBN: 978-1-5090-3704-9,19th – 20th May 2017 , pp 58-64, SVCE, Bengaluru
10. S. Bharath, R. H. Praneeth, M. Tejesh and C. Gururaj, "Customer Experience Enhancement Using Artificial Intelligence," 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), 2019, pp. 305-309, doi: https://doi.org/10.1109/ICATIECE45860.2019.9063847.
11. Benhur, Sean, Roshan Nayak, Kanchana Sivanraju, Adeep Hande, Subalalitha Chinnaudayar Navaneethakrishnan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. "Hypers at ComMA@ ICON: Modelling Aggressiveness, Gender Bias and Communal Bias Identification." arXiv preprint arXiv:2112.15417 (2021).
12. C Gururaj, Satish Tunga, “AI based Feature Extraction through Content Based Image Retrieval”, Journal of Computational and Theoretical Nanoscience, February 2020, volume 17, Issue 9-10, pp. 4097-4101, ISSN: 1546-1955 (Print): EISSN: 1546-1963 (Online), DOI: https://doi.org/10.1166/jctn.2020.9018
13. C Gururaj, D Jayadevappa, Satish Tunga, “Content Based Image Retrieval System Implementation through Neural Network”, IOSR Journal of VLSI and Signal Processing (IOSR-JVSP), Volume 6, Issue 3, Ver. 3 (May – June 2016), e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197, pp 42 – 47
14. C. Agarwal and A. Sharma, "Image understanding using decision tree based machine learning," ICIMU 2011 : Proceedings of the 5th international Conference on Information Technology & Multimedia, 2011, pp. 1-8, doi: https://doi.org/10.1109/ICIMU.2011.6122757.
15. Jianxin Wu, (2012). Efficient HIK SVM Learning for Image Classification. IEEE Transactions on Image Processing, 21(10), 4442–4453. doi: https://doi.org/10.1109/TIP.2012.2207392
16. F. Sultana, A. Sufian and P. Dutta, "Advancements in Image Classification using Convolutional Neural Network," 2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2018, pp. 122-129, doi: https://doi.org/10.1109/ICRCICN.2018.8718718 .
17. Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017). "Attention Is All You Need".
18. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et.al “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.“ ICLR 2021

19. Veena Nayak, Sushma P.Holla, AkshayaKumar K. M., C. Gururaj, "Automatic number plate recognition", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 9, No. 3, pp 3783 – 3787, ISSN 2278-3091, May – June 2020
20. Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. 2000. "Image inpainting". In Proceedings of the 27th annual conference on Computer graphics and interactive techniques (SIGGRAPH '00). ACM Press/Addison-Wesley Publishing Co., USA, 417–424. DOI: https://doi.org/10.1145/344779.344972
21. Dong ping Tianj, "A Review on Image Feature Extraction and Representation Techniques", International Journal of Multimedia and Ubiquitous Engineering Vol. 8, No. 4, July, 2013.
22. Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar, "Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text," European Language Resources Association (ELRA), May, 2020, pp. 32-41.
23. Apeksha Aggarwal, Vibhav Sharma, Anshul Trivedi, and Mayank Yadav, "Two-way feature extraction using sequential and multimodal approach for hateful meme classification," Complexity, vol. 2021, 19 April, 2021, doi: https://doi.org/10.1155/2021/5510253.
24. Rachana Jadhav and Prof. Vikas N. Honmane, "Memes classification system using computer vision and NLP techniques," International Journal of Engineering, Applied Sciences and Technology, vol. 6, Issue 02, June 2021, pp. 164-168, doi: https://doi.org/10.33564/ijeast.2021.v06i02.025
25. Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," 7 April, 2021, arXiv:2005.04790, 2020.
26. Velioglu Riza and Jewgeni Rose, "Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge"2020, arXiv:2012.12975
27. Shervin Malmasi, and MarcosZampieri, "Detecting hate speech in social- media," Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pp. 467-472, doi: https://doi.org/10.26615/978-954-452-049-6_062
28. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". ArXiv, abs/1910.01108.
29. Liu, Ze., Lin, Yutong., Cao, Yue., Hu Han., Wei, Yixuan., Zhang, Zheng., Lin, Stephen., Guo, Baining. 2021. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows".
30. Islam, Md Saiful. (2018). "A Deep Recurrent Neural Network with BiLSTM model for Sentiment Classification."
31. Ganda, Raouf & Mahmood, Ausif. (2017). "Deep learning for sentence classification". 1-5. https://doi.org/10.1109/LISAT.2017.8001979.
32. Islam, Riadul & Shahjalal, Md. (2019). "Soft Voting-Based Ensemble Approach to Predict Early Stage DRC Violations". 1081-1084. https://doi.org/10.1109/MWSCAS.2019.8884896.
33. S. D. Bhattacharjee, W. J. Tolone, R. Cheria and U. Sarka, "Multimodal, Context-Aware, Feature Representation Learning for Classification and Localization," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 1034-1039, doi: https://doi.org/10.1109/BigData47090.2019.9005499.

AUTHORS PROFILE

Roshan Nayak is currently a final year undergraduate student in the Electronics and Communication department at BMS College of Engineering, Bangalore, Karnataka, India. His research work mostly lies in the field of Computer Vision and Natural Language Processing, applying deep learning techniques for social media analysis. His research has been accepted in a National Conference. He has worked in a couple of internships and several intermediate-level projects.

B S Ullas Kannantha is currently a final year undergraduate student in the Electronics and Instrumentation Engineering department at BMS College of Engineering, Bangalore, Karnataka, India. His research work is in the field of Computer Vision, and he is currently working on a research project in Neuroscience. He is a recipient of the Summer Research Fellowship from the Indian Academy of Sciences.

Kruthi S is currently a final year undergraduate student in the department of Electronics and Instrumentation Engineering at BMS College of Engineering, Bangalore, Karnataka, India. She has worked on a few projects relating to Computer Vision including object and motion detection. She has carried out several other hobby projects and is concurrently working on an R&D internship project.

C. Gururaj received his B.E. degree in ECE and MTech in Electronics from Visvesvaraya Technological University, Belagavi his PhD from Jain University, Bengaluru. He is currently working in the department of Electronics and Telecommunication Engineering, BMS College of Engineering, Bengaluru. He has more than 50 publications to his credit with high citations that are indexed in portals such as Scopus, Web of Science, Google scholar, Vidwan etc. He has received multiple awards and grants throughout his 18 years career. His areas of interest are Image Processing, VLSI Design, Machine Learning, Deep Learning, Artificial Intelligence and Engineering Education