Hindawi
Complexity
Volume 2021, Article ID 5510253, 7 pages
https://doi.org/10.1155/2021/5510253

Research Article

Two-Way Feature Extraction Using Sequential and Multimodal Approach for Hateful Meme Classification

Apeksha Aggarwal,[1] Vibhav Sharma,[1] Anshul Trivedi,[1] Mayank Yadav,[1] Chirag Agrawal,[1] Dilbag Singh,[1] Vipul Mishra,[1] and Hassène Gritli[2,3]

[1]*Department of Computer Science and Engineering, Bennett University, Greater Noida 201310, India*
[2]*Higher Institute of Information and Communication Technologies, University of Carthage, Tunis, Tunisia*
[3]*RISC Lab (LR16ES07), National Engineering School of Tunis, University of Tunis El Manar, Tunis, Tunisia*

Correspondence should be addressed to Hassène Gritli; grhass@yahoo.fr

Received 13 February 2021; Revised 6 April 2021; Accepted 8 April 2021; Published 19 April 2021

Academic Editor: M. Irfan Uddin

Millions of memes are created and shared every day on social media platforms. Memes are a great tool to spread humour. However, some people use it to target an individual or a group generating offensive content in a polite and sarcastic way. Lack of moderation of such memes spreads hatred and can lead to depression like psychological conditions. Many successful studies related to analysis of language such as sentiment analysis and analysis of images such as image classification have been performed. However, most of these studies rely only upon either one of these components. As classifying meme is one problem which cannot be solved by relying upon only any one of these aspects, the present work identifies, addresses, and ensembles both the aspects for analyzing such data. In this research, we propose a solution to the problems in which the classification depends on more than one model. This paper proposes two different approaches to solve the problem of identifying hate memes. The first approach uses sentiment analysis based on image captioning and text written on the meme. The second approach is to combine features from different modalities. These approaches utilize a combination of glove, encoder-decoder, and OCR with Adamax optimizer deep learning algorithms. Facebook Challenge Hateful Meme Dataset is utilized which contains approximately 8500 meme images. Both the approaches are implemented on the live challenge competition by Facebook and predicted quite acceptable results. Both approaches are tested on the validation dataset, and results are found to be promising for both models.

1. Introduction

In the present era, social media is the most important activity that directly or indirectly affects people [1]. Although social media is a great platform to masses for developing skills, reach to experts, and for expressing talent, this platform has helped many people to gain success by sharing and escalating their work around the globe with the Internet. Sharing of memes on social media is increasing rapidly. Memes spread humour on a positive side. However, technology comes with a boon and a bane. These memes on the negative side can hurt any group or an individual. Internet memes can most commonly be defined as still images with text that spread rapidly among people and become a craze. They attempt to make us laugh at the expense of a theme or a person. They often carry a deeper meaning. Memes can be made by anyone. A section of audience may find them funny while another section may find them offensive. Memes are widely spread in social media sites such as Quora, Instagram, Twitter, Facebook, Snapchat, and WhatsApp. Memes are a great tool to spread humour; however, some people use it to target an individual or a group and to offend them in a polite and sarcastic way. Such memes spread hatred, and their excess may lead to depression. Nowadays, memes are made on countless topics like politics, movies, games, college life, and comic book characters.

In this work, we are addressing a real-world problem by using multiple techniques of deep learning. Most research is targeting a particular domain, viz, text recognition [2], image classification [3], object detection [4], and natural language processing [5]. As the problem has a direct impact on the society, we are trying to manage odds here and are thus trying to provide the best possible solution by combining the aforementioned techniques so that it is beneficial for society. Although a vast amount of research is related to sentiment analysis [6, 7], combining it with image [3] makes the problem itself novel. Memes can be a great content to have a laugh or two. However, a content which is hilarious to one can also be a loathe to another. Some people also deliberately create a meme whose salient purpose is to spread hate towards a community or a person. Since the reach of content in social media is limited to no one [1], it has also gained the attention of political parties to promote their agenda with the help of memes. Some political parties use memes to spread fallacious information to people about their oppositions, which indirectly affect the elections. However, a lot of people are sharing distasteful memes and encouraging their ideas on social media sites. Such memes try to make fun of a target individual or group. Ideas and statements of such memes should be banned before it is too late. A lot of people read such memes and may accept that idea is acceptable. Data analysts from different parts of the world are trying to solve the problem of identifying such memes. Millions of memes are created and shared every day on social media platforms. It is not possible to remove hateful memes manually. In this research, we propose an algorithm that identifies such memes so that social media platforms like Facebook or Quora could delete such memes.

Our research contributions can be summarized as follows:

(i) A two-way analysis is proposed covering textual as well as image component of the memes.

(ii) Data cleaning, preprocessing, transformation, and text extraction from images are performed over the dataset to improve the generalization.

(iii) Instead of focusing on one domain, ensemble of techniques from multiple domains is proposed.

(iv) Feature extraction over the dataset is performed considering textual component as well as image component features from the memes.

(v) Two novel sequential and multimodal approaches are proposed, and we are successfully able to carry out comparative analysis on both. In sequential approach, image modality is converted to text modality using image captioning and then classification is done using textual features. In multimodal approach, image features and textual features are extracted and combined to classify memes.

(vi) Results show that the proposed approach outperforms the ground truth remarkably.

This work is organized as follows. Section 2 presents an extensive literature survey of DL techniques for hate meme classification problem. In Section 3, we have proposed our two-way model for classifying memes. Section 4 presents, in-detailed comparative analysis, and discussions over the results. Finally, the work is concluded with Section 5.

2. Related Works

Image captioning is one of the ongoing research field. It is very difficult to extract context of a particular image by just looking at it. You et al. [8] proposed a solution to deal with such problems using encoder and decoder architectures of DL.

Anderson et al. [9] proposed a similar method to handle image captioning problem using concatenation of both top-down and bottom-up attention mechanism. This enabled user to calculate the salient image portions. They have used faster R-CNN on image portions each attached to its feature vector which helps them to determine the appropriate weights of the features required as per the architecture demand.

Optical character recognition is one of the most studied fields in AI and DL. Many researchers have performed various model architectures but none of them can be generalized as it all depends on the dataset they are utilizing and helps us get a broad understanding on how we can tackle similar problems. In [10], authors tried to explain one of the approaches to deal with such problems in efficient way. They have explained in detail about how they have taken pretrained weights of Google_Incpetion_V3. The model was trained on some random 54k + noisy char images which helped them in overall 21.5% reduction of error rate compared with the existing OCR's model.

Similarly, authors in [11] used a more rigorous approach rather than relying only on the pretrained model. They developed their custom CNN model by fine-tuning the pretrained model weights with additional layers of LSTM and DNN to achieve better results. However, in order to train such large network, there is a requirement of very strong GPU and VRam. Authors have tried and managed this and provided the results with error rate of only 0.11% on famous UW3 dataset.

Later in a few studies "Multimodal Approach" was proposed, which comes very handy while solving the problems that are dependent on multiple modules for hybrid architecture. In [12], authors used a similar approach to provide solution for the emotion recognition in video where they explained image feature extraction by capturing visual information of the detected human faces and the extraction of audio stream for that particular movement and converted them to similar feature vectors further solving the problem relatively. The results obtained were 15% better than the sequential approach which was discussed in [13] where the author converted the audio stream to text also video stream face to text using the encoder-decoder model and further treated that problem as sentiment analysis [14].

Image classification is one of the most prominent research fields, and lots of advanced level research studies have been published in the past few years related to this domain. Generally, the areas that handle these types of research are computer vision, image processing, and ML. Krishna et al.

[15] explained how they have studied DCNN for image classification and used AlexNet Architecture with CNN for this purpose. Thus, the results obtained from the research were quite promising as the test accuracy on MNIST and CIFAR-100 is around 76.24%.

A new and interesting approach for image classification has been observed by the referred paper [16] as the authors show the implementation of transfer learning in the field of DL and computer vision. Lin et al. [16] discussed how the approach of the pretrained model (Google_Inception_V3) can be used in classification for the custom dataset and explained the procedure for change in the last layer of the architecture and match it to the required number of classes in the output layer.

Object detection [17] is one the most popular fields whose devastating advancement of research results can be seen since 2012. Authors in [18] explained the use of an alternative of CNN-YOLO. The proposed algorithm is very fast with respect to CNN as its FPS 155, and its mAP [18] can also reach up to 78.6%, both of which have been way ahead from F-RCNN.

It has been observed in the web that the growth of text data is rising exponentially from past few decades. In today's world, every data user wants very precise results. Nevertheless, retrieval of relevant information from given text has always been a challenge in terms of AI. Therefore, in the paper [19], authors discussed the approach of tokenization, and then the inference time will be shorter and accurate. Most of the proposed research works specifically targeted either image-based methods or textual methods. In the present work, we have proposed a two-way approach for addressing the problem of hate memes classification.

3. Methodology

To specifically target the aim of predicting whether a given meme is hateful or not, the proposed ML model utilizes information from both the image of meme and the text written on it to give a prediction. This is a binary classification problem. The present work explores two different approaches to solve the problem of identifying hate meme. The first approach uses sentiment analysis based on image captioning and text written on meme. The second approach utilizes features extraction and combination from different modalities. To solve this problem, we have used Facebook Challenge Hateful meme Dataset [20] containing approximately 8500 meme images. Both approaches are tested on the same validation set, and results are quite acceptable for both models.

3.1. Data Preprocessing. As we have used Facebook hateful meme dataset containing 8500 images with unique id and each labelled as 0 or 1 (0: not toxic; 1: toxic). In the dataset, we found that all the images were of different size. Thus, we have applied transformation technique to transform all the given images to size $(224 * 224 * 3)$ where 224 is used for height and width and 3 specifies RGB channel. Further, we normalized given images and converted them into a vectorized form.

After transforming each image, we extracted the text from each image. For this, we have used a 3rd party OCR tool to pull out the required text from the given image. Furthermore, to pass our text data into any given specific model of neural network, we require some separate data preprocessing to make it in a suitable format. In the multimodal approach, we have used FastText [21] which is a built-in library model developed by Facebook developer to make our task easy that covers all the preprocessing steps of text as discussed in approach 2 and creates the required feature vector available to us which can be passed into our final NN classifier. In the image caption-based approach, we have embedded the text using the glove embedding algorithm [22], which is developed by Stanford. It is an unsupervised learning algorithm which is trained on 400 thousand words. Using glove, we obtain feature vector which is passed in the final NN-based sentiment analysis model as discussed in detail in approach 1.

3.2. Model Architecture

3.2.1. Sequential Approach. In this model, the basic procedure followed is to first find semantic meaning of meme image in textual format using imaging captioning. This is done by an encoder-decoder model. The encoder model comprises passing the image through a pretrained resnet-152 [23, 24] model (trained on ImageNet dataset) in which we take the last layer (dimension 2048) as output vector. This vector is passed through a linear layer (whose input dimensions were same as the resnet-152 output layer dimension, and its output dimension was equal to embedding input dimension of LSTM component in decoder) [25, 26]. In the decoder, start token is given along with image vector features to predict the first word. The word with the highest probability of being first word is used along with image features to predict the second word. This process goes on till finish token is not given by LSTM [27, 28].

After that, we perform some basic image processing techniques on the image, and then we use the Tesseract API [29, 30] developed by Google to extract text written on image. This extracted OCR text is concatenated with sentence generated by the image caption model [21, 31]. This text is embedded using glove embedding and then passed through an NN-based sentiment analysis model. This NN model consists of convolution layers, max pooling layers, global max pooling layer, fully connected layer, and a sigmoid layer. If the sigmoid function value is greater than the threshold value (0.5), then we classify meme as hateful, else we classify meme as not hateful. Figure 1 describes in detail the step-by-step methodology for the sequential approach.

3.2.2. Multimodal Approach. In this model, we approached the problem differently rather converting our image into text and then solving it as a sentiment analysis problem. Here, we first perform some preprocessing and after that in case of image vector we passed it through our pretrained resnet-152 model in which we take the last layer as output vector for our feature representation. However, the output feature vector of

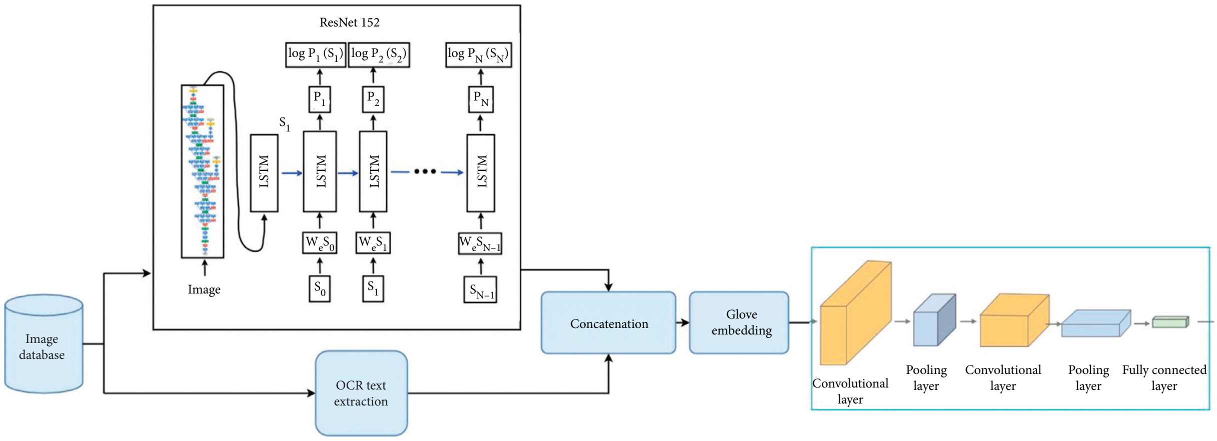

Figure 1: Working of sequential approach.

resnet-152 is of dim 2048. Hence, one more linear layer is added whose input dimensions will be same as resnet-152 output dimension, and its output dimension was similar to that of our language feature dimension. As for our text data rather than fine-tuning manually, we have directly used FastText built-in library to extract the desired features by adding an additional embedding layer whose embedding is kept fixed for simplicity. The output generated after from embedding layer is then passed through a trainable linear layer as a way of fine-tuning our feature represented vector.

Finally, these features received from our vision as well as from our language model are concatenated and transformed into another single feature vector. Later these extracted features are passed to one fully connected layer for classification. Figure 2 depicts the complete model architecture for multimodal approach.

4. Results and Discussion

Both the proposed approaches are tested on the same validation set, and results are found to be quite acceptable for both models. In-detailed results are described further in this section.

4.1. Multimodal Approach. We first address this problem by only using OCR meme text as input. Thus, we obtained the results illustrated in Table 1. This was done by training an NN-based sentiment analysis model on Wikipedia Toxic comments dataset. Adamax optimizer gave the best validation accuracy of 0.55. This was done by training an NN-based sentiment analysis model on Wikipedia Toxic comments dataset. Adamax optimizer gave the best validation accuracy of 0.55.

Table 1 depicts some of the best results obtained in each category after some rigorous amount of training and testing and optimization of hyperparameters. We have used lr_scheduler [32] to automatically determine learning rate value according to the number of epochs. Advantage of using scheduler is while reaching to global minima, step size reduces. Initially scheduler is able to take large steps with higher learning rate value; as it keeps on reaching to minima, step side is reduced. Thanks to PyTorch [32], we were able to use this feature to find appropriate learning rate with respect to each iteration.

We have also used "early stopping" as one of our hyperparameters so that whenever loss tries to go above during training, it stops the model further thus providing the optimum global minimum point and avoid any type of high variance problem leading to overfitting of the model. Figure 3 illustrates the AUC curve with a score of 56.83%, and Figure 4 depicts the confusion matrix for the same score.

4.2. Sequential Approach. In the sequential approach, the results obtained after training and validation on Facebook Hate Meme dataset are shown in this section. We have optimized the hyperparameters to gain the best results possible from this approach. We have used learning rate scheduler (functionality available in keras library) to dynamically adjust learning rate. In other words, as the optimal weights are at a distance from the minima, the learning rate will be high, and when close to optimal weights, then the learning rate will be low. We have also used early stopping to avoid any type of high variance problem leading to overfitting of the model. Table 2 shows the validation accuracy for a variety of optimizers used for training and validation.

We have used dropout regularization of 20% to avoid overfitting. Dropout regularization ensures that the model weights are not affected by noise data while training.

Figure 5 shows the validation set ROC curve obtained for a variety of optimizers, while using the two-way approach combining text and image captioning text.

From Figure 6, it is quite clear that the sequential model (using Adamax optimizer) performed better on validation set than the multimodal (using Adam optimizer). One of the possible reasons for this is that the multimodal approach-based model takes image input features as it is which also contain a lot of noise. In contrast, the model uses the image input features to form a sentence based on highest sequential probability that filters out the noise. However, the sequential model also has a limitation that its accuracy is almost 70%, which means that it may give wrong output sentence that can lead to wrong prediction when this sentence is

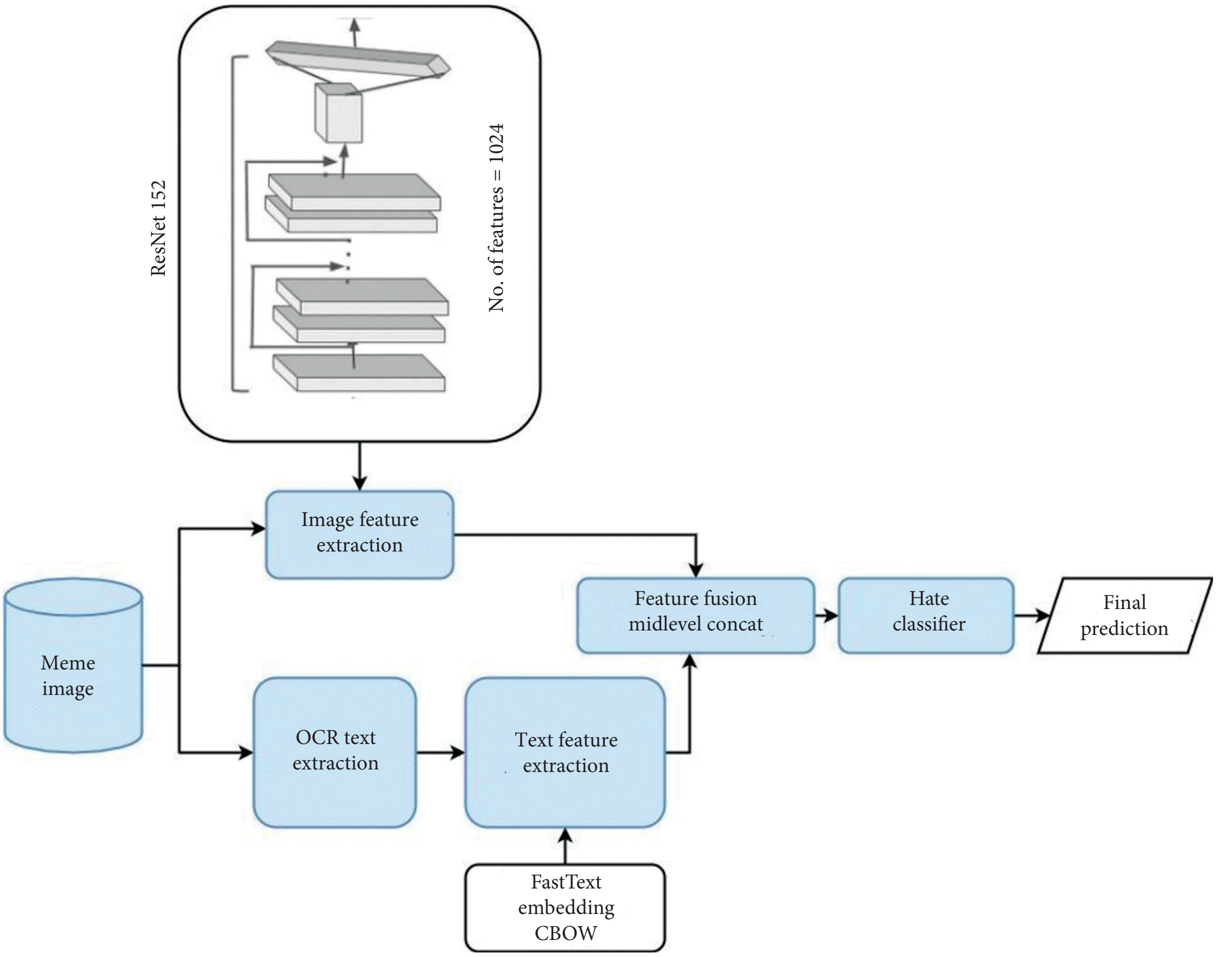

Figure 2: Model architecture of multimodal approach.

Table 1: Some of the best results obtained while training the model using the multimodal approach.

No. of epochs	Optimizer	Batch size	Validation accuracy
10	RMSProp	32	0.51
8	Adadelta	16	0.53
6	Adamax	8	0.55

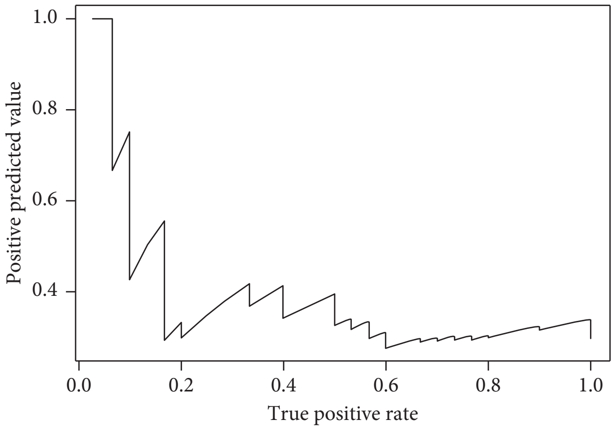

Figure 3: Graph showing AUC curve with the score of 56.83%.

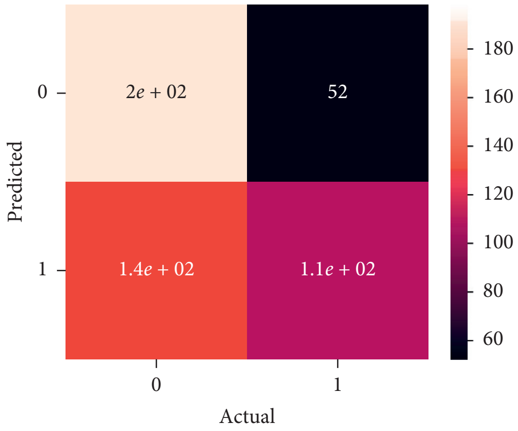

Figure 4: Confusion matrix obtained by using multimodal approach on validation set.

Table 2: Some of the best results obtained while training the model using sequential approach.

No. of epochs	Optimizer	Batch size	Validation accuracy
16	RMSProp	32	0.62
20	Adamax	32	0.64
16	Adam	20	0.59

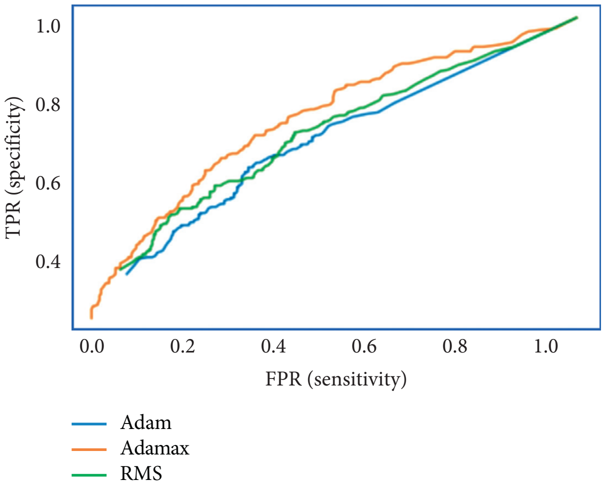

Figure 5: Validation set ROC curve obtained for different optimizers on using both OCR text and image captioning text.

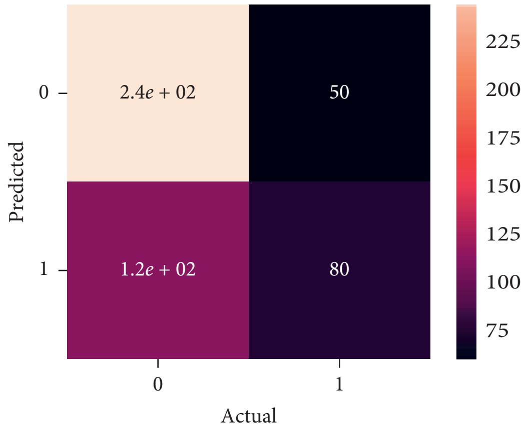

Figure 6: Confusion matrix obtained on using Adamax optimizers used in sequential approach on validation set.

Table 3: Comparative analysis.

S. no.	Model	Validation accuracy
1	Text based	0.55
2	Multimodal	0.59
3	Sequential	0.64

concatenated with OCR text and passed through the sentiment analysis model.

Finally, for the hateful meme classifier, the comparative analysis of accuracies is depicted in Table 3. The best model obtained from the sequential model had a validation accuracy of 0.64, and the best model obtained from the multimodal approach had a probability of 0.59 as seen from Tables 1 and 2. Human accuracy on this prediction problem is approximately 80%, provided by Facebook for Hateful Memes Challenge and dataset [20]. We achieved a decent accuracy on this problem when compared with human accuracy.

5. Conclusion and Future Works

Memes on social media are one of the most popular ways to send false and hateful information to the masses. This work classifies hateful memes targeted at a particular audience so as to modify their opinions on certain issues. Dataset of memes is provided by Facebook in the open challenge. In the present work, we have specifically proposed a sequential approach and a multimodal one to extract information from image captions as well as text in the memes. Thus, a two-way feature extraction was performed, and deep learning models including a combination of OCR, glove, and encoder-decoder architecture are applied in addition to tools like Tesseract API for training. Furthermore, the two approaches are compared over benchmarks, as well as with the dataset collected from other sources. For this work, results obtained were found to be quite comparable to human accuracy. In future, we plan to extend this work to other multimodal feature extraction methods so as to improve the training over the given dataset. Further, social media trends and patterns are fast changing, so there is a need of real time capturing of memes with respect to a particular domain so as to find the influential entities. This work can be extended to capture such real time data and train deep learning models for identifying hateful memes.

Abbreviations

AI:	Artificial intelligence
CNN:	Convolution neural network
CV:	Computer vision
DCNN:	Deep convolutional neural network
DL:	Deep learning
F-RCNN:	Faster R-CNN
DNN:	Deep neural network
GPU:	Graphics processing unit
ML:	Machine learning
NLP:	Natural language processing
NN:	Neural network
LSTM:	Long short-term memory
OCR:	Optical character recognition
R-CNN:	Region-based convolution neural network
RGB:	Red green blue
YOLO:	You only look once.

Data Availability

The authors have participated in the Hateful Meme Classification challenge hosted by Facebook. While registering for this competition, they have agreed to not outsource the dataset as per the restricted content it may have for general public and can only be used for research purpose to provide solution for the given problem.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank Bennett University for providing platform and infrastructure for carrying out this research.

References

[1] Y. Liu, D. Liu, and Y. Chen, "Research on sentiment tendency and evolution of public opinions in social networks of smart city," *Complexity*, vol. 2020, Article ID 9789431, 13 pages, 2020.

[2] B. Xu, G. Fan, and D. Yang, "Topic modeling based image clustering by events in social media," *Scientific Programming*, vol. 2016, Article ID 5283471, 7 pages, 2016.

[3] C. Zhao, H. Zhao, G. Wang, and H. Chen, "Hybrid depth-separable residual networks for hyperspectral image classification," *Complexity*, vol. 2020, Article ID 4608647, 17 pages, 2020.

[4] G. Modwel, A. Mehra, N. Rakesh, and K. K. Mishra, "A robust real time object detection and recognition algorithm for multiple objects," *Recent Advances in Computer Science and Communications*, vol. 14, no. 1, pp. 330–338, 2021.

[5] X. Zhang, S. Wang, G. Cong, and A. Cuzzocrea, "Social big data: mining applications and beyond," *Complexity*, vol. 20192 pages, 2019.

[6] W. Park, Y. You, and K. Lee, "Detecting potential insider threat: analyzing insiders' sentiment exposed in social media," *Security and Communication Networks*, vol. 2018, Article ID 7243296, 18 pages, 2018.

[7] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting Of The Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, OR, USA, June 2011.

[8] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659, Las Vegas, NV, USA, June 2016.

[9] P. Anderson, X. He, C Buehler et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, June 2018.

[10] A. Chaudhuri, K. Mandaviya, P. Badelia, and S. K. Ghosh, "Optical character recognition systems," *Optical Character Recognition Systems for Different Languages with Soft Computing*, vol. 352, pp. 9–41, 2017.

[11] C. Wick, C. Reul, and F. Puppe, "Calamari-a high-performance tensorflow-based deep learning package for optical character recognition," 2018, https://arxiv.org/abs/1807.02004.

[12] S. E. Kahou, X. Bouthillier, P. Lamblin et al., "Emonets: multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[13] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, and T. S. Huang, "How deep neural networks can improve emotion recognition on video data," in *IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, September 2016.

[14] R. Yan, Z. Xia, Y. Xie, X. Wang, and Z. Song, "Research on sentiment classification algorithms on online review," *Complexity*, vol. 2020, Article ID 5093620, 6 pages, 2020.

[15] M. Krishna, M. Neelima, H. Mane, and V. Matcha, "Image classification using deep learning," *International Journal of Engineering & Technology*, vol. 7, p. 614, 2018.

[16] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, Honolulu, HI, USA, July 2017.

[17] E. Torres-Pereira, H. Martins-Gomes, A. E. Monteiro-Brito, and J. M. De Carvalho, "Hybrid parallel cascade classifier training for object detection," *Advanced Information Systems Engineering*, vol. 8827, 2014.

[18] V. Singh and B. Saini, "An effective tokenization algorithm for information retrieval systems," *Computer Science & Information Technology (CS & IT)*, vol. 4, 2014.

[19] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 4, no. 34–47, p. 4, 2001.

[20] Hateful Memes Challenge and dataset for research on harmful multimodal content, https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/.

[21] fastText: Facebook Open Source. https://fasttext.cc/.

[22] Z. H. Kilimci and S. Akyokus, "Deep learning-and word embedding-based heterogeneous classifier ensembles for text classification," *Complexity*, vol. 2018, Article ID 7130146, 10 pages, 2018.

[23] R. Wang, H. Yu, G. Wang, G. Zhang, and W. Wang, "Study on the dynamic and static characteristics of gas static thrust bearing with micro-hole restrictors," *International Journal of Hydromechatronics*, vol. 2, no. 3, pp. 189–202, 2019.

[24] Tensorflow documentation: https://www.tensorflow.org/api_docs/python/tf/keras/applications/ResNet152.

[25] S. Osterland and J. Weber, "Analytical analysis of single-stage pressure relief valves," *International Journal of Hydromechatronics*, vol. 2, no. 1, pp. 32–53, 2019.

[26] S. Ghosh, P. Shivakumara, P. Roy, U. Pal, and T. Lu, "Graphology based handwritten character analysis for human behaviour identification," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 1, pp. 55–65, 2020.

[27] B. Gupta, M. Tiwari, and S. Singh Lamba, "Visibility improvement and mass segmentation of mammogram images using quantile separated histogram equalisation with local contrast enhancement," *CAAI Transactions on Intelligence Technology*, vol. 4, no. 2, pp. 73–79, 2019.

[28] T. Wiens, "Engine speed reduction for hydraulic machinery using predictive algorithms," *International Journal of Hydromechatronics*, vol. 2, no. 1, pp. 16–31, 2019.

[29] H. S. Basavegowda and G. Dagnew, "Deep learning approach for microarray cancer data classification," *CAAI Transactions on Intelligence Technology*, vol. 5, no. 1, pp. 22–33, 2020.

[30] O. C. R. Tesseract: Google Open Source. https://opensource.google/projects/tesseract.

[31] H. Memes, "Phase 2: Facebook," 2021, https://www.drivendata.org/competitions/70/hateful-memes-phase-2/.

[32] P. Package, "Torch Optim," 2021, https://pytorch.org/docs/stable/optim.html.