Ensemble Deep Learning (EDL) for Cyber-bullying on Social Media

Zarapala Sunitha Bai[1]*, Sreelatha Malempati[2]
Department of Computer Science and Engineering-Y.S.R University College of Engineering & Technology,
Acharya Nagarjuna University, Guntur 522510, Andhra Pradesh, India[1]
Department of Computer Science and Engineering, R.V.R & J.C College of Engineering,
Chowdavaram, Guntur-522019, India[2]

***Abstract*—Cyber-bullying is a growing problem in the digital age, affecting millions of people worldwide. Deep learning algorithms have the potential to assist in identifying and combating Cyber-bullying by detecting and classifying harmful messages. This paper uses two Ensemble deep learning (EDL) models to detect Cyber-bullying on text data, images and videos—and an overview of Cyber-bullying and its harmful effects on individuals and society. The advantages of using deep learning algorithms in the fight against Cyber-bullying include their ability to process large amounts of data and learn and adapt to new patterns of Cyber-bullying behaviour. For text data, firstly, a pre-trained model BERT (Bidirectional Encoder Representations from Transformers) is used to train on cyber-bullying text data. The next step describes the data pre-processing and feature extraction techniques required to prepare data for deep learning algorithms. We also discuss the different types of deep learning algorithms that can be used for Cyber-bullying detection, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and deep belief networks (DBNs). This paper combines the sentiment analysis model, such as Aspect-based Sentiment Analysis (ABSA), for classifying bullying messages. Deep Neural network (DNN) used the classification of Cyber-bullying images and videos. Experiments were conducted on three datasets such as Twitter (Kaggle), Images (Online), and Videos (Online). Datasets are collected from various online sources. The results demonstrate the effectiveness of EDL and DNN in detecting Cyber-bullying in terms of detecting bullying data from relevant datasets. The EDL and DNN obtained an accuracy of 0.987, precision of 0.976, F1-score of 0.975, and recall of 0.971 for the Twitter dataset. The performance of Ensemble CNN brought an accuracy of 0.887, precision of 0.88, F1-score of 0.88, and recall of 0.887 for the Image dataset. For the video dataset, the performance of Ensemble CNN is an accuracy of 0.807, precision of 0.81, F1-score of 0.82, and recall of 0.81. Future research should focus on developing more accurate and efficient deep learning algorithms for Cyber-bullying detection and investigating the ethical implications of using such algorithms in practice.**

Keywords—Cyber bullying; ensemble deep learning (EDL); convolutional neural networks (CNNs); recurrent neural networks (RNNs); deep belief networks (DBNs)

I. Introduction

Sentiment analysis can be used to detect instances of cyber-bullying by analyzing the language and tone used in online messages, comments, or posts [1]. The process involves using natural language processing (NLP) techniques to identify the sentiment expressed in a text, whether it is positive, negative, or neutral [2]. One approach to using sentiment analysis for cyber-bullying detection is to look for negative sentiments expressed towards an individual or group, such as derogatory or offensive language, insults, or threats [3]. These can be identified using NLP techniques, such as part-of-speech tagging, named entity recognition, and sentiment analysis algorithms. Another approach is to analyze the context in which the message is being conveyed, such as the topic being discussed and the online community it is being shared in [4] [5]. For example, if a message contains negative sentiments towards a specific group of people or individual, and is being shared in an online community known for hostile or aggressive behaviour, it may be a sign of cyber-bullying. It is important to note that sentiment analysis is not a foolproof method for detecting cyber-bullying and should be used in combination with other techniques, such as human moderation and reporting mechanisms. Additionally, it is important to ensure that the use of sentiment analysis does not infringe on individuals' privacy rights or result in false accusations.

Cyber-bullying (CB) is a default model for publicly abusing a person. Many online social media networking (OSMN) like Facebook, Twitter, and Instagram act as a medium for people based on cyber-bullying attacks [6]. Several automated models aim to develop to classify cyber-bullying in terms of text messages, audio, and videos [7]. Sometimes based on the topic modeling, cyber-bullying attacks occur in several datasets belonging to topic modeling. Twitter has become more popular for cyber-bullying by using various types of attacks. Exemplary-grained automated models were developed to detect cyber-bullying regarding topic modeling [8]. It is essential to complain the cyber-bullying attacks in OSMS if the user violates the ITE Law, which is considered a crime in OSMS like Twitter [9]. The victim should act if any abusive language is used on Twitter. The aim of cyber-bullying mainly focuses on classifying the tweets present on Twitter. This paper describes the automated approach for detecting cyber-bullying attacks by using the sentiment analysis on text messages, videos, and audio. Sentiment analysis helps the proposed automated approach to find the cyber-bullying attacks in multi-media. An aspect based sentiment analysis model combined with automated classification approach used to classify the cyber-bullying words from given input data. It shows the improved

*Corresponding Author.

performance in terms of accuracy, sensitivity, specificity, F1-score and precision.

The organization of the paper is as follows. Section II explains the literature survey about various methods of cyberbullying. Section III and Section IV give the significance of the work and training and feature extraction techniques. Section V describes the experimental results with comparative performance. Section VI describes the conclusion.

A. *Significant Points of Proposed System*

Cyber-bullying has become a common problem in today's digital age, with severe psychological and emotional consequences for victims. Detecting and preventing cyber-bullying is critical for ensuring individuals' well-being in online communities. Manual monitoring of online content, on the other hand, is a time-consuming and inefficient process. As a result, there is a need to create an automated system that uses deep learning models to detect cyber-bullying. This project aims to develop and test a deep-learning model to detect cyber-bullying in online text, images, and videos. The model should be able to classify messages, comments, or posts as cyber-bullying or non-cyber-bullying based on their content. The detection system's accuracy, precision, and recall should be high, with low error rates.

II. Literature Survey

Semantic-enhanced marginalized denoising auto-encoder (SEDMA) is a type of neural network that is trained to reconstruct clean data from noisy data by learning the underlying distribution of the input data. It has been enhanced with semantic information to improve its ability to capture the context and meaning of the text [10]. To use SEDMA for cyber-bullying detection, the first step is to train the model on a dataset of labelled examples of cyber-bullying and non-cyber-bullying text. During training, the SEDMA learns to identify patterns and features that distinguish between cyber-bullying and non-cyber-bullying text. Once the SEDMA is trained, it can be used to detect cyber-bullying in new text. The input text is first pre-processed to remove noise and convert it into a numerical representation that can be input to the SEDMA. The SEDMA then reconstructs the clean version of the input text, and the reconstruction error is used to determine whether the input text is cyber-bullying or not. The advantage of using SEDMA for cyber-bullying detection is that it can capture the semantic meaning of the text, which is critical for detecting subtle forms of cyber-bullying. Additionally, it is more robust to noise and can handle variations in the input text. In conclusion, the use of semantic-enhanced marginalized denoising auto-encoder is a promising approach for cyber-bullying detection, and it has the potential to improve the accuracy of current cyber-bullying detection systems. Zhang et al. [11] proposed the novel pronunciation-based CNN to solve issues in detecting cyber-bullying based on the pronunciation of misspelled words. The proposed approach corrects the errors that occur by spellings that didn't change in accent because of its noise and data sparsity imbalance present in the dataset. To solve these issues, the proposed model applied to two datasets, Twitter and Form spring. The proposed approach's comparative performance shows more effective results than existing models. Zhang et al. [12] developed a fine-grained model to detect cyber-bullying messages based on linguistic analysis. The proposed model focused on finding the patterns based on Linguistic Inquiry Word Count (LIWC) to detect fine-tuned cyber-bullying detection from different social media datasets. Dalvi et al. [13] proposed an ML model to see cyber-bullying from social media posts like Twitter. The proposed ML model is used to prevent bullying on Twitter. Using the Twitter API, the tweets are extracted and classified whether the tweets are bullied or not. Zhao et al. [14] proposed a new learning model to solve the issues in cyber-bullying detection—the proposed approach combined with a deep learning model to denoising auto-encoder (SDA). The SDA model contains semantic dropout noise and sparsity. These features focused on knowledge and the word embedding method. The performance of the proposed model improved by combining it with semantic-enhanced marginalized (SEM) to find the hidden features of the bullying content. The version of the proposed model was analyzed using two datasets such as Twitter and MySpace, and achieved better performance in terms of classification. Luo et al. [15] proposed the BiGRU-CNN for classifying cyber-bullying messages. BiGRU mainly focused on extracting the global features that significantly impact organizing bullying messages. The CNN consists of a convolution method with 128 kernels of length 5; this is used to extract the features that improve the learning rate of the model better. Adav et al. [16] introduced the BERT model, which is suitable for creating contextual embeddings that produce the particular embeddings for classifying cyber-bullying detection in social media. Ahmed et al. [17] introduced the cyber-bullying model that classifies the bullying words belonging to Bangla and Romanized Bangla texts utilizing ML and DL models. Iwendi et al. [18] performed various DL models that detect cyber-bullying in social media. The comparison between different DL algorithms shows high performance. Aind et al. [19] introduced the novel Q-Bully model that sees cyber-bullying automatically from social media platforms. The proposed performance is improved by combining with Reinforcement Learning gives better accuracy. Ketsbaia et al. [20] introduced the DL models that detect cyber-bullying automatically. Pradhan et al. [21] proposed the new DL model that sees the cyber-bullying from Wikipedia, Formspring, and Twitter cyber-bullying datasets.

III. How Cyber-Bullying Affects the Social Media

Cyber-bullying can have a significant impact on social media use, both for individuals and as a whole. Here are some of the ways it can affect social media introduction:

1) Fear of harassment: Cyber-bullying can create a climate of fear on social media platforms. Individuals who have been bullied in the past or who fear being bullied may be hesitant to join social media or may limit their use of these platforms.

2) Damage to reputation: Cyber-bullying can damage an individual's reputation, making them less likely to want to be active on social media. This can also affect the reputation of the platform itself if it becomes known as a place where bullying is rampant.

3) Decreased engagement: Cyber-bullying can lead to decreased engagement on social media, as individuals may avoid posting or interacting with others out of fear of being targeted. This can have a negative impact on the platform as a whole, as it relies on user engagement to generate revenue.

4) Reduced trust: If social media platforms are seen as a place where cyber-bullying is common, users may lose trust in these platforms and be less likely to use them. This can affect the growth and sustainability of social media as a whole.
Overall, cyber-bullying can have a significant impact on social media use and adoption, and it is important for individuals and social media companies to take steps to prevent and address this issue, Fig. 1.

Fig. 1. Types of cyber-bullying.

IV. Cyber-bullying Detection model for Text Messages

A. BERT (Bidirectional Encoder Representations from Transformers) for Training on Cyber-Bullying Data

Cyber-bullying can have a significant impact on social media use, both for individuals and as a whole. Here are some of the ways it can affect social media introduction:

1) Fear of harassment: Cyber-bullying can create a climate of fear on social media platforms. Individuals who have been bullied in the past or who fear being bullied may be hesitant to join social media or may limit their use of these platforms.

2) Damage to reputation: Cyber-bullying can damage an individual's reputation, making them less likely to want to be active on social media. This can also affect the reputation of the platform itself if it becomes known as a place where bullying is rampant.

3) Decreased engagement: Cyber-bullying can lead to decreased engagement on social media, as individuals may avoid posting or interacting with others out of fear of being targeted. This can have a negative impact on the platform as a whole, as it relies on user engagement to generate revenue.

4) Reduced trust: If social media platforms are seen as a place where cyber-bullying is common, users may lose trust in these platforms and be less likely to use them. This can affect the growth and sustainability of social media as a whole.
Overall, cyber-bullying can have a significant impact on social media use and adoption, and it is important for individuals and social media companies to take steps to prevent and address this issue.

5) Here are the equations for the fine-tuning process: First, we add a classification layer on top of the pre-trained BERT model. The classification layer consists of a fully connected layer and a soft-max activation function.

$$\mathrm{h_{cls}} = \mathrm{W_{cls}} \times [\mathrm{CLS}] + \mathrm{b_{cls}} \quad (1)$$

$$\mathrm{y_{hat}} = \mathrm{softmax}(\mathrm{h_{cls}}) \quad (2)$$

Where $\mathrm{h_{cls}}$ is the hidden state of the [CLS] token, $\mathrm{W_{cls}}$ and $\mathrm{b_{cls}}$ are the weight and bias parameters of the fully connected layer, and $\mathrm{y_{hat}}$ is the predicted probability distribution over the two classes (cyber-bullying and non-cyber-bullying).

We then define the loss function as the cross-entropy loss between the predicted and true labels:

$$\mathrm{L} = -\sum \mathrm{y_i} \times \log\left(\mathrm{y_{hat_i}}\right) + (1 - \mathrm{y_i}) \times \log(1 - \mathrm{y_{hat_i}})) \quad (3)$$

Where $\mathrm{y_i}$ is the true label (1 for cyber-bullying, 0 for non-cyber-bullying), and $\mathrm{y_{hat_i}}$ is the predicted probability for the i^{th} message.

We optimize the parameters of the classification layer by minimizing the loss function using gradient descent:

$$\mathrm{W_{cls}}, \mathrm{b_{cls}} = \mathrm{argmin}(\mathrm{L}) \quad (4)$$

The above equations outline the fine-tuning process for BERT in cyber-bullying detection. Note that this process requires a labelled dataset of cyber-bullying and non-cyber-bullying messages, as well as appropriate pre-processing and tokenization of the input text.

B. Pre-processing Techniques for Cyber-Bullying

1) Tokenization: Tokenization divides the text into smaller units known as tokens, which can be words, phrases, or symbols. It is a common step in pre-processing natural language processing (NLP) tasks such as sentiment analysis and topic modelling. Consider the following example sentence to demonstrate tokenization for cyber-bullying data:

- "I hate you and wish you were never born. You're worthless and nobody likes you."
- To tokenize this sentence, we could use a straightforward method of separating the text by whitespace and punctuation marks.
- The tokenized sentence would look like this:
- ["I", "hate", "you", "and", "wish", "you", "were", "never", "born", ".", "You're", "worthless", "and", "nobody", "likes", "you", "."]

Each element in the resulting list is a token that can be processed and analyzed further using various NLP techniques, Fig. 2.

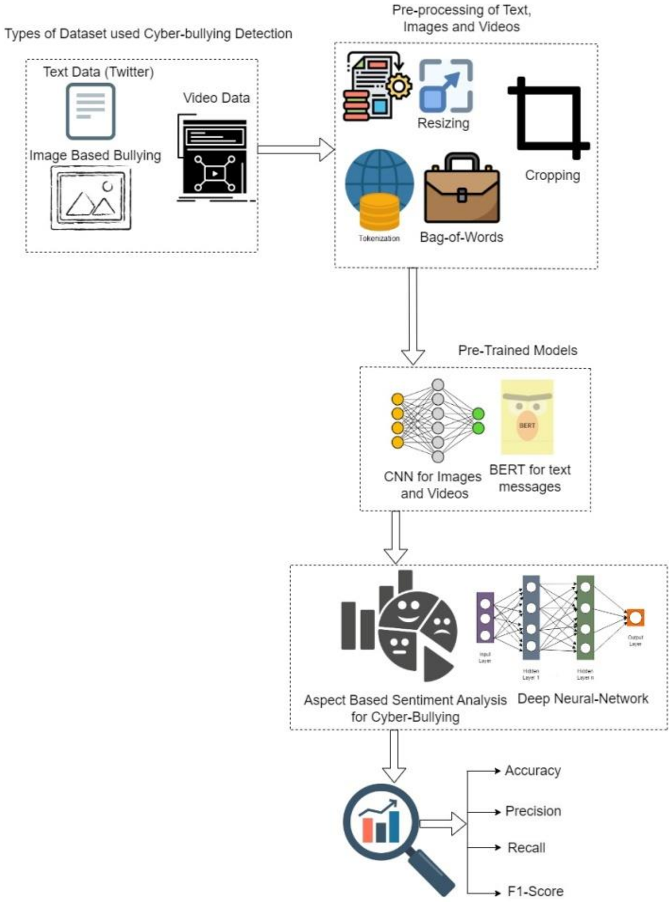


Fig. 2. Overall system architecture.

2) Stop-words removal: Stop words are commonly used in a language but have little meaning and can be removed from text without affecting the overall message. Stop word removal can be used in cyber-bullying to filter out irrelevant or offensive words from the text to identify and prevent bullying behavior.

- Here's an example of how to use stop-word removal in the context of cyber-bullying:
- Assume a social media platform wants to look for instances of cyber-bullying in user posts.
- The platform could include a stop word filter that removes common words and phrases that are unlikely to be used in cyber-bullying incidents, such as "the," "and," "is," "in," "a," "of," and "on."
- For example, if a user writes, "I hate you, and I hope you die," the stop word filter will remove the words "I," "you," "and," "hope," and "die." The filtered text would then be "hate," raising a red flag and prompting a review by the platform's moderators.
- Stop word removal is a valuable tool for detecting cyber-bullying and promoting a safer online environment.
- However, it is essential to note that stop-word removal alone may not be enough to identify instances of cyber-bullying accurately and that other techniques, such as sentiment analysis and machine learning algorithms, may be required.

3) Feature extraction techniques for cyber-bullying: Feature extraction is a crucial step in natural language processing (NLP) tasks such as cyber-bullying detection. Here are some techniques for feature extraction from cyber-bullying text data:

a) Bag of Words (BoW): It is a simple and effective method to extract features from text data. It involves counting the frequency of occurrence of each word in the document. For instance, consider the following sentence: "You're such a loser. Nobody likes you." The BoW representation of this sentence would be: {'you': 2, 're': 1, 'such': 1, 'a': 1, 'loser': 1, 'nobody': 1, 'likes': 1}.

b) TF-IDF (Term Frequency-Inverse Document Frequency): It is another technique that helps to extract features from text data. It assigns a weight to each word based on its frequency in the document and its frequency in the entire corpus. For example, in the sentence "You're such a loser. Nobody likes you," the word "you" appears twice in the document but is likely to appear in many other documents too. So, the weight assigned to "you" will be relatively low.

c) N-grams: N-grams are a sequence of N words in a sentence. For example, a bigram of the sentence "You're such a loser. Nobody likes you" would be "you're such," "such a," "a loser," "loser nobody," "nobody likes," and "likes you." N-grams help capture the context of words in a sentence.

d) Word embeddings: Word embeddings are vector representations of words that capture semantic and syntactic relationships between them. They are learned using neural networks trained on large amounts of text data. Word2Vec and GloVe are some examples of popular word embedding techniques.

Example:

- Let's say you have a dataset containing cyber-bullying text data, and you want to use these techniques to extract features from it. Here is an example of how you can use these techniques to extract features:
- Suppose you have a sentence in your dataset like this: "You are ugly and nobody likes you."
- BoW representation: {'you': 2, 'are': 1, 'ugly': 1, 'and': 1, 'nobody': 1, 'likes': 1}.
- TF-IDF representation: {'you': 0.276, 'are': 0.276, 'ugly': 0.385, 'and': 0.385, 'nobody': 0.385, 'likes': 0.385}.

- N-gram representation: {('you', 'are'): 1, ('are', 'ugly'): 1, ('ugly', 'and'): 1, ('and', 'nobody'): 1, ('nobody', 'likes'): 1, ('likes', 'you'): 1}.
- Word embeddings: [-0.456, 0.678, -0.234, 0.987, 0.678, -0.567] (this is just an example of a vector representation of the sentence using word embeddings, and the values are random).

V. Aspect-based Sentiment Analysis (ABSA) for Cyber-bullying

ABSA is a natural language processing technique used to identify and extract aspects or features of a given text and determine their sentiment polarity (positive, negative, or neutral) [22]. In the context of cyber-bullying, ABSA can be used to identify the specific aspects or topics that are associated with negative or abusive comments, messages, or posts.

A mathematical model for ABSA in cyber-bullying detection could be formulated as follows:

Let D be a set of documents containing potentially abusive or negative content, and A be a set of aspects or topics that may be associated with cyber-bullying. Each document d ∈ D can be represented as a set of sentences $\{s_1, s_2, \ldots, s_n\}$ and each sentence si can be further represented as a set of words $\{w_1, w_2, \ldots, w_n\}$. Let $P(w_i|s_i)$ be the probability of word w_i occurring in sentence s_i, and let $P(si|d)$ be the probability of sentence s_i occurring in document d.

The sentiment polarity of each aspect a ∈ A can be determined based on the sentiment scores of the words that are associated with that aspect. Let S(a) be the sentiment score of aspect a, which can be calculated as follows:

$$S(a) = \sum w_i \in a\, P(w_i|a) * \text{ Polarity}(w_i) \qquad (5)$$

Where $\text{Polarity}(w_i)$the polarity scores of is word w_i (e.g., +1 for positive, -1 for negative, 0 for neutral).

To detect cyber-bullying, we can use a threshold value T to determine whether a document d contains abusive or negative content. Let B(d) be a binary variable that indicates whether document d is abusive or not, where B(d) = 1 if d is abusive and B(d) = 0 otherwise. We can define B(d) as follows:

$$B(d) = \{1 \text{ if} \max a \in A\, S(a) \geq \ T;\ 0 \text{ otherwise}\} \qquad (6)$$

Where $\max a \in A\, S(a)$ is the maximum sentiment score of all aspects in document d. The threshold value T can be determined empirically based on the distribution of sentiment scores in a training dataset of labeled cyber-bullying and non-cyber-bullying documents.

Overall, the mathematical model for ABSA in cyber-bullying detection involves identifying the aspects or topics associated with cyber-bullying, calculating the sentiment scores of those aspects based on the sentiment polarity of the words associated with them, and using a threshold value to determine whether a document is abusive or not.

A. *Convolutional Neural Networks (CNN) for Training on Images and Videos*

CNNs are a type of deep learning model that are particularly effective at processing visual data, making them a popular choice for image and video classification tasks, including cyber bullying detection. The basic architecture of a CNN consists of several layers, including convolutional layers, pooling layers, and fully connected layers. Each layer performs a specific function, and the output of one layer is fed as input to the next layer.

The equations used to train a CNN for cyber bullying image and video classification involve the use of back-propagation and gradient descent to update the weights and biases of the network. The overall goal is to minimize the error between the predicted output and the actual output.

The general equation for computing the output of a convolutional layer can be expressed as follows:

$$y_i = f(\sum j = 1\hat{n}\; w_j \times\ x_{ij} + b_i) \qquad (7)$$

Where:

- 'y_i' is the output of the ith neuron in the layer.
- 'f()' is the activation function.
- 'n' is the number of input neurons.
- 'w_j' is the weight connecting the jth input neuron to the ith output neuron.
- 'x_{ij}' is the activation of the jth input neuron at the ith location of the receptive field.
- 'b_i' is the bias term for the ith output neuron.

The pooling layer reduces the dimensionality of the input by aggregating nearby activations. The most common pooling operation is max pooling, which selects the maximum value from each local neighborhood of activations.

The fully connected layer takes the flattened output from the previous layer and applies a matrix multiplication operation to produce the final output. The equation for the fully connected layer can be expressed as:

$$y = f(Wx + b) \qquad (8)$$

Where:

y is the output vector.

f() is the activation function.

W is the weight matrix.

x is the input vector.

b is the bias vector.

During training, the weights and biases of the network are updated using the back-propagation algorithm. The gradient of the loss function with respect to each weight and bias is computed, and the weights and biases are updated in the opposite direction of the gradient to minimize the loss function.

The equations for back-propagation and gradient descent are as follows:

$$\frac{\mathrm{dL}}{\mathrm{dw}} = \frac{\mathrm{dL}}{\mathrm{dy}} \times \frac{\mathrm{dy}}{\mathrm{dw}} \tag{9}$$

$$\mathrm{w} = \mathrm{w} - \mathrm{Ir} \times \frac{\mathrm{dL}}{\mathrm{dw}} \tag{10}$$

$$\frac{\mathrm{dL}}{\mathrm{db}} = \frac{\mathrm{dL}}{\mathrm{dy}} \times \frac{\mathrm{dy}}{\mathrm{db}} \tag{11}$$

$$\mathrm{b} = \mathrm{b} - \mathrm{Ir} \times \frac{\mathrm{dL}}{\mathrm{db}} \tag{12}$$

Where:

‘L’ is the loss function.

‘w’ and ‘b’ are the weights and biases of the network.

‘y’ is the output of the network.

‘lr’ is the learning rate.

Overall, the use of CNNs with back-propagation and gradient descent provides an effective way to train models for cyber bullying image and video classification tasks.

B. Cropping of Images and Videos

Cropping is the removal of unwanted parts of an image or video to focus on a specific area or subject.

Here are some typical image and video cropping techniques:

1) The rule of thirds: This technique entails dividing the image or video into thirds horizontally and vertically and then positioning the subject along the intersections or lines.
As a result, the composition is more balanced and visually appealing.

2) Center crop: This technique involves cropping an image or video to center the subject. It works well when the issue is the main focus and there is no distracting background. Cropping an image or video to a specific aspect ratio, such as 4:3 or 16:9, is an example of this technique. It comes in handy when creating content for specific platforms or devices.

3) Pan and zoom: This technique involves cropping an image or video and animating it to simulate a camera pan or zoom. It can add motion to the image or video or emphasize specific parts.

4) Content-aware crop: This technique uses software tools to determine the best cropping based on the image or video content. It can be helpful when the subject is not in a fixed position or when complex background content needs to be removed.

C. Deep Neural Network (DNN)

DNN is specifically CNNs and recurrent neural networks (RNNs), can be used to solve the problem of classifying cyber-bullying in images and videos (RNNs).

D. Data Gathering and Pre-processing

Preprocess a large dataset of images and videos containing Cyber-bullying content to extract features like color, texture, shape, and motion.

E. Image Classification using CNN

Train a CNN on the image data to determine whether or not each image contains cyber-bullying content.

Multiple convolutional and pooling layers are followed by fully connected layers and a softmax output layer in a CNN. To improve the model's performance, employ data augmentation, dropout, and early stopping techniques.

F. Dataset Description

The experiments use a Python programming language with three datasets: Twitter, Images, and Videos dataset. The dataset is aged between 15-40 years from schools to job holders. Among these, 88% of data is analyzed as cyber-bullying. These tweets contain more than 47656 with two attributes such as tweet_text and tweet_type. Python libraries such as Keras, Pandas, and TensorFlow were used to analyze the performance of the proposed model. Table I shows the various types of bullying text messages for training and testing is given.

Table II shows the types of bullying images that affects the human personally and mentally. These images are JPAG images with standard size. These images are classified based on comments, captions, and topics.

Table III shows the various types of videos belong to different categories. Three types of bullying videos are present for experimental analysis. These videos such as hate speech, personal abuse and normal are shown in Table IV.

TABLE I. TYPES OF CYBER-BULLYING TEXT MESSAGES

Types of Cyber-bullying	**Tweets**
Age	The girl who bullied you in high school but now wants to sell you Arbonne
Ethnicity	I said dont put north west in coffee fuck the diddy call fifty i said there no to assassinate out the door of the air port dumb niggers
Gender	Don't call bitches females. That's mad disrespectful. Bitches hate when you call them females.
Religion	@UmarMal And I'm not sure how you can yammer about homelessness when Muslims are still murdering people for apostacy and blasphemy.
Other type of Cyber-bullying	@Eleoryth I sometimes envy those who don't have retarded parents
Not Cyber-bullying	Rebecca Black Drops Out of School Due to Bullying

TABLE II. TYPES OF TWITTER TEXT DATASET

Message Type	**Training**	**Testing**
Religion based bullying	3000	4997
Age based bullying	3000	4992
Ethnicity	3000	4959
Gender	3000	4948
Other cyber-bullying	3000	4823
Not cyber-bullying	3000	4937
Total	18000	29656

TABLE III. TYPES OF IMAGE DATASET

Image Type	Training	Testing
Morphing Images	1500	1500
Personal Abuse	500	500
Adult	1k	1k
Total Messages	3k	3k

TABLE IV. VIDEOS DATASET

Video Type	Training	Testing
Hate speech	25	25
Personal Abuse	25	25
Normal Videos	10	10
Total Videos	60	60

G. Performance Metrics

A confusion matrix is an approach that analyzes the performance of the proposed model. The classification of images will specify the model performance on test data. The confusion matrix mainly focused on two attributes such as predicted and original values, see Fig. 3.

True Negative (TN): The predicted input is bullied and actual input is also bullied.

True Positive (TP): The predicted input is not bullied and actual value is not bullied.

False Positive (FP): The predicted input is bullied and actual input is not bullied.

False Negative (FN): The predicted input is not-bullied and actual input is bullied.

Precision: This parameter gives the overall correct outputs given by the proposed model.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{13}$$

F1 Measure: It is the parameter that combines the recall and precision.

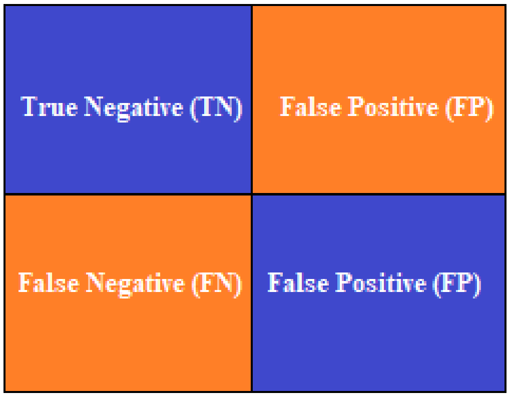
$$\text{F1 Measure} = 2 \times \frac{\text{precision}*\text{recall}}{\text{precision+recall}} \tag{14}$$

Accuracy: The overall accuracy of proposed model is measured as

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{15}$$

Recall: This metric is mainly focused on reducing the false negatives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{16}$$



True Negative (TN)	False Positive (FP)
False Negative (FN)	False Positive (FP)

Fig. 3. Confusion matrix.

TABLE V. COMPARATIVE PERFORMANCES OF EXISTING AND PROPOSED APPROACHES FOR ANALYSIS OF TWITTER DATA

		Precision	F1 Measure	Accuracy	Recall
2DCNN [23]	Religion based bullying	0.68	0.68	0.68	0.68
	Age based bullying	0.69	0.698	0.68	0.69
	Ethnicity	0.67	0.68	0.68	0.67
	Gender	0.678	0.687	0.698	0.685
	Other cyber-bullying	0.68	0.68	0.687	0.685
	Not cyber-bullying	0.68	0.675	0.675	0.68
InceptionV3 [24]	Religion based bullying	0.87	0.87	0.87	0.87
	Age based bullying	0.886	0.885	0.876	0.868
	Ethnicity	0.88	0.87	0.873	0.871
	Gender	0.878	0.876	0.879	0.873
	Other cyber-bullying	0.875	0.878	0.897	0.867
	Not cyber-bullying	0.884	0.878	0.876	0.875
Proposed Approach	Religion based bullying	0.965	0.956	0.966	0.964
	Age based bullying	0.961	0.966	0.968	0.967
	Ethnicity	0.976	0.975	0.978	0.967
	Gender	0.956	0.967	0.976	0.956
	Other cyber-bullying	0.976	0.976	0.968	0.976
	Not cyber-bullying	0.978	0.972	0.976	0.967

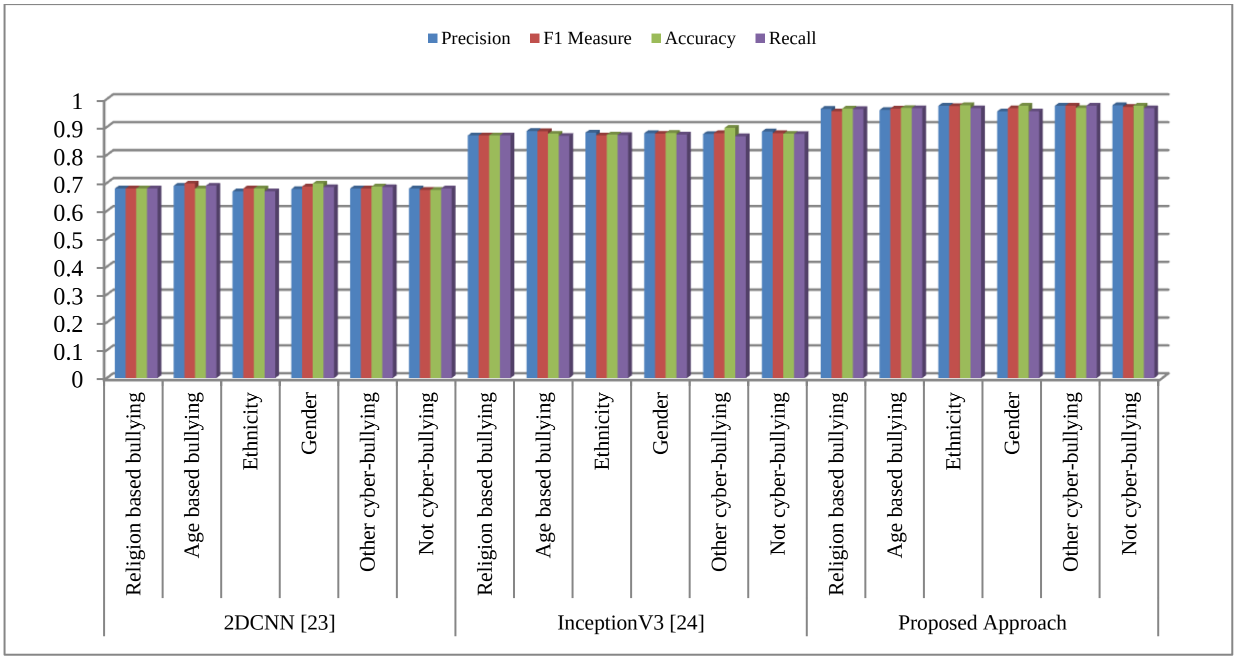


Fig. 4. Comparative performances of existing and proposed approaches for analysis of Twitter data.

Table V shows the comparative results among the existing 2DCNN [23], InceptionV3 [24] and proposed approach. Fig. 4 also compares text based cyber-bullying models based on the given data in twitter dataset.

Table VI, Table VII and Fig. 5, Fig. 6 shows a comparison of image-based cyberbullying. MoSI is the existing model, and Ensemble CNN is the proposed model. Ensemble CNN classifies images and videos of cyber-bullying. They are creating a diverse and representative dataset of cyberbullying incidents, including various types such as harassment, hate speech, or threats, and creating deep learning model architecture suitable for detecting cyberbullying. The model should be trained and optimized for high performance using the collected dataset. They are addressing the issue of imbalanced data, as cyber-bullying incidents are relatively rare compared to non-cyber-bullying incidents. These techniques need methods like oversampling, under sampling, and generating synthetic data.

TABLE VI. COMPARATIVE PERFORMANCES OF EXISTING AND PROPOSED APPROACHES FOR ANALYSIS OF IMAGES

	Types	**Precision**	**F1 Measure**	**Accuracy**	**Recall**
Motion from Static Images (MoSI)	Morphing Images	0.78	0.78	0.78	0.78
	Personal Abuse	0.79	0.698	0.78	0.79
	Adult	0.77	0.78	0.78	0.77
	Normal	0.77	0.78	0.78	0.78
Ensemble CNN	Morphing Images	0.88	0.88	0.88	0.88
	Personal Abuse	0.89	0.898	0.88	0.89
	Adult	0.87	0.88	0.88	0.87
	Normal	0.8856	0.879	0.8876	0.876

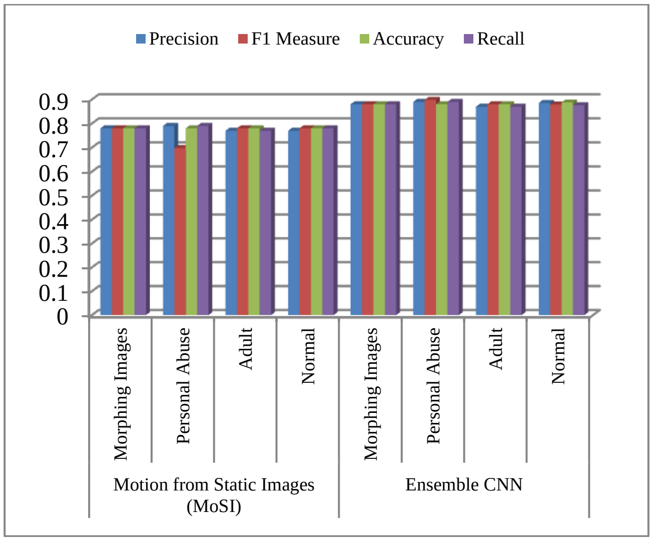


Fig. 5. Comparative performances of existing and proposed approaches for analysis of image data.

TABLE VII. COMPARATIVE PERFORMANCES OF EXISTING AND PROPOSED APPROACHES FOR ANALYSIS OF VIDEOS

		Precision	**F1 Measure**	**Accuracy**	**Recall**
Motion from Static Images (MoSI)[25]	Hate speech	0.68	0.78	0.78	0.78
	Personal Abuse	0.69	0.698	0.78	0.79
	Normal	0.67	0.78	0.78	0.77
Ensemble CNN	Hate speech	0.80	0.81	0.81	0.81
	Personal Abuse	0.80	0.821	0.82	0.81
	Normal	0.81	0.81	0.81	0.82

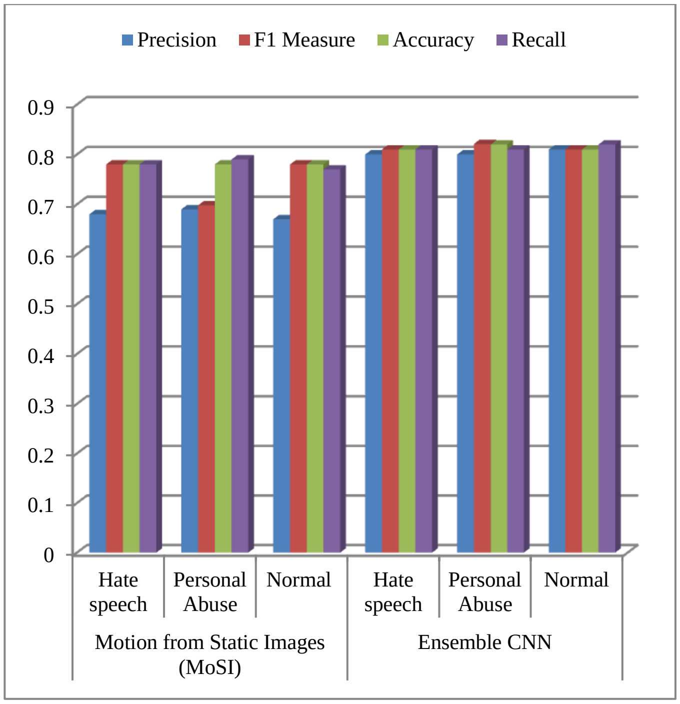


Fig. 6. Comparative performances of existing and proposed approaches for analysis of video data.

VI. Conclusion

Finally, an ensemble deep neural network can classify cyber-bullying on Twitter using text, images, and videos. The ensemble method combines multiple models' outputs to produce a more accurate and robust prediction. An Aspect-based Sentiment Analysis (ABSA) for Cyber-bullying is introduced for text classification. This model learned the patterns and features that distinguish cyber-bullying tweets from non-cyber-bullying tweets using large datasets of labeled text. The text model's output can then be combined with the results of the image and video models via a weighted voting scheme. A deep neural network, such as a recurrent neural network (RNN) or a convolutional neural network (CNN), can classify images and videos. These models trained on large datasets of labeled images and videos to learn the features and patterns that differentiate cyber-bullying from non-cyber-bullying content. A weighted voting scheme is used for the output of the image and video models that can be combined with the output of the text model. Because different models can capture various aspects of the data, using an ensemble deep neural network allows for a more accurate and robust classification of cyber-bullying on Twitter. The ensemble model is better equipped to handle the complexity and variability of cyber-bullying content on Twitter because their outputs are combined. Overall, using an ensemble deep neural network is a promising approach for addressing the problem of cyber-bullying on Twitter and can contribute to creating a safer and more positive online environment.

References

[1] Mahlangu, T., Tu, C.: Deep learning Cyber-bullying detection using stacked embbedings approach. IEEE: 2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI), pp. 45–49 (2019).

[2] Alam, K.S., Bhowmik, S., Prosun, P.R.K.: Cyber-bullying detection: an ensemble based machine learning approach. IEEE: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 710–715 (2021).

[3] S. Agrawal and A. Awekar, "Deep learning for detecting Cyber-bullying across multiple social media platforms," in Advances in Information Retrieval (Lecture Notes in Computer Science), vol. 10772, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds. Cham, Switzerland: Springer, 2018, pp. 141–153.

[4] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect Cyber-bullying," in Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops (ICMLA), vol. 2, Dec. 2011, pp. 241–244, doi:10.1109/ICMLA.2011.152.

[5] A. Muneer and S. M. Fati, "A comparative analysis of machine learning techniques for Cyber-bullying detection on Twitter," Futur. Internet, vol. 12, no. 11, pp. 1–21, 2020, doi: 10.3390/fi12110187.

[6] N. Yuvaraj, K. Srihari, G. Dhiman, K. Somasundaram, A. Sharma, S. Rajeskannan, M. Soni, G. S. Gaba, M. A. AlZain, and M. Masud, "Nature-inspired-based approach for automated Cyber-bullying classification on multimedia social networking," Math. Problems Eng., vol. 2021, pp. 1–12, Feb. 2021, doi: 10.1155/2021/6644652.

[7] N. Yuvaraj, V. Chang, B. Gobinathan, A. Pinagapani, S. Kannan, G. Dhiman, and A. R. Rajan, "Automatic detection of Cyber-bullying using multi-feature based artificial intelligence with deep decision tree classification," Comput. Electr. Eng., vol. 92, Jun. 2021, Art. no. 107186, doi: 10.1016/j.compeleceng.2021.107186.

[8] Y. Zhang and A. Ramesh, "Fine-grained analysis of Cyber-bullying using weakly-supervised topic models," in Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA), Oct. 2018, pp. 504–513, doi:10.1109/DSAA.2018.00065.

[9] N. M. G. D. Purnamasari, M. A. Fauzi, Indriati, and L. S. Dewi, "Cyber-bullying identification in Twitter using support vector machine and information gain based feature selection," Indones. J. Electr. Eng. Comput. Sci.,vol. 18, no. 3, pp. 1494–1500, 2020, doi: 10.11591/ijeecs.v18.i3.pp1494-1500.

[10] Zhao and K. Mao, "Cyber-bullying detection based on semanticenhanced marginalized denoising auto-encoder," IEEE Trans. Affect. Comput., vol. 8, no. 3, pp. 328–339, Jul. 2017, doi:10.1109/TAFFC.2016.2531682.

[11] X. Zhang, J. Tong, N. Vishwamitra, E. Whittaker, J. P. Mazer, R. Kowalski, H. Hu, F. Luo, J. Macbeth, and E. Dillon, "Cyber-bullying detection with a pronunciation based convolutional neural network," in Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2016, pp. 740–745, doi: 10.1109/ICMLA.2016.0132.

[12] Y. Zhang and A. Ramesh, "Fine-grained analysis of Cyber-bullying using weakly-supervised topic models," in Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA), Oct. 2018, pp. 504–513, doi: 10.1109/DSAA.2018.00065.

[13] R. R. Dalvi, S. B. Chavan, and A. Halbe, "Detecting a Twitter Cyber-bullying using machine learning," in Proc. 4th Int. Conf. Intell. Comput. Control Syst. (ICICCS), May 2020, pp. 297–301, doi: 10.1109/ICICCS48265.2020.9120893.

[14] R. Zhao and K. Mao, "Cyber-bullying detection based on semantic-enhanced marginalized denoising auto-encoder," IEEE Trans. Affect. Comput., vol. 8, no. 3, pp. 328–339, Jul. 2017, doi: 10.1109/TAFFC.2016.2531682.

[15] Luo, Y., Zhang, X., Hua, J., Shen, W.: Multi-featured Cyber-bullying detection based on deep learning. IEEE: 2021 16th International Conference on Computer Science & Education (ICCSE), pp. 746–751 (2021).

[16] Adav, J., Kumar, D., Chauhan, D.: Cyber-bullying detection using pre-trained bert model. IEEE: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1096–1100 (2020).

[17] Ahmed, M.T., Rahman, M., Nur, S., Islam, A., Das, D.: Deployment of machine learning and deep learning algorithms in detecting Cyber-bullying in bangla and romanized bangla text: A comparative study. IEEE: 2021 International Conference on Advances in Electrical,

Computing, Communication and Sustainable Technologies (ICAECT), pp. 1–10 (2021).

[18] Iwendi, C., Srivastava, G., Khan, S., Maddikunta, P.K.R.: Cyber-bullying detection solutions based on deep learning architectures. Multimedia Systems, 1–14 (2020).

[19] Aind, A.T., Ramnaney, A., Sethia, D.: Q-bully: a reinforcement learning based Cyber-bullying detection framework. IEEE: 2020 International Conference for Emerging Technology (INCET), pp. 1–6 (2020).

[20] Ketsbaia, L., Issac, B., Chen, X.: Detection of hate tweets using machine learning and deep learning. In: 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 751–758 (2020).

[21] Pradhan, A., Yatam, V.M., Bera, P.: Self-attention for Cyber-bullying detection. In: 2020 IEEE International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), pp. 1–6 (2020). IEEE

[22] Sahana, B., Sandhya, G., Tanuja, R., Ellur, S., Ajina, A.: Towards a safer conversation space: Detection of toxic content in social media (student consortium). In: 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), pp. 297–301 (2020).

[23] Kumari K, Singh JP, Dwivedi YK, Rana NP (2020) Towards Cyber-bullying-free social media in smart cities: a unified multimodal approach. Soft Comput 24(15):11059–11070

[24] Roy PK, Tripathy AK, Das TK, Gao X-Z. A framework for hate speech detection using deep convolutional neural network. IEEE Access. 2020;8:204951–204962. doi: 10.1109/ACCESS.2020.3037073.

[25] ZHuang, S. Zhang, J. Jiang, M. Tang, R. Jin, and M. H. Ang, "Self-supervised motion learning from static images," in Proceedings of the ieee/cvf conference on computer vision and pattern recognition, Nashville, TN, USA, June 2021.