A Deep Multi-modal Neural Network for the Identification of Hate Speech from Social Media

Gunjan Kumar, Jyoti Prakash Singh, Abhinav Kumar

A deep multi-modal neural network for the identification of hate speech from social media

Gunjan Kumar[1], Jyoti Prakash Singh[1], and Abhinav Kumar[2]

[1] National Institute of Technology Patna, India
[2] Institute of Technical Education & Research, Siksha 'O' Anusandhan, Bhubaneswar, India
gunjank.phd20.cs@nitp.ac.in,
jps@nitp.ac.in,
abhinavanand05@gmail.com

Abstract. Hate speech can be particularized as an intentional and chronic act to harm a single person or a group of individuals. This act can be performed via social networking websites such as Twitter, YouTube, Facebook, and more. Most of the existing approaches for finding hate speech are concentrated on either textual or visual information of the posted social media contents. In this work, a multi-modal system is proposed that uses textual as well as the visual contents of the social media post to classify it into *Racist, Sexist, Homophobic, Religion-based hate, other hate* and *No hate* classes. The proposed multi-modal system uses a convolutional neural network-based model to process text and a pre-trained VGG-16 network to process imagery contents. The performance of the proposed model is tested with the benchmark dataset and it achieved significant performance in classifying social media posts into six different hate classes.

Keywords: Hate-speech · Multi-modal · Twitter images · VGG-16 · CNN

1 Introduction

Social media such as Twitter, Instagram, YouTube, and Facebook encourages users to share ideas, thoughts, and information through virtual networks and communities[5]. The present electronic era makes it more popular and comfortable to access and communicate over it. These communications include blogging, reviews, social gaming, sharing of photos, video, audio, text, and business networks. It can connect and share information worldwide at the same time with many people. In recent years, social media users are tremendously increased. Currently, social media is also being utilized by governments to engage with constituents and voters. For businesses, social media is an essential tool for companies to find and hold with customers, increase sales through promotion, advertisements, and offering customer service or support. Social media has the capability to gather information from every user which helps to focus on research

in many areas. It has numerous advantages but some of the severe challenges are also associated with it. Every user has the freedom to express their views without revealing their real identity, but some users are misusing this freedom to write the offensive language. Gomez1et al. [3] defined hate-speech as an "aggressive, intentional act carried out by a group or individual using electronic forms of contact, repeatedly or overtime against a victim that cannot easily defend him or herself". Social media users are targeted by Hate-Speech and Offensive language such as abusive, hurtful, derogatory, or unlawful user-generated content by some mischievous users [10]. As a result of the misuse of online interactions, many people have fallen into depression, anxiety, other mental illness, and they feel poor in their position to react to Internet violence or harassment [10] [12] [17]. In severe cases, if the victim cannot reply and motivate himself/herself, then they commit suicide too. All these incidents encourage researchers to propose a practical solution to eliminate the negative impact of social media. Identifying hate speech and removing it from social media or preventing the writing of these posts is an important task.

Nowadays, social media users are frequently using text, images, videos, audios, and a combination of these media for their social interaction. A number of works have been reported by researchers that use textual contents of the social media posts to identify hate contents [7] [6] [17] [10] [11]. The role of imagery contents with the textual contents is important because sometimes by seeing a single modality of the post it is very difficult to recognize hate contents. For example, Figure 3 not related to hate speech if someone only sees the imagery content whereas the combination of imagery and textual content make it hate content. Therefore, when designing an automated hate speech detection system it is important to take care of other modalities of the posts also to make online social media platform vigorous and secure [12][7] [6]. A few works [2] [14] [12] [3] have been reported where researchers tried to use textual and imagery contents of the social media posts to train a system for binary classification (Hate or, Not-hate). In line with their works, in this work, a multi-modal system is developed to classify social media posts into six different classes such as *Racist, Sexist, Homophobic, Religion-based hate, other hate* and *No hate*. To process textual contents, a convolutional neural network-based model is developed whereas to process imagery contents, a fine-tuned VGG-16 network is used. After getting the textual and imagery features from the said convolutional network and VGG-16 networks, the features are concatenated and pass through a softmax layer to classify posts into different classes. To validate the proposed system, a benchmark dataset [3] is used. The overall contributions of this paper are as follows:

- To extract features from the convolutional neural network and pre-trained VGG-16 network from textual and imagery contents, respectively.
- To propose a deep multi-modal neural network-based model for the classification of social media posts into the six different hate classes.

The rest of the paper is structured as follows: Section 2 presents related works for the detection of Hate Speech while Section 3 presents our methodology for identification of Hate Speech. Section 4 lists the results of the proposed model. Finally, Section 5, concludes the paper and has discussed the future directions for this work.

2 Related work

Hate speech is the most prominent problem on social media, and an ample amount of research is going in this field. Identification and prevention of hate speech on social media are essential to avoid harm and injury in society. Salawu et al. [16] broadly classify the existing hate-speech detection methods into four classes, specifically supervised learning, rule-based, lexicon-based, and mixed-initiative methods. Supervised learning-based methods generally using Naive Bayes and Support Vector Machine(SVM) classifier to build predictive models for hate-speech detection. Lexicon-based processes are using word lists and find the existence of words inside the lists to identify hate speech. The rule-based method compares the text to pre-determine rules to recognize offensive and mixed-initiatives strategies to amalgamate human-based reasoning with one or more of the methods mentioned above. Kumari et al. [10][11] worked on multi-lingual (Hindi, English and Bangla) code-mixed text. Also focus on finding the aggression level of the comment posted on social media. Each comment is marked as Non-aggressive, Covertly aggressive, or Overtly aggressive. They proposed two deep learning systems: Long Short Term Memory (LSTM) and Convolutional Neural Network (CNN), with two separate inputs in text representations, One-hot, and FastText embedding. It was found that for Hindi and Bangla datasets, LSTM is performing better with FastText embedding, and CNN is performing better for English. Chan et al. [1] identified hate-speech based on the social cognitive theory, which focuses on the known and needs to know, and also the reciprocal relationships between perpetrators, victims, and bystanders.

Another group of researchers [19] [14] [15] [13] proposed multi-modal systems for the identification of hate speech from social media platform. Wang et al. [19] propose a modal having multi-model encoder-decoder using bi-directional LSTM on two datasets from one of the famous social networking site Instagram (video and photo sharing) and Vine (Small video sharing) platform. Several textual features involve word-level TF-IDF vectors, character-level TF-IDF vectors, and intellectual characters from Linguistic Inquiry Word Count (LIWC). They have also used several deep learning models as the standard, including LSTM, Text-CNN, and an accuracy of 0.864 and F_1-score of 0.86 on Instagram and an accuracy of 0.838 and F_1-score of 0.841 on the Vain dataset. Cheng et al. [2] proposed XBully, one of the hate-speech detection frameworks, which first redevelop multi-modal social networking website data as a heterogeneous network and then focus on learning node embedding representations upon it. Extensive experimental evaluations on real-world multi-modal social networking website datasets showed that the XBully architecture performs better than the existing

hate-speech detection models. Yang et al. [20]present several fusion methods to integrate text and image signals. They adopt the baseline convolutional text differentiator and the image characteristics of photos. Also described multiple approaches to fuse texts and pictures, involving elementary concatenation, gated aggregated, bi-linear modification, and noticed with various alternations. Pretrain a deep, Cheng et al. [2] proposed XBully, one of the hate-speech detection frameworks, which first redevelop multi-modal social media data as a heterogeneous network and then focus on learning node embedding representations upon it. Extensive experimental evaluations on real-world multi-modal social media datasets showed that the XBully framework performs better than the existing hate-speech detection models. Finally, using attention fusion with deep cloning performs 84.8, an improvement over basic concatenation is statistically significant at the 99% confidence level.

Kumari et al. [14] has tried to find the bullying comment over social media posts containing text as well as image. They represent the text and image together and form a module that harmoniously learns the image and text, eliminates the need for independent learning. Single-layer Convolution Neural Network (CNN) is performing better with the 2-layered convolutional neural network. They used 3 channels of the word and three channels of the color photo to present the input and achieved a recall value of 74% for the abusive comment classification. Paul et al. [15] designed a deep learning-based multi-modal architecture that helps in the early detection of hate speech. They predict a posted comment is hate or not as early as possible. The multi-modal features fusion-based experimental analysis achieved a 0.75 F-measure using the Residual BiLSTM-RCNN model, reflecting the efficiency of the proposed framework. Kumari et al. [13] proposed a model based on a Binary Particle Swarm Optimization (BPSO) and Convolutional Neural Network (CNN) to categorize the social networking website posts having a photo with cognates text-message into 3 classes (non-aggressive, medium-aggressive, and high-aggressive). The dataset that they are using having symbolic photos and text messages to validate the proposed model. VGG-16 model has been used to find out the image characteristic and a three-layered CNN to find out the text characteristic. The combined characteristic set is attained by adding the characteristics of image and text and enhanced using the BPSO algorithm to gain the more appropriate factor and achieve a weighted F1-Score of 0.74. Most of the developed models are classifying the post in binary class.

In most of the earlier works, researchers proposed the model using textual contents of social media for the identification of hate speech. A few potential works such as [19] [14] [15], [13] that uses multi-modal content of the social media platform to identify hate speech, but most of them are a binary classification task only, i.e., they used either hate or not-hate class to develop their system. In the current work, the proposed deep multi-modal system is trained for the six different hate classes such as *Racist, Sexist, Homophobic, Religion-based hate,*

other hate and *No hate* to see the efficiency of the proposed system in the granular level of hate speech classification.

3 Methodology

This section describes the details about the dataset and proposed methodology. The framework of the proposed multi-modal system is shown in Figure 1. The design incorporates two parallel deep neural network structures: (i) Convolutional Neural Network (CNN) for processing tweet-text, and (ii) VGG-16 for images. The text is fed through CNN layers to extract text features whereas, for the image, Convolutional Neural Network (CNN)-based pre-trained VGG-16 model is used to extract imagery features. For the pre-trained VGG-16 network, the weights of the last two layers were trained and all other weights are marked as non-trainable to transfer the pre-trained weights for the current task. Then the extracted imagery features are mapped to the dense layer containing 128-neurons to get the 128-dimensional imagery feature vector (see Figure 1). Similarly, the text features extracted through the convolution and max-pooling operation are mapped to the dense layer containing 128-neurons to get a 128-dimensional textual feature (see Figure 1). Then the extracted textual and imagery contents are concatenated and input into a softmax layer to classify tweets into six different classes *Racist, Sexist, Homophobic, Religion-based hate, other hate* and *No hate*.

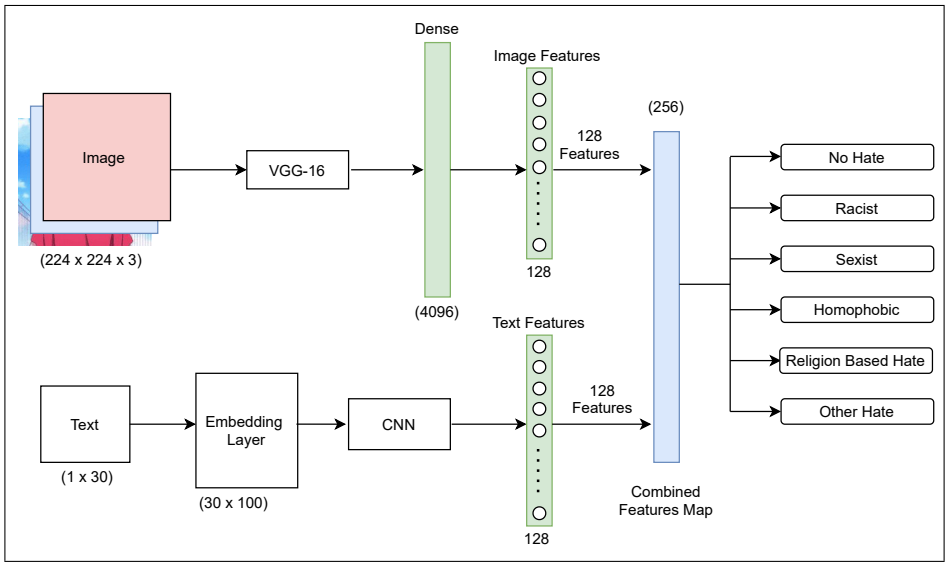

Fig. 1. Flow diagram of the proposed model for hate speech classification

3.1 Data description and pre-processing

The dataset[3] published by Gomez et al. [3] is used to validate the proposed system. The dataset contains tweets from six different categories (i) Not-hate, (ii) Racist, (iii) Sexist, (iv) Homophobic, (v) Religion-based-hate, and (vi) Other-hate. Figures 2, 3, 4, 5, 6, and 7, represents the data samples for each of the classes, *Not-hate, Racist, Sexist, Homophobic, Religion-based-hate*, and *Other-hate*, respectively. The data samples for each of the classes can be seen in Table 1. For the pre-processing of images, it is resized to (224 x 224 x 3), where (224 x 224) is the height and width of the image, and 3 is the RGB component. The pixel value of images was normalized between 0 to 1 by dividing all pixel values by 255. After pre-processing, images are directly inputted to the pre-train model to extract the relevant feature from them. Each tweet text is represented in 30 words; for the tweets having less than 30 words, we used padding and for more than 30 words, we curtailed out the words to make it into an equal length of 30.

Fig. 2. Nohate: "When a android nigga wanna talk about phone battery:"

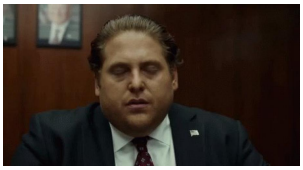

Fig. 3. Racist: "@washingtonpost @Sabriyyah54 More stupid racist white TRASH..."

Fig. 4. Sexist: "@Doggin-Trump No sympathy for the twat waffle"

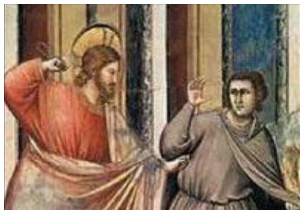

Fig. 5. Homophobic: "@eeeeeeeeeekkk Me: Mental Illness: shut up faggot"

Fig. 6. Religion-based-hate: "The only POS is the Muzzie who will be arrested for treason"

Fig. 7. Other-hate: "My horses are retarded"

[3] https://github.com/gombru/multi-modal-hate-speech

Table 1. The description of the data samples for each of the classes

Tweets categories	Numbers of tweet per class
No attacks to any community	112845
Racist	11925
Sexist	3495
Homophobic	3870
Religion-based	163
Attacks to other communities	5811
Not in any categories	11714

3.2 Image classification (VGG16)

VGG-16 is based on convolutional neural network architecture and it is trained on ImageNet dataset to classify it into 1,000 classes. It has 16 layers out of this 13 are convolutional layers and 3 layers are fully connected. It takes input as the image of size (224 224 3) and filter (3 x 3) for performing convolution operation. The detailed information regarding the parameters and layers of the VGG-16 can be seen in Simonyanet et al. [18]. VGG-16 has attractive architecture; it is the best option for pull-out the characteristics from the images [9]. Due to the diverse popularity of VGG-16 networks in extracting features from the images, this work also utilizes the VGG-16 network for extracting imagery features. The VGG-16 network can be modified at the last layer according to the tasks. In this work, weights between the first 14 layers are frozen and the weights of the last two layers are trained.

3.3 Text classification (CNN)

The convolution neural network (CNN) has the ability to identify the pattern and provide meaningful sense from the textual content of tweets. In the convolution layer, the dot product is performed between the weights and input. The size of the resultant is depended upon the filter size and number of filters used. ReLu is used as the activation function, whereas the Max-pooling layer is used to find the most important feature from a pooling window [18]. The detailed description of the CNN network can be seen in [4] [8].

To pull out the textual features from text messages, we first embedded each word into an embedding vector using pre-train Glove embedding vectors. We set 2-gram, 3-gram, and 4-gram filters over the 1st, 2nd, and 3rd convolution layers. Then we applied max-pooling of size 5 over it to pull out the best features from it. These features are then concatenated with the image features (Figure 1). This concatenated feature vector is then passed through the softmax layer to classify it into six different hate classes. The extensive experiments were performed to choose the best set of parameters. A learning rate of 0.001, epochs equals 16 and the batch size of 32 performed best. Since the model is a categorical classifier, we

Table 2. Hyper-parameter settings for the proposed multi-modal system

Model	Hyper-Parameters	Value
VGG-16	Image size	224 x 224 x 3
	Optimizer	Adam
	Loss Function	Categorical crossentropy
	Activation function	ReLU, Softmax
	Epochs	16
	Batch size	32
	Learning rate	0.001
CNN	Maximum sequence length	30
	Maximum No. of words	4000
	No. of filter	1024
	Filter size	2,3,4
	Optimizer	Adam
	Dropout rate	0.2
	Pooling size	5
	Loss Function	Categorical crossentropy
	Activation function	ReLU
	Epochs	16
	Batch size	32
	Learning rate	0.001

use Categorical cross-entropy as our loss function. Adam is used for optimization and the Softmax activation function is used in the output layer. The detailed hyper-parameters used in the experimental analysis have been shown in table 2.

4 Result

This section describes the result obtained from the proposed model for the Dataset MMHS150K. The proposed multi-modal system classifies the posted hate Twitter contents into six different hate classes. The provided dataset[4] have 138109 tweets out of which 10,000 tweets are for testing the model and 5,000 tweets are for validating the model and the remaining data samples are for training the model.

The proposed model categories the Twitter contents into six different hate classes and for each class, we get the precision, recall, and F_1-score (see Table 3). The proposed system achieved a precision of 0.84, 0.76,0.67, 0.72, 0.60, and 0.76 for No-hate, Racist, Sexist, Homophobic, Religion, and Other-hate classes, respectively. The recall values achieved by the proposed systems are 0.97, 0.29, 0.26, 0.47, 0.23, and 0.42 for No-hate, Racist, Sexist, Homophobic, Religion and Other-hate classes, respectively. Similarly, the proposed system achieved an F_1-scores of 0.90, 0.41, 0.37, 0.57, 0.33 and 0.54 for No-hate, Racist, Sexist,

[4] https://github.com/gombru/multi-modal-hate-speech

Table 3. Result of proposed deep multi-modal system to classify Twitter contents into six different hate classes

Class	Precision	Recall	F_1-score
Not-hate	0.84	0.97	0.90
Racist	0.76	0.29	0.41
Sexist	0.67	0.26	0.37
Homophobe	0.72	0.47	0.57
Religion	0.60	0.23	0.33
Other-hate	0.76	0.42	0.54
Weighted average	**0.82**	**0.83**	**0.81**

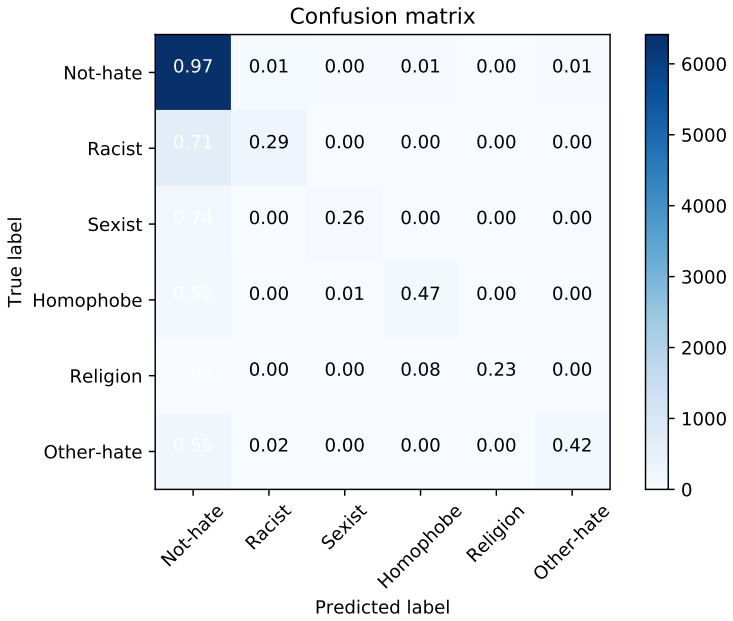

Fig. 8. Confusion matrix of the proposed deep multi-model neural network for hate speech classification

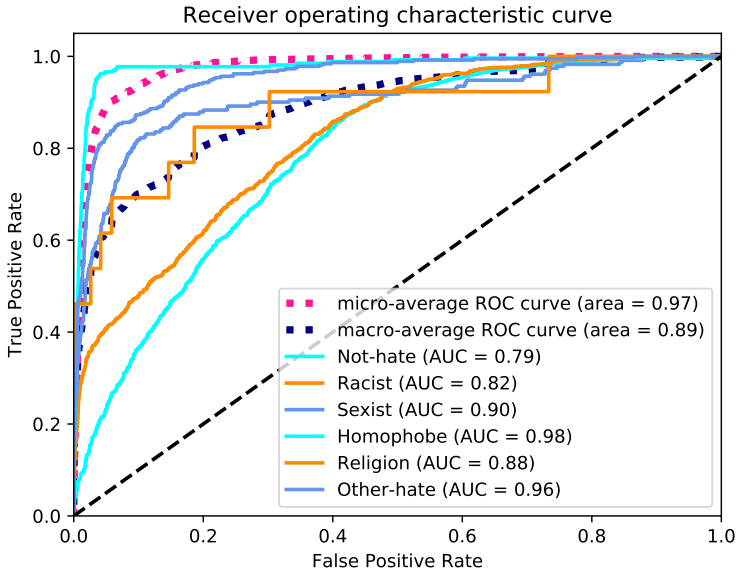

Fig. 9. Receiver operating characteristic curve of the proposed deep multi-modal neural network for hate speech classification

Homophobic, Religion and Other-hate classes, respectively. The proposed system achieved the weighted precision, weighted recall, and weighted F_1-scores of 0.82, 0.83, and 0.81, respectively. The confusion matrix and ROC curve for the proposed system can be seen in Figures 8 and 9, respectively.

5 Conclusion

Hate speech on social media can cause people to suffer from depression, anxiety, and other mental illnesses. Most of the earlier works are reported for the identification of hate speech uses textual contents only. A few works are reported where researchers tried to train their system into the multi-modal setting, but they tested their system for the binary class classification only (i.e., whether hate or not hate). In this work, we have proposed a deep multi-modal system to classify Twitter hate contents into *Racist, Sexist, Homophobic, Religion-based hate, other hate* and *No hate* classes. The textual content of the tweets was processed using a convolutional neural network and imagery content was processed using fine-tuned VGG-16 model. The proposed deep multi-modal neural network achieved promising performance with the weighted precision of 0.82, weighted recall of 0.83, and weighted F_1-score of 0.81. The benchmark dataset used in this study suffers from the data imbalance problem. Therefore, in the future data augmentation, under-sampling, and over-sampling techniques can be applied to get better performance throughout all of the hate classes. The

performance of the model is only tested with convolutional neural network and pre-trained VGG-16 network, therefore, in the future other deep learning models such as BERT for textual contents whereas ResNet50, VGG-19, Xception networks for imagery contents can be tested for better performance.

References

1. Chan, T.K., Cheung, C.M., Lee, Z.W.: Cyberbullying on social networking sites: A literature review and future research directions. Information & Management p. 103411 (2020)
2. Cheng, L., Li, J., Silva, Y.N., Hall, D.L., Liu, H.: Xbully: Cyberbullying detection within a multi-modal context. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. pp. 339–347 (2019)
3. Gomez, R., Gibert, J., Gomez, L., Karatzas, D.: Exploring hate speech detection in multimodal publications. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1470–1478 (2020)
4. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751. ACL (2014). https://doi.org/10.3115/v1/D14-1181
5. Kumar, A., Rathore, N.C.: Relationship strength based access control in online social networks. In: Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2. pp. 197–206. Springer (2016)
6. Kumar, A., Saumya, S., Singh, J.P.: Nitp-ai-nlp@ hasoc-fire2020: Fine tuned bert for the hate speech and offensive content identification from social media. FIRE (Working Notes), CEUR (2020)
7. Kumar, A., Saumya, S., Singh, J.P.: Nitp-ainlp@ hasoc-dravidian-codemix-fire2020: A machine learning approach to identify offensive languages from dravidian code-mixed text. FIRE (Working Notes), CEUR (2020)
8. Kumar, A., Singh, J.P.: Location reference identification from tweets during emergencies: A deep learning approach. International journal of disaster risk reduction **33**, 365–375 (2019)
9. Kumar, A., Singh, J.P., Dwivedi, Y.K., Rana, N.P.: A deep multi-modal neural network for informative twitter content classification during emergencies. Annals of Operations Research pp. 1–32 (2020)
10. Kumari, K., Singh, J.P.: Ai ml nit patna at hasoc 2019: Deep learning approach for identification of abusive content. In: FIRE (Working Notes). pp. 328–335 (2019)
11. Kumari, K., Singh, J.P.: Ai_ml_nit_patna@ trac-2: Deep learning approach for multi-lingual aggression identification. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying. pp. 113–119 (2020)
12. Kumari, K., Singh, J.P.: Identification of cyberbullying on multi-modal social media posts using genetic algorithm. Transactions on Emerging Telecommunications Technologies **32**(2), e3907 (2021)
13. Kumari, K., Singh, J.P., Dwivedi, Y.K., Rana, N.P.: Multi-modal aggression identification using convolutional neural network and binary particle swarm optimization. Future Generation Computer Systems **118**, 187–197 (2021)
14. Kumari, K., Singh, J.P., Dwivedi, Y.K., Rana, N.P.: Towards cyberbullying-free social media in smart cities: a unified multi-modal approach. Soft Computing **24**(15), 11059–11070 (2020)

15. Paul, S., Saha, S., Hasanuzzaman, M.: Identification of cyberbullying: A deep learning based multimodal approach. Multimedia Tools and Applications pp. 1–20 (2020)
16. Salawu, S., He, Y., Lumsden, J.: Approaches to automated detection of cyberbullying: A survey. IEEE Transactions on Affective Computing **11**(1), 3–24 (2017)
17. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the fifth international workshop on natural language processing for social media. pp. 1–10 (2017)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
19. Wang, K., Xiong, Q., Wu, C., Gao, M., Yu, Y.: Multi-modal cyberbullying detection on social networks. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)
20. Yang, F., Peng, X., Ghosh, G., Shilon, R., Ma, H., Moore, E., Predovic, G.: Exploring deep multimodal fusion of text and photo for hate speech classification. In: Proceedings of the Third Workshop on Abusive Language Online. pp. 11–18 (2019)