

# Cyberbullying detection on multi-modal data using pre-trained deep learning architectures

*Detección de acoso cibernético en datos multimodales utilizando arquitecturas de aprendizaje profundo pre-entrenado*

*Detecção de cyberbullying em dados multimodais usando arquiteturas de aprendizagem profunda pré-treinadas*

Subbaraju Pericherla<sup>1</sup>  
E Ilavarasan<sup>2</sup>

**Received:** May 10<sup>th</sup>, 2021

**Accepted:** July 30<sup>th</sup>, 2021

**Available:** September 6<sup>th</sup>, 2021

**How to cite this article:**

S. Pericherla, E. Ilavarasan, "Cyberbullying Detection on Multi-Modal Data Using Pre-Trained Deep Learning Architectures," *Revista Ingeniería Solidaria*, vol. 17, no. 2, 2021. doi: <https://doi.org/10.16925/2357-6014.2021.03.09>

---

Research article. <https://doi.org/10.16925/2357-6014.2021.03.09>

<sup>1</sup> Research Scholar, Department of CSE, Pondicherry Engineering college, Puducherry, India- 605014

Email: [raju.pericherla74@gmail.com](mailto:raju.pericherla74@gmail.com)

**ORCID:** <https://orcid.org/0000-0002-0701-6377>

<sup>2</sup> Professor, Department of CSE, Pondicherry Engineering college, Puducherry, India- 605014

Email: [eilavarasan@pec.edu](mailto:eilavarasan@pec.edu)

**ORCID:** <https://orcid.org/0000-0001-6258-2243>



## Abstract

*Introduction:* The present article is the product of the research "Cyberbullying Detection on Multi-Modal Data Using Pre-Trained Deep Learning Architectures.", developed at Pondicherry Engineering College in the year 2020.

*Problem:* Identification of cyberbullying activities on multi-modal data of social media

*Objective:* To propose a model that can identify cyberbullying activity for text and image data.

*Methodology:* This paper has extracted the features of using two pre-trained architectures for text data and image data, in order to identify cyberbullying activities on multi-modal data, concatenated text features and image features, before supplying them as inputs to the classifier.

*Results:* An analysis has been performed on the proposed approach implemented on multi-modal data with Recall, and F1-Score as measures. Grad-cam visualization is presented for images to show highlighting regions.

*Conclusion:* The results indicate that the proposed approach is efficient when compared with the baseline methods.

*Originality:* The proposed approach is effective and conceptualized to improve cyberbullying detection on multi-modal data.

*Limitations:* There is a need to develop a model which can identify bullying graphical images and videos

**Keywords:** Cyberbullying, Deep learning, Xception, RoBERTa, Natural language processing, Social Media.

## Resumen

*Introducción:* El presente artículo es producto de la investigación "Detección del ciberacoso en datos multimodales utilizando arquitecturas de aprendizaje profundo pre-entrenadas", desarrollada en Pondicherry Engineering College en el año 2020.

*Problema:* Identificación de actividades de ciberacoso en datos multimodales de redes sociales.

*Objetivo:* Proponer un modelo que pueda identificar la actividad de ciberacoso para datos de texto e imágenes.

*Metodología:* Este artículo ha extraído las características del uso de dos arquitecturas previamente entrenadas para datos de texto e imágenes, con el fin de identificar actividades de ciberacoso en datos multimodales, características de texto concatenadas y características de imagen, antes de suministrarlas como entradas al clasificador.

*Resultados:* Se ha realizado un análisis sobre el enfoque propuesto implementado en datos multimodales con Recall y F1-Score como medidas. La visualización Grad-cam se presenta para que las imágenes muestren regiones destacadas.

*Conclusión:* Los resultados indican que el enfoque propuesto es eficiente en comparación con los métodos de referencia.

*Originalidad:* el enfoque propuesto es eficaz y está conceptualizado para mejorar la detección del ciberacoso en datos multimodales.

*Limitaciones:* existe la necesidad de desarrollar un modelo que pueda identificar imágenes gráficas y videos de intimidación

**Palabras clave:** Cyberbullying, Deep learning, Xception, RoBERTa, Procesamiento del lenguaje natural, Social Media.

## Resumo

**Introdução:** O presente artigo é o produto da pesquisa “Detecção de cyberbullying em dados multimodais usando arquiteturas de aprendizado profundo pré-treinadas”, desenvolvida na Pondicherry Engineering College no ano 2020.

**Problema:** Identificação de atividades de cyberbullying em dados multimodais de mídia social

**Objetivo:** propor um modelo que possa identificar a atividade de cyberbullying para dados de texto e imagem.

**Metodologia:** Este artigo extraiu as características do uso de duas arquiteturas pré-treinadas para dados de texto e dados de imagem, a fim de identificar atividades de cyberbullying em dados multimodais, recursos de texto concatenados e recursos de imagem, antes de fornecê-los como entradas para o classificador.

**Resultados:** Foi realizada uma análise da abordagem proposta implementada em dados multimodais com Recall e F1-Score como medidas. A visualização Grad-cam é apresentada para imagens para mostrar as regiões de destaque.

**Conclusão:** Os resultados indicam que a abordagem proposta é eficiente quando comparada aos métodos basais.

**Originalidade:** a abordagem proposta é eficaz e conceituada para melhorar a detecção de cyberbullying em dados multimodais.

**Limitações:** É necessário desenvolver um modelo que possa identificar imagens gráficas e vídeos de bullying

**Palavras-chave:** Cyberbullying, Deep learning, Xception, RoBERTa, Processamento de linguagem natural, Social Media.

## [1] INTRODUCTION

Social media networks such as Facebook, Twitter, Instagram etc... assembled a large audience at one place irrespective of boundaries all over the world. According to the statistics reported by Statista\*, a web-based report, the current social media population stands at 4.14 billion, which is more than half of the total world population. On one side, there is popularity and increase in the number of users on the social media platforms, on the other side, it also attracts illicit, criminal, unlawful activities executed by illegitimate users, such as online hate speech, online trolling, cyberbullying etc. One of the most harmful activities that affect the teenagers and youth of social media is cyberbullying. Teenage girls are more affected victims than boys (34.5% boys and girls 38.5%)[1]. Cyberbullying happens when the digital platforms are used as a medium to bully someone through shaming, degrading and demeaning which can lead to mental breakdowns. It creates severe psychological disorders [2] and sometimes leaves the victims with suicidal tendencies. Since 2010, the instances of cyberbullying have increased rapidly as more children become targets of bullying. Cyberbullying is a more heinous crime than traditional bullying as it happens anytime and anywhere. One of the biggest challenges faced when dealing with cyberbullying is its

rapid propagation. Many times, cyberbullying victims do not to share their experiences with others, which creates another problem.

With the rapid increase of digital technologies over the last two decades, netizens share their opinions through different formats such as texting, images, videos, and emojis. Due to the existence of abundant forms of expression of data, cyberbullying becomes a challenging task. Cyberbullying text messages might involve short texts, misspelt texts, texts with embedded symbols. Cyber-bullying through images might comprise complex facial expressions, animals, or some embarrassing images [13]. Other modes of cyberbullying are encompassed using a combination of text and image, image and video, text and emoji, etc. Most of the research works pursued in the literature place an emphasis on cyberbullying detection with text data only. Only a limited number of studies are in existence on cyberbullying detection using image data and multi-modal data. These scenarios motivated researchers to mitigate cyberbullying activities on social media for multi-modal data.

The contribution to this research work is twofold.

1. To propose a neural network-based method to handle multi-modal data for cyberbullying detection.
2. To perform an extensive analysis on multi-modal dataset.

The rest of the paper is organized as follows. Section 2 describes related works in cyberbullying detection. Section 3 details the methodology adopted for the proposed work. Section 4 deals with datasets and the experimental results. Section 5 concludes the work with directions on future work.

## [2] RELATED WORKS

In this section, the earlier works related to cyberbullying detection with text data and image data are discussed.

Reynolds et al [3] used language-based methods to identify cyberbullying messages. The authors crawled the data from Formspring.me website. They manually labeled the tweets with the help of Amazon's Mechanical Turk service, based on the bad words in each record. They used 'NUM' and 'NORM' as two features to train the model. The feature 'NUM' represents the number of bad words in a message and 'NORM' represents intensity of the bad words. Their proposed model was able to accurately identify 78% of bullying posts. Yin et al [4] employed a supervised learning approach

for the detection of online harassment. The datasets were collected from three popular social media websites; namely MySpace, Konegragate, and Slashdot. They used three special features called sentiment features, contextual features, and local features. Support Vector machine with linear kernel is used for the classification task. The proposed model with special features outperformed the basic term frequency and inverse document frequency (tf-idf) models. Rusa et al [5] implemented three deep learning architectures viz., a simple CNN (convolutional neural network), a hybrid-CNN-LSTM (Long Short Term Memory) and a mixed CNN-LSTM-DNN for cyberbullying detection problem. The dataset was collected from Formspring.me website. A Word2Vec word embedding technique is used to generate features from the input. Bu S and Cho S [6] (2018) proposed an ensemble-based hybrid deep learning system that consists of two deep learning models. They used CNN to capture the information at the syntactic level and LRCN (Long term Recurrent Convolutional Network) to capture semantic level information. The dataset was collected from Kaggle Website. The proposed model achieved 87% accuracy when identifying cyberbullying. They used the t-SNE algorithm to analyze the relation between semantic and syntactic models. Van Bruwane et al [7] (2020) created a multi-platform dataset called VISR dataset for cyberbullying and cyber aggression. They collected more than half a million posts from seven different social media platforms by various methods and compared them with different machine learning models. CNN and XGBoost outperformed other techniques in general and support vector machine (SVM) in specific. Liu et al [8] (2020) proposed a hybrid neural network to detect emotional analysis based on text data. They implemented a model with two deep learning architectures: 1) CNN to extract local features from the text and 2) Bi-LSTM to extract global features from the text. They used word2vec word embeddings for training corpus. The experimental results were tested on the popular IMDB dataset. The proposed models attained better accuracy than single CNN and Bi-LSTM models. Shylaja SS et al [9] (2018) conducted several experiments on cyber aggressive comments using recurrent neural networks such as simple RNN, LSTM, Bi-LSTM, and GAN. A Doc2vec word embedding technique is used for feature representations. The data was collected from Kaggle and various GitHub repositories. LSTM achieves the highest accuracy (93%), surpassing the other methods. Frederick F et al [10] (2019) presented a study that suggested an unsupervised algorithm to find the patterns between cyberbullying words and keywords from unstructured data. The data was extracted from Twitter by supplying frequent keywords used by the teenagers. The FP-Growth association rule algorithm was applied to find relations between the keywords. J. de-la-Peña-Sordo et al [11] (2016) proposed a novel method to detect online trolling comments from social media websites. Authors used a combination of

syntactic and opinion features for identification of troll comments. The proposed approach was evaluated on a dataset from 'Meneame', a popular Spanish news website, Gutiérrez-Esparza et al [12] (2019) classified cyberbullying in social media by applying OneR, Variable Important measures (VIMs) and Random Forest to classify cyber aggression comments. The comments were collected from Facebook social networking site. They used the VIM process to identify key features in cyber aggression. Haoti Zhong et al [13] (2016) presented a study on cyberbullying detection with images that are shared on Instagram; another popular social networking site. Over 3000 images from Instagram were crawled using API to conduct the experiments. They used feature extraction techniques such as Bag of Words (BoW), Word2vec for text data and CNN for image data. R Zhao et al [14] (2017) proposed a novel method called Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA) to tackle cyberbullying attacks. The smSDA method was able to learn efficient features from BoW representation. The datasets for experiments were taken from Twitter and MySpace. Akshi kumar and Nitin Sachedeva [15] studied the significance of soft computing techniques for cyberbullying detection. Zhang X et.al [16] proposed novel convolutional neural network based on pronunciation (PCNN). Threshold movement, cost function adjustment and yielding hybrid solutions were utilized to handle class imbalance problems. Cyberbullying datasets from Twitter and Formspring.me social networking sites comprised the corpus. Tripathi K et al [17] proposed an ALBERT-based fine-tuning model for cyberbullying detection; the model achieved better performance than GRU and BERT. Sayanta Paul and Sriparna Saha [18] presented an innovative method of BERT for cyberbullying identification. The proposed method was able to classify bullying tweets on three real world datasets. K kumari et al [19, 20] proposed a single layer convolutional neural network for cyberbullying identification on multi-modal data. The proposed method achieved 74% of recall on bullying class. The genetic algorithm was used to reduce the features and further improved 9% of weighted F1-score. Satya Prakash Yadav and Sachin Yadav [25] explored fusion techniques for multimodal data on medical images. They discussed the involvement of various medical entities like medical resonance imaging (MRI), positron emission tomography (PET), and computed tomography (CT).

From the above literature survey, we have identified research gaps. Most of the research work carried out for cyberbullying detection is at the text level only. Very few studies have been done on cyberbullying detection with image data. Word embedding techniques are key features for any natural language processing tasks, but in the case cyberbullying detection approaches, most of the research carried out has used basic approaches like BoW and tf-idf techniques. A few studies used next-level advanced

word embeddings techniques like Word2vec, Glove and Doc2Vec. These word embedding techniques increase the computation time and complexity of the system when training the model. To fill these gaps, we proposed a model to identify cyberbullying for multi-modal data.

### [3] METHODOLOGY

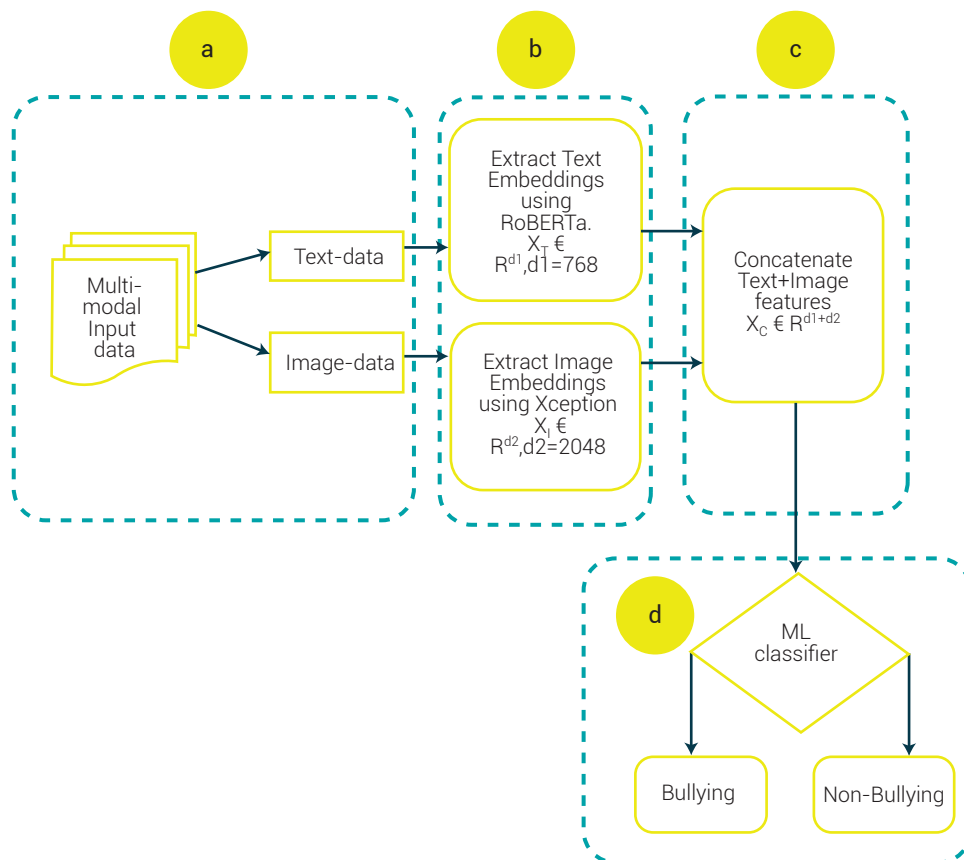
The proposed architecture is presented and discussed in detail in the preceding section. The proposed technique to detect cyberbullying in text and image data fundamentally consists of four components as illustrated in Fig. 1 viz., a) Input data, b) Pre-processing, c) Proposed approach and d) Classification.

#### *Problem statement*

Let  $T$  and  $I$  represent text and images of a given tweet or comment. Let  $X_T \in \mathbb{R}^{d_1}$  and  $X_I \in \mathbb{R}^{d_2}$  denote the features of text and images respectively, the goal is to design a model that can predict the label of a given comment using features  $X_I$  and  $X_T$ . This is formulated as a minimization loss function as provided below in Equation-1.

$$\min f(\theta) = f(y | X_I, X_T; \theta) \quad (1)$$

where  $\theta$  represents the model parameters and  $y$  symbolizes the combined label.



**Fig. 1** Architecture of Proposed method





Source: own work

### *a) Input data description*

The dataset employed in [19] is utilized as input data, which is an amalgamation of data from different social media platforms; namely Facebook, Instagram, and Twitter. Finally, 2100 samples were collected for training the model. Each record comprises two fields namely text comment and image. The combination of text and image yields six possible cases for bullying and non-bullying records. Table-1 enumerates the four possible cases for bullying when the text and image data are conjointly used. Table-2 shows the two possible cases for non-bullying when combined text and image together. For the text comments, 884 records are labelled as bullying and 1216 records are labelled as non-bullying. For the images, 464 are labeled as bullying images and 1636 are labelled as non-bullying images. For combined data, 1481 are labelled as bullying and 619 are labelled non-bullying. Table-3 gives an overview of the data samples.





**Table 1.** Four Possible Cases for Bullying and for Combined Data

Case	Comment Text	Image	Combined Text +Image
Case 1	Comment text: <b>Bullying</b> Example: You can't enter the gate Fatty acid	Image: <b>Bullying</b> 	Combined: <b>Bullying</b>
Case 2	Comment text: <b>Bullying</b> You are like this trees which have nothing.	Image: <b>Non-Bullying</b> 	Combined: <b>Bullying</b>
Case 3	Comment text: <b>Non Bullying</b> you're so tall, you remind me of this.	Image: <b>Non-Bullying</b> 	Combined: <b>Bullying</b>
Case 4	Comment text: <b>Non-Bullying</b> Stop Eating Fatty	Image: <b>Bullying</b> 	Combined: <b>Bullying</b>

**Source:** own work

**Table 2.** Two possible cases for Non-bullying for combined data

Cases	Comment Text	Image	Combined Text +Image
Case 1	Comment text: <b>Non-Bullying</b> You are not fat you are just chubby.	Image: <b>Bullying</b> 	Combined: <b>Non-Bullying</b>
Case 2	Comment text: <b>Non-Bullying</b> I am taller than you.	Image: <b>Non-Bullying</b> 	Combined: <b>Non-Bullying</b>

Source: own work

**Table 3.** Overview of the input data

	Text data	Image data	Combined Text+ Image
Non-Bullying	1216	1636	1481
Bullying	884	464	619
Total	2100	2100	2100

Source: own work

### *b) Pre-processing*

The data pre-processing is an essential procedure performed while building a machine learning model. The quality of application of pre-processing techniques affects the results of prediction. In this phase, various existing pre-processing techniques such as lemmatization, stemming and stop word removals are applied to clean and structure the text data. Various image pre-processing methods such as image segmentation and edge enhancements are utilized to clean the image data.

### *c) Proposed Approach*

Most of the research works conducted for cyberbullying detection principally focus on text-based features. Though text-based data serves as a primary source of information to classify the comments as bullying and non-bullying, a more refined and richer representation would improve the detection system to conduct the classification more precisely. Hence, a robust detection system may require the presence of several modalities to resolve ambiguities, if any, that may arise by inferring the multiple-modal information fusion. A multimodal detection system normally performs better than any one of its individual components. With this hypothesis, we propose a novel approach for cyberbullying detection by leveraging the information from two modalities; i.e., the text and images. The input data comprises both the text and images for any given tweet or comment and is provided with a binary label, either as bullying or non-bullying. In the proposed approach, the features are first extracted from the text using RoBERTa, a neural network architecture, and the Xception model to extract the features from images. The contribution of the proposed approach is that it leverages the information from both the text and images to enhance the performance of the system. Also, we demonstrate the neural networks attention on images that help the model to decide bullying vs non-bullying.

#### *RoBERTa to Extract the Text Embeddings*

RoBERTa [22] is an enhanced version of Bi-directional Encoder Representations from Transformers (BERT) [23] language model by modifying a hyperparameter of BERT, namely the Next Sentence Pre-training (NSP) objective, and trained with a larger number of datasets and learning rates. The RoBERTa base uses 12 layer, 768 hidden, 12 heads and 125 million parameters. It is built upon the dynamic masking strategy of BERT model. It uses 160GB of text for training with various databases like English Wikipedia, Books and News datasets whereas BERT uses 16GB for training the model. Further, in this work, a total of 786 dimension vectors for feature extraction are used.

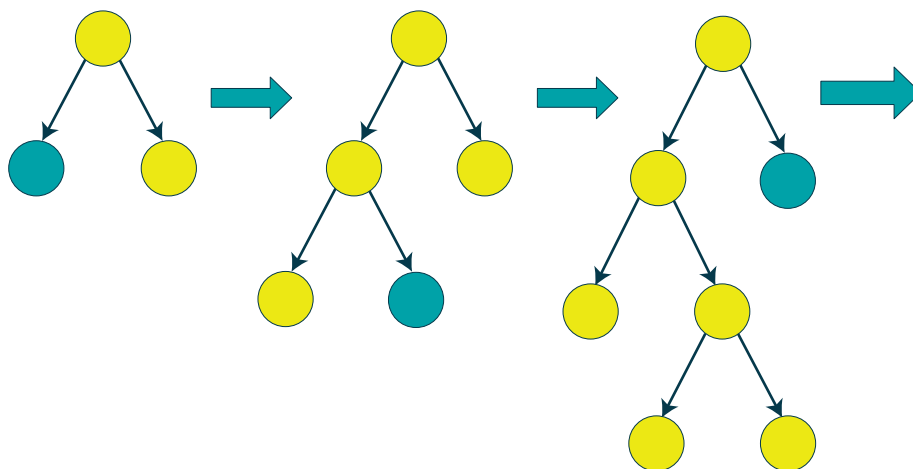
#### *Xception Model to Extract the Image Embeddings*

The Xception model [21] stands for extreme version of Inceptions and is fundamentally designed based on modified depth-wise separable convolution. This design is derived from inception-v3 networks motivation, where inception blocks are introduced in the model. The architecture consists of 36 convolutional layers that are structured into 14 modules. These layers are basically a stack of depth-wise separable convolutional

layers with residual connections. In this work, the pre-trained Xception network which is trained on Imagenet databases that have 1.2 million images of 1000 classes, is used. The structure has 2048 dimensions and upon providing the PIP image, the activations are extracted from the layer before reaching the final soft-max layer.

#### *d) Machine Learning classifier*

A Light Gradient Boosting Machine (LightGBM) classifier is employed in the proposed approach to classify the tweets as bullying or non-bullying. LightGBM grows the tree vertically, while other tree algorithms grow trees horizontally. Leaf-wise algorithms can reduce losses over level-wise algorithms. It will choose the leaf with max delta loss to grow. The LightGBM classifier is trained with text feature vectors generated by RoBERTa and image feature vectors generated by Xception. The LightGBM has overcome several shortcomings of Gradient Boosting Decision Tree and Decision Tree algorithms. Hence, it is very useful to draw greater insights from Cyberbullying Data. The LightGBM classifier follows leaf-wise tree growth which is advantageous for the classification task. Fig.2 shows the leaf-wise growth in LightGBM classifier. 1000 estimators (Sequential Decision Trees) with maximum depth of 5 were set as input parameters for LightGBM classifier. Each node indicates a decision to classify the tasks.



**Fig 2.** Leaf-wise growth in Light GBM.

Source: own work

**Table 4.** Algorithm of the proposed classification scheme

<b>Input:</b>	Raw-text and Image of Tweets or comments.
<b>Output:</b>	Binary Output – Bullying or Non-bullying
<i>Step 1</i>	Extract the embeddings for the text using RoBERTa [22] – $X_t$
<i>Step 2</i>	Extract the embeddings for the images using Xception [21] – $X_i$
<i>Step 3</i>	Concatenate $X_t$ and $X_i$
	$X_c = \begin{bmatrix} X_t \\ X_i \end{bmatrix}$
<i>Step 4</i>	$\min f(\theta) = \min f(y, f_{\theta}(X_c))$ $= \min \ y - f_{\theta}(X_c)\ _2^2 \text{ such that } \ R\ _2 = 1.$
<i>Step 5</i>	Test: Given test text ( $Z_t$ ) and Image ( $Z_i$ ), $Y_{\text{Prediction}} \leftarrow \text{Model}_{\theta}(Z_t, Z_i)$

**Source:** own work

## [4] RESULTS

The following section presents the experimental methodology, metrics employed for evaluation and the respective outcomes. The proposed technique is implemented in Python using packages such as Numpy, Pandas, matplotlib, Scikit-Learn, LightGBM, and Tensorflow in the Linux operating system. The methods were run on an Intel i7 8th Gen 12core CPU processor and Nvidia Max-Q 1070 32GB RAM.

### 4.1 Evaluation metrics

To evaluate the proposed approach, the evaluation metrics such as precision, recall and F1-score are considered.

The precision is defined as the ratio of correct predictions of bullying to total number of predictions of bullying and is calculated as given in Eq. (2).

$$P = \frac{C}{C+B} \quad (2)$$

where C indicates the number of correct predictions of bullying, B indicates the number of non-bullying classes that are incorrectly classified as bullying.

The recall is defined as the ratio of correct predictions of bullying to the total number of actual bullying classes and is given in Eq. (3)

$$R = \frac{C}{C+NB} \quad (3)$$

Where NB refers to the total number of instances of actual bullying classes wrongly predicted as non-bullying

F1-score is the weighted average of recall and precision and is computed as given in Eq. (4)

$$\text{F1-score} = \frac{2 * P * R}{(P + R)} \quad (4)$$

## 4.2 Experiment methodology

We conducted experiments with no sampling, over-sampling and under-sampling modes using five-fold cross validation. The proposed approach achieved a weighted average F1-score of 80% as compared to existing approaches. The proposed method is able to achieve recall and F1-scores of 92% and 86% for bullying class respectively. Fig. 3 shows recall of bullying class which is able to classify bullying tweets 13% more efficiently than existing methods and Fig. 4 shows F1-scores for bullying class which is 6% more efficient compared to the existing approaches. Fig. 5 shows the weighted F1-score. Table 4 shows experimental results of Kumari et.al, [20]. The authors initially proposed a CNN architecture to identify bullying on multi-modal data. Later, they proposed a genetic algorithm to find the best features from text and image data and be able to classify the multi-modal bullying data efficiently.

**Table 4.** Precision, Recall and F1-score on different features on CNN [18]

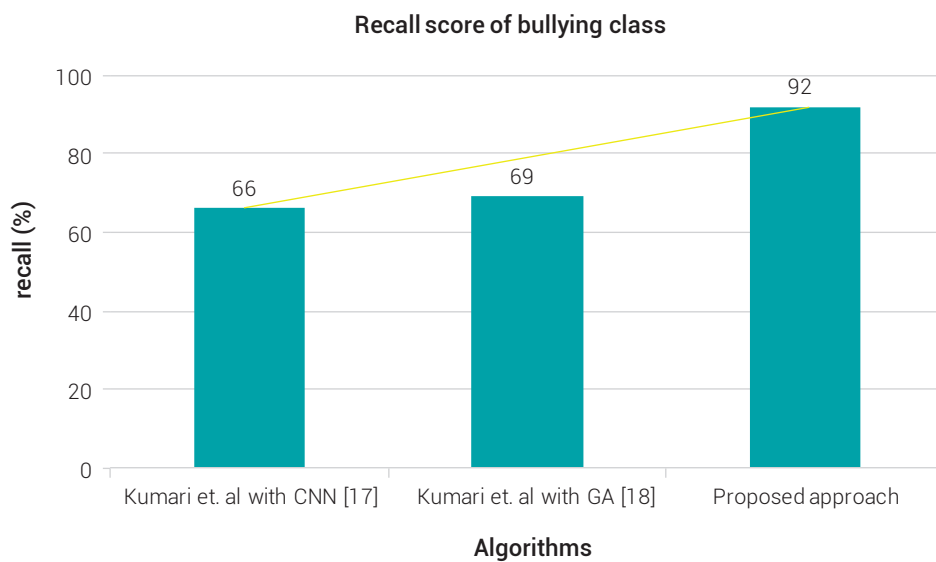
Image features size	Text features size	Class	Results		
			Precision	Recall	F1-score
128	128	Non-bullying	0.66	0.84	0.74
		Bullying	0.80	0.58	0.67
		Weighted Average	0.73	0.71	0.71
128	256	Non-bullying	0.74	0.90	0.81
		Bullying	0.86	<b>0.69</b>	<b>0.76</b>
		Weighted Average	0.80	0.79	0.78
256	128	Non-bullying	0.73	0.81	0.76
		Bullying	0.76	0.67	0.71
		Weighted Average	0.74	0.74	0.74

(continúa)

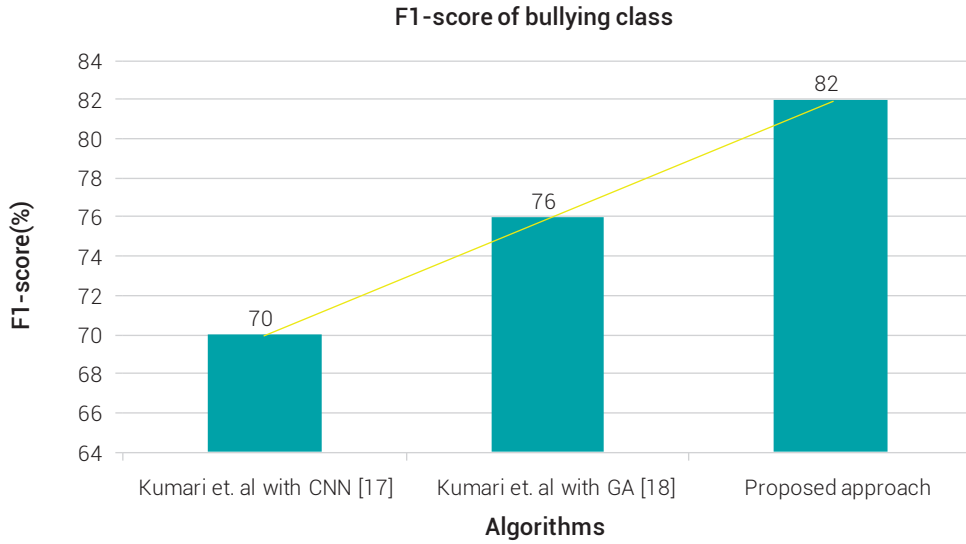
(viene)

Image features size	Text features size	Class	Results		
			Precision	Recall	F1-score
256	256	Non-bullying	0.69	0.83	0.75
		Bullying	0.80	0.64	0.71
		Weighted Average	0.75	0.73	0.73
256	512	Non-bullying	0.68	0.92	0.78
		Bullying	0.88	0.59	0.71
		Weighted Average	0.78	0.75	0.74
512	256	Non-bullying	0.70	0.84	0.76
		Bullying	0.75	0.59	0.66
		Weighted Average	0.73	0.72	0.72
512	512	Non-bullying	0.70	0.90	0.79
		Bullying	0.87	0.62	0.73
		Weighted Average	0.79	0.76	0.76

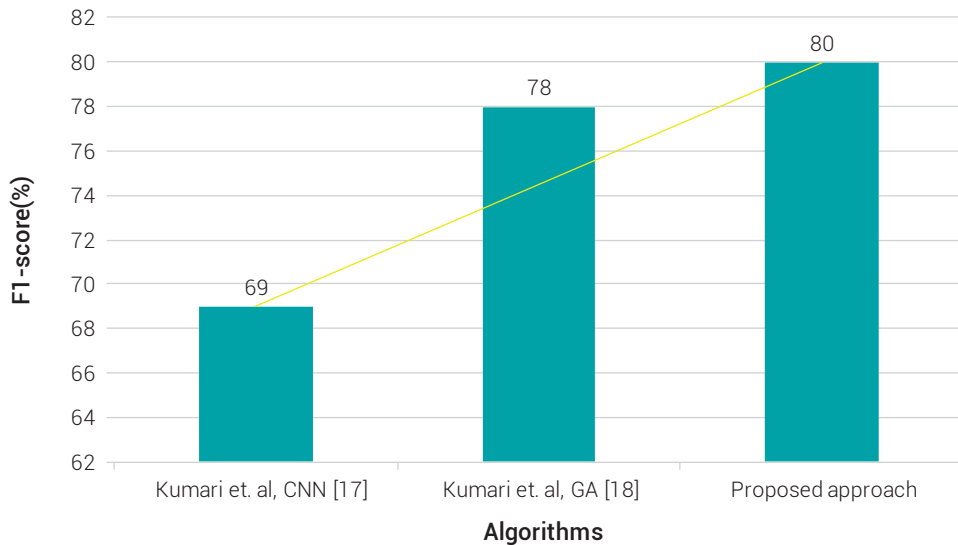
Source: own work

**Fig 3.** Recall score of Bullying class

Source: own work



**Fig 4. F1-score of bullying class**  
Source: own work



**Fig 5. Weighted F1-score**  
Source: own work

We apply Gradient-weighted Class Activation Mapping (Grad-CAM) [23] to highlighting important regions which causes bullying in images. Table 5 shows Grad-CAM images of bullying images where the Xception model was able to identify the highlighted features.



Table 5. Grad-CAM visualizations

Original image	Grad-CAM visualization
	
	
	

Source: own work

## 5. CONCLUSION AND FUTURE DIRECTIONS

The occurrence of cyberbullying increases in proportion with technological growth and high-speed digital devices in general and sophisticated online social media platforms in specific. The existing approaches to detect cyberbullying rely on data at the text level. In order to improve the performance of the cyberbullying detectors, a more

refined approach was presented in this work by considering combinatorial data that include text as well as image data. The proposed approach was effective in identifying cyberbullying using combinatorial data and achieved 92% recall, and an 82% F1-score for bullying class. The weighted F1-score of proposed model achieves 80%. Though the work proposed addresses the detection of cyberbullying, it considered the text-level features that are in the English language and requires additional mechanisms to apply the detection process in other languages. As part of the future work, the detection of cyberbullying is aimed to be attempted in Indian regional languages.

## REFERENCES

- [1] J.A. Pater, A.D. Miller, E.D. Mynatt, "This digital life: a neighborhood-based study of adolescents' lives online," *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 2305-2314, 2015, doi: 10.1145/2702123.2702534
- [2] D. Halpern, M. Pina and J. Vasquez, "Loneliness, personal and social well-being: towards a conceptualization of the effects of Cyberbullying," *Cult.y Educ.*, vol-29, no.4, pp.703- 727, oct-2017, doi: <https://doi.org/10.1080/11356405.2017.1370818>
- [3] K. Reynolds, A. Kontostathis A, L. Edwards, "Using machine learning to detect cyberbullying," *2011 10th International conference on machine learning and applications and workshops (ICMLA)*, vol. 2, pp 241-244. 2011, doi: 10.1109/ICMLA.2011.152
- [4] D. Yin, Z. Xue, L. Hong, B.D. Davison, A. Kontostathis, L. Edwards, "Detection of harassment on web 2.0.," *Proc Content Anal WEB 2*, 1-7, 2009.
- [5] H. Rosa, D. Matos, R. Ribeiro, L. Coheur and J. P. Carvalho, "A "Deeper" Look at Detecting Cyberbullying in Social Networks," *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, pp. 1-8, Oct-2018, doi: 10.1109/IJCNN.2018.8489211
- [6] S. Bu., S. Cho., "A Hybrid Deep Learning System of CNN and LRCN to Detect Cyberbullying from SNS Comments," *Hybrid Artificial Intelligent Systems*, 2018, doi: 10.1007/978-3-319-92639-1\_47
- [7] D. Van Bruwaene, Q. Huang, D. Inkpen, "A multi-platform dataset for detecting cyberbullying in social media," *Lang Resources & Evaluation*, 54, pp.851-874, 2020, doi : <https://doi.org/10.1007/s10579-020-09488-3>

- [8] Zx. Liu, Dg. Zhang, Gz. Luo, et al., "A new method of emotional analysis based on CNN- Bi-LSTM hybrid neural network," *Cluster Computing*, 23, pp.2901–2913, 2020. doi : <https://doi.org/10.1007/s10586-020-03055-9>
- [9] S.S. Shylaja, Abhishek Narayanan, Abhijith Venugopal, Abhishek Prasad, "Recurrent Neural Network Architectures with Trained Document Embeddings for Flagging Cyber-Aggressive Comments on Social media," *International Conference on Advanced Computing and Communications (ADCOM)*, 2018.
- [10] F. Frederick, Patacsil, "Analysis of Cyberbullying Incidence among Filipina Victims: A Pattern Recognition using Association Rule Extraction," *International Journal of Intelligent Systems and Applications (IJISA)*, vol.11, no.11, pp.48-57, 2019, doi: 10.5815/ijisa.2019.11.05
- [11] J. de-la-Peña-Sordo, I. Pastor-López, X. Ugarte-Pedrero, I. Santos and P. G. Bringas, "Anomaly-based user comments detection in social news websites using troll user comments as normality representation," *Logic Journal of the IGPL*, vol. 24, no. 6, pp. 883-898, 2016, doi: 10.1093/jigpal/jzw043.
- [12] G.O. Gutiérrez-Esparza, M. Vallejo-Allende, J. Hernández-Torruco, "Classification of Cyber-Aggression Cases Applying Machine Learning," *Appl. Sci.* 9, no. 9, 1828., 2019, dpoi: <https://doi.org/10.3390/app9091828>
- [13] Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea, "Content-driven detection of cyberbullying on the Instagram social network," *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, (IJCAI'16), AAAI Press, pp,3952–3958, 2016.
- [14] R. Zhao and K. Mao, "Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 328-339, 2017, doi : 10.1109/TAFFC.2016.2531682.
- [15] A. Kumar, Nitin Sachdeva. "Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis," *Multimedia Tools and Applications*, pp. 1-38, 2019, doi: <https://doi.org/10.1007/s11042-019-7234-z>
- [16] X. Zhang, X. et al., "Cyberbullying Detection with a Pronunciation Based Convolutional Neural Network," *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 740-745, 2016, doi : 10.1109/ICMLA.2016.0132.

- [17] J.K. Tripathy, S.S. Chakkaravarthy, S.C. Satapathy, et al., "ALBERT-based fine-tuning model for cyberbullying analysis," *Multimedia Systems*, 2020, doi: <https://doi.org/10.1007/s00530-020-00690-5>
- [18] S. Paul, S. Saha, "CyberBERT: BERT for cyberbullying identification," *Multimedia Systems*, 2020, doi : <https://doi.org/10.1007/s00530-020-00710-4>
- [19] K. Kumari, J.P. Singh, Y.K. Dwivedi, et al., "Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach," *Soft Computing*, pp.11059-11070, 2019, doi: <https://doi.org/10.1007/s00500-019-04550-x>
- [20] K. Kumari, J.P. Singh, "Identification of cyberbullying on multi-modal social media posts using genetic algorithm," *Wiley*, 2020, doi : <https://doi.org/10.1002/ett.3907>
- [21] F. Chollet, "Xception: Deep learning with depth wise separable convolutions," *Conference Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1800-1807, 2017, doi: 10.1109/CVPR.2017.195.
- [22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov, "RoBERTa: A robustly optimized BERT pre-training approach," 2019.
- [23] J. Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," pp. 4171-4186, 2018, doi:10.18653/v1/N19-1423
- [24] Ramprasaath R selvaraju, Michael Cogswell, Abhisek Das, Ramakrishna Vedantam, Devi Parikh & Dhruv Batra, " Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of computer vision*, pp. 336–359, 2020, doi: <https://doi.org/10.1007/s11263-019-01228-7>
- [25] Satya Prakash Yadhav and Sachin Yadav, "Image fusion using hybrid methods in multimodality medical images," *Medical & Biological Engineering & Computing, Springer*, 2020, doi : <https://doi.org/10.1007/s11517-020-02136-6>