EasyChair Preprint

№ 10743

Multimodal Hate Speech Detection from Videos and Texts

Nishchal Prasad, Sriparna Saha and Pushpak Bhattacharyya

August 19, 2023

Multimodal Hate Speech Detection from Videos and Texts

Nishchal Prasad[1], Sriparna Saha[1], and Pushpak Bhattacharyya[2]

[1] Indian Institute of Technology (IIT) Patna, Bihar, India
prasadnishchal.np@gmail.com, sriparna@iitp.ac.in
[2] Indian Institute of Technology (IIT) Bombay, Mumbai, India
pushpakbh@gmail.com

Abstract. Since social media posts also consist of videos with associated comments, and many of these videos or their comments impart hate speech, detecting them in this multimodal setup is crucial. We have focused on the early detection of hate speech in videos by exploiting features from an initial set of comments. We devise Text Video Classifier (TVC), a multimodal hate classifier, based on four modalities which are character, words, sentence, and video frame features, respectively, and develop a Cross Attention Fusion Mechanism (CA-FM) to learn global feature embeddings from the inter-modal features. We report the architectural details and the experiments performed. We use several sampling techniques and train this architecture on a Vine dataset of both video and their comments. Our proposed architectural design attains performance improvement on the models previously constructed on the chosen dataset, for an output probability threshold of 0.5, showing the positive effect of using the CA-FM and TVC.

Keywords: TVC · Multimodal · Cross Attention Fusion Mechanism

1 Introduction

Hate speech can be understood as any kind of communication in speech, writing, or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of religion, ethnicity, nationality, race, color, descent, gender, or other identities factor[3]. According to The American Experience 2020: Online Hate and Harassment Report,[4], 44% of Americans have been subjected to some type of internet harassment.

Hate speech encompasses a wide range of expressions that advocate, encourage, promote, or excuse hatred, violence, or prejudice against an individual or a group of individuals for a variety of causes. It poses serious threats to democratic society's cohesion, human rights protection, and the rule of law. If neglected, it can evolve into larger-scale acts of violence and war. Hate speech is, in this view, an extreme kind of bigotry that contributes to hate crimes.

[3] https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml
[4] https://www.adl.org/online-hate-2021

Hate speech has been connected to a rise in violence against minorities around the world, including ethnic cleansing, lynching, and mass shootings. These have serious implications on the well beings of individuals, both emotionally and mentally, that lead to depression, suicidal tendencies, or an inclination to impart hate speech or violent actions. One way to appease such an effect is to look for a way to spot hate speech actively and as automatically as possible. There have been several developments in deep learning and machine learning-based approaches to detect hate speech. And since today's social media posts are multimodal in nature (i.e., a combination of text, image, video, audio, etc.), hence the multimodal approach.

In this study, we create and analyze a deep learning-based approach for detecting hate speech in videos with accompanying comments. Our work focuses on the early detection of hate speech taking into account the first few comments along with their associated video.Using all the available information; i.e., the inputs from multiple modalities (text, video); provide more information about content, thus helping the deep learning architecture to learn information representation capturing the whole context. The following lists our contributions:

- We have proposed a multimodal architecture, called Text Video Classifier (TVC), for early identification of hate speech over both videos and their associated comments.
- We developed a Cross Attention Fusion Mechanism, to learn from multimodal features i.e. having combinations of word, sentence, and character embeddings with video frame features.
- We train the architectures with sampling techniques to see their importance in such a multimodal scenario.
- We have experimented with different frameworks (unimodal, multimodal) with several architectures having various combinations of text and image features with various attention fusion mechanisms on the vine dataset [10] to concur with our proposed architecture, with significant performance improvements in comparison to the previous benchmark frameworks ([9]).

2 Related works

Unimodal approaches: Many solutions and architectures are developed and experimented on different unimodal datasets. Such as Reynolds et al. [11], wherein they experimented with a decision tree and an instance-based learner to detect cyberbullying on a dataset collected from the website "Formspring.me". Karthik et al. [5] analyzed several binary and multi-class classification architectures on a dataset of 4.5k YouTube comments, aiming to detect cyberbullying (upon two classes, i.e., sexual or racism) using SVM and Naive Bayes classifiers. Badjatiya et al. [17] analyzed multiple architectures including Random Forest, Gradient Boosted Decision Trees, SVMs, Logistic Regression, and Deep Neural Networks to detect cyberbullying on a dataset of 16000 tweets.

Kumari et al. [18] developed a deep learning-based architecture for detecting aggressive posts on a dataset of symbolic images gathered from Google searches

to query hostile photos. Djuric et al. [19] presented a two-step procedure. They employed a paragraph-to-vector (paragraph2vec [20]) approach to model the comments and words together, giving a text embedding of lower dimension in which the words are grouped together with their semantically similar counterpart comments. These embeddings were further utilized for training models to classify between the hate and non-hate instances on Yahoo comments.Soumitra et al. [1] created a multi-domain hate speech corpus (MHC) of English tweets and used a stacked ensemble-based hate speech classifier (SEHC) by stacking the existing SOTA models to detect hate over MHC.

Multimodal approach: In a multimodal setup trying to detect our objective using a single modality is inefficient, because a single mode of information misses out on rich information that can be learned from other modalities.Poria et al. [21] demonstrated how different modes of information can be effectively used to yield a more fine-grained decision. Cambria et al. [22] used a multimodal fusion approach for combining information from different modalities for the analysis of semantics and sentics. To recognize cyberbullying from Instagram posts Zhong et al. [14] used the features from the text(comments), and images. Raul et al. [6] created the multimodal MMHS150K dataset [5], and trained their transformer models upon it to target the problem of hate speech classification in a multimodal setup by using the features extracted from both text and images.

Kumari et al. [7] developed a unified multimodal strategy to identify cyberbullying based on a single representation of text and images combined, obviating the need for separate learning modules for images and text. They discovered that encoding information using text in comparison to visuals is a better model. By establishing a genetic algorithm-based multimodal framework employing a pre-trained VGG-16 network and convolutional neural network to extract features from images and text, Kumari et al. [23] achieved an increased F1 score of 78% over the dataset introduced in [7]. In our past work [2] we proposed Character Text Image Classifier (CTIC) to detect hate over a Twitter dataset consisting of image+text.

3 Architectures

A set of three feed-forward networks (of dimensions $1024, 512, 1$) with their respective selu (Scaled Exponential Linear Unit [3]) activation and a final sigmoid activation are used to process the resultant feature vector for every model in this section.

Text Classification:

- **GloVe text-model (GloVe-text):** GloVe[6] vector representations (both 100 and 200 dimensions) of words are used along with a single layer BiLSTM

[5] https://gombru.github.io/2019/10/09/MMHS/
[6] https://nlp.stanford.edu/projects/glove/

(100 hidden states). As our dataset is significantly small and GloVe is pre-trained over a large corpus, we don't retrain.
- **BERT text-model (BERT-text):** The pre-trained BERT-base [4] (uncased) model (BERT input dimension = 512) is fine-tuned on the training dataset, and output from the last layer of BERT is max pooled to get the sentence embedding. The sentence embedding is used as the feature vector.
- **XLNet text-model (XLNet-text):** The pre-trained XLNet-base [13] (cased) model (input dimension = 512) is fine-tuned on the training dataset, and output from the last layer of XLNet-base is max pooled to get the sentence embedding.
- **Residual-BiLSTM (ResBiLSTM) [9]:** We have experimented with the proposed model Residual-BiLSTM (ResBiLSTM) from the paper [9]. It consists of two residual blocks which can be reduced or increased as per need. The model architecture is the same as in [9] with BiLSTM layers in residual blocks having 512 hidden dimensions. The model was trained on the word embeddings of comments, generated from the Universal Sentence Encoder[7].

Image processing: For the final output, the image features extracted from the video are fed into a fully connected feed-forward neural network layer with sigmoid activation and a global average pooling layer.

EfficientNet [12] is used for the feature extraction from video frames. Fine-tuning its variants B5, B3, and B0 with our objective of hate speech classification resulted in similar performance with slight variations. Owing to lesser parameters, EffecientNet-B0 was selected as the base model for feature extraction from video frames.

Multimodal feature combination: The extracted features from individual modalities are combined together with different combination techniques for final classification.

- **Baseline Multimodal Architecture (Base-mul):** Simple concatenation is used for combining multiple modalities. Along with the max-pooled video frame features from EffecientNet-B0, the max-pooled sentence embedding (768 dimensional) from BERT-base/XLNet-base, is used as an input to the concatenation layer.
- **Modified Approach:** We make certain modifications to the baseline multimodal architecture and incorporate some architectural changes which are:
  - Along with the simple concatenation, the attention mechanism is used to combine and focus the information extracted from image features to text features and vice versa, by learning to relate a portion of the associated images with relevant words.
  - Embedding at the character level is used along with the word, and sentence embeddings, helping to learn more words with spelling mistakes, called out-of-vocabulary(OOV), and having non-trivial syntax which is common in social media posts.

[7] https://tfhub.dev/google/universal-sentence-encoder/4

Cross Attention Fusion Mechanism (CA-FM): Luong attention [8], a form of dot product attention is used in this mechanism. Some idea for this mechanism is taken from [9]. The attention is computed as

$$f_{\text{attention}}\ (T_f, I_f) = T_f I_f^T \tag{1}$$

where T_f, is the feature matrix extracted from text, and I_f is the feature matrix extracted from image frames (extracted from video). We then apply an activation function $g_{\text{activation}}$ over $f_{\text{attention}}$ as $G_{\text{a}} = g_{\text{activation}}\ (f_{\text{attention}}\ (T_f, I_f))$.

G_{a} is attended over I_f using dot product attention, and concatenated with T_f as shown in equation (2).

$$C_{\text{concat}} = Concatenate(f_{\text{attention}}\ (G_a, I_f), T_f) \tag{2}$$

Finally, we extract the cross-attended features from C_{concat} using a pooling mechanism such as global average pooling. These extracted features can be used further in a classification architecture to jointly learn from a multimodal input.

Character Embedding (CE): An embedding layer is trained over the character embeddings, helping to learn character-level vector representations. This can be processed by a 1D convolution neural network or an RNN (such as LSTM, GRU, BiLSTM, etc.). An RNN is used on the sequences, extracting the local one-dimensional patches as shown in Figure 1. By running an RNN along the character embeddings, character *n-grams* are learned which are mapped sequences of n characters that aid in capturing the morphology of words.

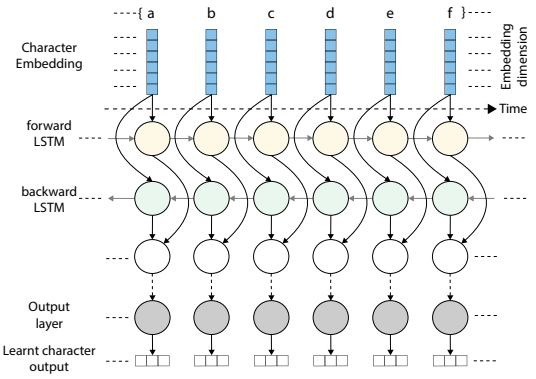

Fig. 1. BiLSTM in action with character embedding.

Modified Multimodal Architectures: Alongside the Baseline Multimodal architecture (Base-mul) we use some combinations of the modified approach (Section 3).

– **Base-mul with Character Embedding (Base-mul-CE):** The character level features are extracted using ResBiLSTM from character embeddings instead of word embeddings. We convert the two-dimensional feature matrix

output to a one-dimensional matrix by using a flatten layer. The resultant vector is used alongside the concatenation layer of Base-mul. The remaining architecture is the same as Base-mul.

– **Base-mul-CE with Recurrent-CNN and Attention Fusion (Base-mul-CE + RCNN + Att):** We have used the idea of Recurrent-CNN along with attention fusion from the paper [9]. The image features extracted from EffecientNet-B0 are attended alongside the text features through the attention fusion mechanism used in [9], and the output is fed into a Recurrent-CNN layer of two recurrent blocks. The resultant feature matrix is run through a global average pooling layer and a feed-forward layer of 1024 hidden nodes. Finally, the output from this feed-forward layer is combined with the concatenation layer in Base-mul-CE.

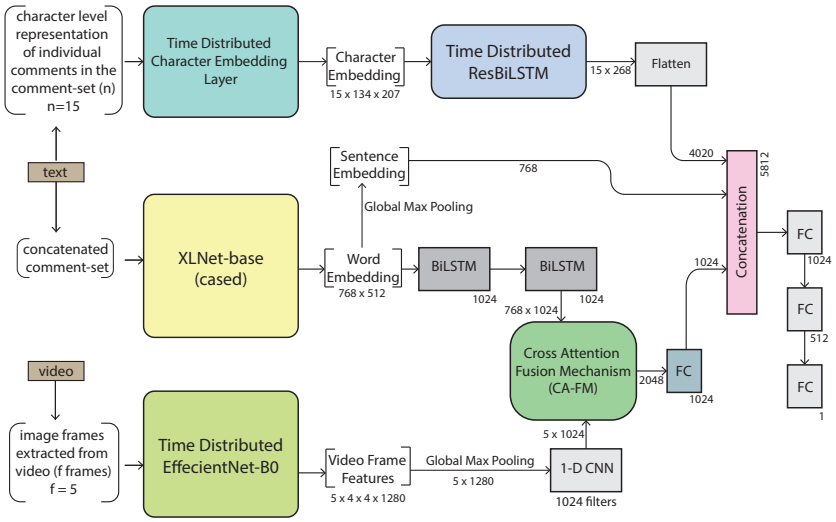

Fig. 2. Proposed Model.

Proposed Architecture: For the image feature extraction from video frames, we use EffecientNet-B0, XLNet-base (cased) for extracting text features from comments, and an embedding layer(learnable) for obtaining the corresponding character level vector representation. The architecture(Figure 2) has the following parts:

– **Sentence Embeddings**: The output from the last layer of XLNet-base (cased) is max pooled to get the sentence embedding O_{sent} (feature vector of 768 dimensions).
– **Processing of characters**: The character-level vector representation (embedding) is passed through a ResBiLSTM layer of two residual blocks, learning character level representations. The character embedding layer has an embedding matrix of size 208×207, which is created from the vine dataset.

The BiLSTM layers inside the Residual-BiLSTM have a dimension of 134, which is the maximum number of characters in the input. The two-dimensional output matrix from the Residual-BiLSTM is converted to a one-dimensional vector O_{char}, using a flatten layer.

- **Video frame processing**: A global average pooling is used over the video frame features extracted from EffecientNet-B0 to get an output feature vector of 1280 dimensions for each frame. This is passed through a 1-D convolution of 1024 hidden dimensions to get an output matrix O_{image} having 1024 feature dimensions for each frame.
- **Cross Attention Fusion Mechanism**: Upon permuting the word embeddings (shape = 512×768) from XLNet-base (cased) we obtain a tensor(text feature matrix) of shape 768×512 which is used alongside the image feature matrix O_{image} in the cross attention fusion mechanism (CA-FM). The context feature matrix from CA-FM is fed into a fully connected feed-forward network of 1024 hidden nodes with a LeakyReLU activation function to obtain a vector O_{CA-FM} of 1024 dimensions.

We concatenate O_{char}, O_{sent} and O_{CA-FM} as

$$O_{\text{concat}} = Concatenate\,(O_{char}, O_{sent}, O_{CA-FM}) \tag{3}$$

to get a feature vector O_{concat} of 5812 dimensions. O_{concat} is fed into three fully connected networks as shown in Figure 2, each having a selu activation layer with corresponding 10% drop-out. Sigmoid activation is used for obtaining the class prediction. We name this model as **Text Video Classifier (TVC)**[8].

Data Sampling: For training multimodal architectures we correct the problem of class imbalance in the dataset and analyze its effect.

- Class Weights (CW): The class weighs are applied while computing the loss function where each class's weights are computed as follows: $w^i = n_{obs}/tn^i_{obs}$ where w^i represents the weight to class i, n^i_{obs} represents the number of observations in class i, n_{obs} represents the number of observations, and t represents the total number of classes. This aids the loss function in penalizing the minority class for misclassification by increasing class weight while decreasing weight for the majority class.
- Random Oversampling (OverS): Populating the training dataset by choosing samples (with replacement) from the minority class at random from the training dataset and replicating them. The oversampling ratio[9] is expressed as $\alpha_{os} = N_{rm}$ (instances in minor class) / N_M (instances in the major class).
- Undersampling + Oversampling (UnderS + OverS): Random undersampling is the removal of examples from the majority class in the training dataset at random. But the random removal of instances from the training dataset also excludes those instances which are critical for training the models to learn the

[8] https://github.com/NishchalPrasad/Text-Video-Classifier
[9] https://imbalanced-learn.org/stable/index.html

decision boundaries. As a result, we combine the method of undersampling and oversampling to have a balance between the classes, as follows:
With $\alpha_{os} = 0.5$, the minor class is oversampled, followed by the major class being undersampled, yielding a $1 : 1$ proportion of $minor\ class : major\ class$.

4 Model Training and Experimental Setup

4.1 Dataset

Table 1. Vine dataset statistics

	Count
Media sessions	970
-hate	304
-non hate	666
Media sessions with missing video files	130
-hate	35
-non hate	95
Number of comments	78250

To test our approach we used a multimodal dataset consisting of videos with their associated captions and comments. Rahat et al. [10] contributed their dataset for this study. Vine was a social networking site in the United States that allowed users to create looping video snippets of up to six seconds in length. It was established in June 2012. Before its formal release on January 24, 2013, it was acquired by Twitter, an American microblogging service. The data was collected using the snowball sampling [15] method, and a profanity test [16] was applied to choose media sessions based on a percentage of comments containing profanity. There are at least 15 comments linked with each video in the dataset's examples. In the dataset, there are 969 media sessions, some of which are missing videos. The dataset statistics can be found in Table 1. We first focused on the unimodal(textual) approach to detect hate speech and tried to develop techniques to improve textual inference models, which will be further used to develop multimodal classification architectures. We split the dataset into test set (97 examples) and use the remaining for training our models.

Hyperparameter settings: We experimented with the chronological comment-set (of 5, 10, or 15 comments) for the unimodal architectures, where we concatenate the comments in the comment-set to form a single sentence separated by a full stop (.) punctuation and remove the emojis present. The emojis needed to be removed because BERT-base, and XLNet-base, which we use, are not trained on emojis. The emojis are taken into account through their character-level representations(character embeddings). We take a comment-set and use it

along a TimeDistributed[10] layer for input to the character embedding layer. For the multimodal approach, we analyze their performance on a comment-set of 15 comments to capture more contextual information as compared to 5 or 10 comment-sets (Table 2). We have extracted 5 frames from the videos and used them along a TimeDistributed[10] layer as input to EffecientNet-B0 to jointly extract individual frame features into a feature matrix. Binary Cross-Entropy is used as the loss function, Adam with weight decay (AdamW)[11] is used as the optimizer, with weight decay as $1e^{-4}$ and learning rate of $3e^{-5}$. The output prediction probability threshold is set to default, i.e., 0.5. Section 3 details the respective architectures. All of the sampling procedures discussed in Section 3 are used for training the multimodal architectures and the best results are reported in Table 3.

5 Experimental Results

For measuring the predictive performance of the model we have used the area under the Receiver Operating Characteristic curve (AUC-ROC), mean accuracy, and F1 score.

Table 2. Unimodal experimental results

	Accuracy (%)			AUC			F1 score		
comments set	5	10	15	5	10	15	5	10	15
GloVe-text									
100 dimensions	65.8	67.71	69.9	0.6448	0.659	0.754	0.496	0.5332	0.5782
200 dimensions	65.48	67.83	70.0	0.6457	0.66	0.7632	0.503	0.54	0.5811
BERT-text	67.45	69.23	75.0	0.6858	0.745	0.7788	0.532	0.549	0.6163
XLNet-text	65.38	73.08	76.92	0.7102	0.7454	0.7706	0.522	0.5468	0.625
ResBiLSTM	63.7	67.94	75.0	0.632	0.7	0.7126	0.48	0.5294	0.5806
EffecientNet-B0	68.0			0.578			0.423		

BERT-base and XLNet-base architectures outperform alternative unimodal architectures, as shown in Table 2, and we choose the best performing of the two in our final model. Character embedding boosts the performance in Base-mul-CE and though XLNet-base covers a large set of out-of-vocabulary words, several words are not captured. This is because all the comments in the dataset are noisy with misspellings, OOV words, various representations of a word, and emojis. "Base-mul-CE + RCNN + Att" performs poorly over the sampling strategies which is because of the Recurrent-CNN layer applied to the context vector obtained after CA-FM. For our model, this creates much extra noise in the context vector thus losing information after several recurrent convolutions.

[10] Keras TimeDistributed layer, https://www.tensorflow.org
[11] AdamW, https://www.tensorflow.org

Table 3. Multimodal experimental results (15 comments)

	Sampling technique	Accuracy (%)	AUC	F1 score
Base-mul	CW	75.85	0.8090	0.621
	OverS	75.3	0.81	0.638
Base-mul-CE	OverS	**80.21**	0.8434	0.678
Base-mul-CE + RCNN + Att	OverS	73.20	0.6701	0.5357
	OverS + UnderS	74.3	0.7023	0.605
Proposed Architecture (TVC)	OverS + UnderS	78.12	**0.8904**	**0.7342**
ResBiLSTM-RCNN [9]	No sampling as per [9]	76.23	0.7751	0.68

Hence we removed the Recurrent-CNN layer in our proposed model and used the context vector from CA-FM directly in the final concatenation layer, attaining a considerable increase in the performance metrics when trained over the OverS + UnderS sampled training set. This shows that training the model on a sampling technique may help the model to attend to the minority class and helps in the optimal training of a multimodal model. Also, a suitable attention mechanism to combine features from different modalities helps the model map features from one modality to the other.

The effect of sampling can be seen in Table 3 with oversampling (OverS) achieving better performance than simple class weights (CW) over Base-mul. So we choose OverS and experimented further with including undersampling too alongside (OverS+UnderS), and as can be seen with "Base-mul-CE + RCNN + Att" reducing the samples from the dominant class in the dataset helps the model generalize more and gives better performance in both the minority and majority class. This shows that a sampling technique tailored to the class imbalance in the dataset has a good effect on the generalization of our multimodal model's prediction overall, leading to a better performance.

Error Analysis: Upon manually checking classification errors for some of the instances, we found out the following few reasons.

- Instances referring to events that are not in the context of any other example of the dataset were incorrectly categorized. For example, if a comment mentions something regarding a topic that is not mentioned in the previous comments, and this topic intends to be hateful in nature.
- Usage of the slur: such as the words "cunt","nigger", or "'nigga'" which take place inside a society without the goal of causing harm, resulted in wrong classification.
- Presence of abusive or swear words in a sentence intended to imply sarcasm. The model fails to understand sarcasm-oriented texts which involve abusive words.
- Absence of any hate word, though the text implies hate: The model learns to identify abusive and swear words to predict hate or non-hate, and when such words are missing in a hate speech comprised of non-abusive words the model misclassifies it to be non-hate.

6 Future work and Conclusion

Our work takes into account the hate speech classification problem in a multimodal setup of videos and texts combined, by utilizing the techniques from deep learning-based methods. For this, we experimented with several architectures(both unimodal and multimodal in nature) with different sampling methods. The idea of cross-attention fusion to jointly learn from the features of individual modalities is also used, which significantly boosts the model's performance. This work will be further extended to the problem of multimodal hate speech prediction in Indian literature.

Acknowledgments: Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India for carrying out this research.

References

1. S. Ghosh, A. Ekbal, P. Bhattacharyya, T. Saha, A. Kumar and S. Srivastava, "SEHC: A Benchmark Setup to Identify Online Hate Speech in English," in IEEE Transactions on Computational Social Systems, vol. 10, no. 2, pp. 760-770, April 2023.
2. Prasad, Nishchal, et al. "A Multimodal Classification of Noisy Hate Speech Using Character Level Embedding and Attention." 2021 International Joint Conference on Neural Networks (IJCNN), 2021, pp. 5998–8.
3. G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-Normalizing Neural Networks," in Advances in Neural Information Processing Systems, 2017, vol. 30.
4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota.
5. Dinakar, K., Reichart, R., and Lieberman, H. (2021). Modeling the Detection of Textual Cyberbullying. Proceedings of the International AAAI Conference on Web and Social Media, 5(3), 11-17.
6. Gomez, R., Gibert, J., Gómez, L., and Karatzas, D. (2019). Exploring Hate Speech Detection in Multimodal Publications. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 1459-1467.
7. Kumari, K., Singh, J., Dwivedi, Y.K., Rana, N.P.: Towards cyberbullying-free social media in smart cities: a unified multi-modal approach. Soft Computing 24,11059–11070 (2020)
8. Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421, Lisbon, Portugal.
9. Paul, S., Saha, S., Hasanuzzaman, M.: Identification of cyberbullying: A deep learning based multimodal approach. Multimedia Tools and Applications pp. 1–20 (092020)

10. Rafiq, R.I., Hosseinmardi, H., Mattson, S.A. et al. Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network. Soc. Netw. Anal. Min. 6, 88 (2016).
11. Reynolds, K., Kontostathis, A., Edwards, L.: Using machine learning to detect cyberbullying. In: 2011 10th International Conference on Machine Learning and Applications and Workshops. vol. 2, pp. 241–244 (2011)
12. Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research), PMLR, 6105–6114.
13. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Advances in Neural Information Processing Systems, Curran Associates, Inc.
14. Zhong, H., Li, H., Squicciarini, A., Rajtmajer, S., Griffin, C., Miller, D., Caragea, C.: Content-driven detection of cyberbullying on the instagram social network. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. p. 3952–3958. IJCAI'16, AAAI Press (2016)
15. L. A. Goodman, "Snowball Sampling," The Annals of Mathematical Statistics, vol. 32, no. 1, pp. 148 – 170, 1961.
16. H. Hosseinmardi, S. A. Mattson, R. Ibn Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network," in Social Informatics, T.-Y. Liu, C. N. Scollon, and W. Zhu, Eds. Cham: Springer International Publishing, 2015, pp. 49–66.
17. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion (2017).
18. Kumari, K., Singh, J.P., Dwivedi, Y.K., Rana, N.P.: Aggressive social media post detection system containing symbolic images. Digital Transformation for a Sustainable Society in the 21st Century. pp. 415–424. Springer International Publishing, Cham (2019)
19. Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N.: Hate speech detection with comment embeddings. In: Proceed- ings of the 24th International Conference on World Wide Web. p. 29–30. WWW '15 Companion, Association for Computing Machinery, New York, NY, USA (2015).
20. Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14). JMLR.org, II–1188–II–1196.
21. Soujanya Poria, Erik Cambria, Amir Hussain, Guang-Bin Huang. Towards an intelligent framework for multimodal affective data analysis. Neural Networks, Volume 63, 2015, Pages 104-116, ISSN 0893-6080
22. Cambria, E., Howard, N., Hsu, J., Hussain, A.: Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics. In: 2013 IEEE Symposium on Computational Intelligence for Human-like Intelligence (CIHLI). pp. 108–117 (2013)
23. Kumari, K., Singh, J.: Identification of cyberbullying on multi-modal social media posts using genetic algorithm. Transactions on Emerging Telecommunications Technologies 32 (02 2020).