Multi-modal Aggression Identification Using Convolutional Neural Network and Binary Particle Swarm Optimization

Kirti Kumari[a,*], Jyoti Prakash Singh[a], Yogesh K. Dwivedi[b], Nripendra P. Rana[c]

[a]*National Institute of Technology Patna, Patna, India*
[b]*School of Management, Swansea University, Bay Campus, Swansea, UK*
[c]*School of Management, University of Bradford, Bradford, UK*

Abstract

Aggressive posts containing symbolic and offensive images, inappropriate gestures along with provocative textual comments are growing exponentially in social media with the availability of inexpensive data services. These posts have numerous negative impacts on the reader and need an immediate technical solution to filter out aggressive comments. This paper presents a model based on a Convolutional Neural Network (CNN) and Binary Particle Swarm Optimization (BPSO) to classify the social media posts containing images with associated textual comments into *non-aggressive*, *medium-aggressive* and *high-aggressive* classes. A dataset containing symbolic images and the corresponding textual comments was created to validate the proposed model. The framework employs a pre-trained VGG-16 to extract the image features and a three-layered CNN to extract the textual features in parallel. The hybrid feature set obtained by concatenating the image and the text features were optimized using the BPSO algorithm to extract the more relevant features. The proposed model with optimized features and Random Forest classifier achieves a weighted F1-Score of 0.74, an improvement of around 3% over unoptimized features.

Keywords: Cyber-aggression; Cyberbullying; Multi-modal Data; Convolutional Neural network; Binary Particle Swarm Optimization

1. Introduction

Social media networks, such as Facebook[1], Instagram[2] and Vine[3], are platforms to share opinions, ideas and information. These platforms help businesses to grow by spreading information about their products and services in a relatively short time. Government agencies are using them as a feedback mechanism in making their policies and regulations. Social media is also helping the economy by providing a potential pathway for sustainable societies to assure its citizen's equality, freedom and a healthy standard of living [1, 2].

*Corresponding author
Email addresses: kirti.cse15@nitp.ac.in (Kirti Kumari), jps@nitp.ac.in (Jyoti Prakash Singh), ykdwivedi@gmail.com (Yogesh K. Dwivedi), nrananp@gmail.com (Nripendra P. Rana)

[1]www.facebook.com

[2]www.instagram.com

[3]https:/vine.co/

 Received: date / Accepted: date

Along with these positive usages to help economic growth and societal development, social media is also impacting our society negatively by spreading hate speech, fake news, negativity about government orders, anti-national activities, defamatory postings and so on. These activities have grown exponentially in the past few years [3]. The circulation of offensive and unacceptable comments on social media is a massive threat to our society. Cyber-aggression [4], hate speech, cyberstalking and cyberbullying [5] are among the most disturbing barriers to a sustainable society. The need of the hour to ensure the flourishing of an open society is to find out an appropriate way to respond to such materials without enforcing strict censorship.

Cyber-aggression is characterized as hostile or violent behaviour with the aim of harming others by using electronic media. It comprises sending, posting or sharing threatening, negative or nasty information about an individual or a group causing character assassination, humiliation, emotional stress, depression, anxiety and suicidal thoughts to the victim or victims. Such aggression occurs in many forms including textual aggression (instant messaging, e-mail, chatting), verbal aggression (verbal posts, phone calls) and visual aggression (sending, posting or sharing embarrassing videos or images). Although cyber-aggression can affect any age group of social networking users, adolescents and youths are the most affected groups. Recent studies have concluded that teens generally make frequent use of online sites for video and image sharing (e.g., Instagram and Vine) and are more vulnerable to these behaviours [6]. The visual contents (video and image), accounting for more than 70% of all Internet traffic[4], is making cyber-aggression more chaotic and damaging [7].

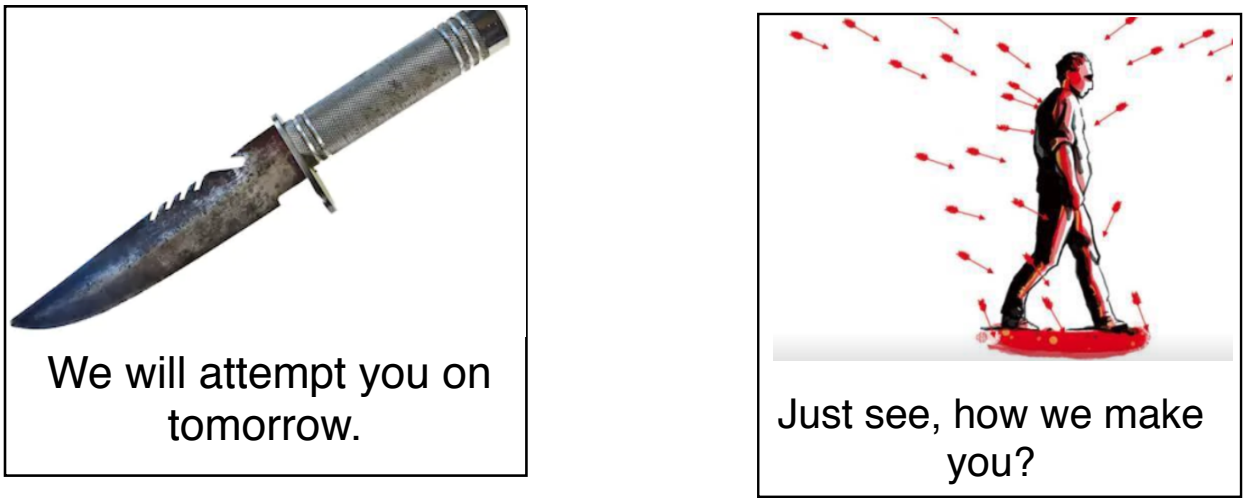
The severity of the problem requires immediate technical attention given that manual monitoring is not practically scalable and also very time-consuming. Hence, it is necessary to develop automated tools to detect these kinds of aggression in the very first instance to minimize mental and physical health problems of Internet users [8].

Most of the earlier works [4, 9, 10, 11, 12, 13] to distinguish between aggressive and non-aggressive posts were concentrated on the text content of posts whereas the posts also very often contain images along with text[17]. A few recent works [15, 16, 17] included images with text to identify the cases of cyberbullying. The images and text of the comments together with the user features like the number of followers and followees were used by Hosseinmardi et al. [15] to predict cyberbullying instances on the Instagram network. Another model to identify the cases of cyberbullying was proposed by Singh et al. [16] using visual and text characteristics for posts on the Instagram network. Kumari et al. [18] presented a model to detect cyber-aggressive posts using symbolic images. They considered images only and ignored the textual part of the post.

To the best of our knowledge symbolic images together with the text of the post were not considered in any of the earlier works for the detection of cyber-aggression. The current work concentrated on the identification of multi-modal aggressive posts containing symbolic images and associated comments. To achieve this task, we created a multi-modal dataset containing 3,600 images with associated comments.

We utilized the pre-trained VGG-16 [41] to extract features from images and three-layers CNN to extract features

[4]https://www.recode.net/2015/12/7/11621218/streaming-video-now-accounts-for-70-percent-of-broadband-usage



(a) (b)

Figure 1: Examples of cyber-aggressive posts containing symbolic images and text

from the text of a post. The BPSO algorithm was used to optimize the features space by eliminating redundant features. Then we applied the Random Forest classifier on optimized features to classify the multi-modal cyber-aggressive posts. Our major contributions can be summarized as:

- We created a multi-modal cyber-aggressive dataset containing symbolic images with composed text comments. From this dataset, we found a peculiar case in which the image and the text of a post separately appear non-aggressive but together they make the post highly aggressive as shown in Figure 1a.
- We proposed a hybrid model to extract features from text and symbolic images in parallel to get combined features for the multi-modal posts.
- We employed the BPSO algorithm to reduce the features space to improve the performance of classification. The proposed work improves performance by 3% compared to the unoptimized feature set.

The rest of the paper is organized as follows. In Section 2, the related works are discussed. The proposed framework for the detection of cyber-aggression is presented in Section 3. The findings of the current system are illustrated in Section 4. In Section 5, a discussion about the findings is provided. Finally, the article is concluded in Section 6 by outlining some future research directions.

2. Related Works

Automatic detection and prevention of cyber-aggressive posts have attracted a lot of attention in recent years [9, 18, 19, 20, 21, 24, 25]. Burnap and Williams [21] presented an ensemble learning-based solution to identify hate-related tweets. They used ensembles of three popular classifiers: (i) Logistic Regression, (ii) Support Vector Machine

(SVM) and (iii) Random Forest with *n-gram* text features to get an F1-Score of 0.77. Al-garadi et al. [22] also employed four classifiers: (i) Naive Bayes, (ii) SVM, (iii) Random Forest and (iv) K-Nearest Neighbours to detect cyberbullying instances for the posts collected from Twitter considering four sets of features: (a) user, (b) activity, (c) network and (d) content features. They reported their best result with the Random Forest classifier to have a recall value of 0.71. A model for online harassment detection for YouTube, Twitter, Formspring, MySpace, SlashDot and Kongregate was proposed by Chen et al. [23] using Naive Bayes and SVM classifiers. They reported their best result for the MySpace dataset having a recall value of 0.78. Waseem and Hovy [24] focused on identifying racist and sexist tweets and achieved an F1 score of 0.74 using character *n-gram* feature. The racist, sexist and homophobic tweets were also the main attention of Davidson et al. [25]. But they got very limited results in terms of recall and precision values of 0.61 and 0.44, respectively. A Semantic-Enhanced Marginalized Denoising Auto-Encoder based approach was used by Zhao and Mao [26] to detect cyberbullying comments. A verbal aggression detection system with sentiment analysis for tweets was proposed by Chen et al. [20] using a CNN model. An active learning method was proposed by Bhattacharjee et al. [27] to detect malicious forum content from web-based social media platforms. Gallo et al. [28] used a network knowledge-based model with a machine-learning classifier to identify the online users' reaction on Twitter. A cloud-based application to promote educational tools aiming at the acquisition of subject-related knowledge for tweets was proposed by Visvizi et al. [29]. The analysis of cyber-aggression conducted on tweets by Chatzakou et al. [4] revealed that combining the network and the user features with text features improves the performance of the model. They achieved the values 0.72 and 0.73 for overall accuracy and recall respectively for different classes. Sadiq et al. [30] developed and compared various models: Multi-Layer Perceptron (MLP) with TF-IDF features, MLP with word embedding, two deep neural network frameworks: CNN with Long Short Term Memory (CNN-LSTM) and CNN with Bidirectional LSTM (CNN-BiLSTM) with word embedding to identify cyber-trolling tweets. They found that MLP with the TF-IDF features-based model achieved 0.92 accuracy and outperformed other models.

Table 1: Summary of related works (wgt. F1 refers to weighted F1-Score)

Article	**Modality**	**Approach**	**Type of dataset**	**Validation method**	**Performance**
Burnap and Williams [21]	Text	Voted ensemble of Logistic Regression, Random Forest and SVM classifiers with *n-gram* features	English tweets	10-fold cross-validation	0.77 (F1-Score)
Al-garadi et al. [22]	Text	Content, Activity, User and Network features with Random Forest classifier	English tweets	10-fold cross validation	0.71 (Recall)
Chen et al. [23]	Text	Naive Bayes and SVM classifiers	YouTube, Twitter, Formspring, MySpace, SlashDot and Kongregate	10-fold cross validation	0.78 (Recall)
Waseem and Hovy [24]	Text	Linguistic and Character *n-gram* features with Logistic Regression classifier	English tweet	10-fold cross validation	0.74 (F1-Score)
Davidson et al. [25]	Text	Logistic Regression, Decision Tree, SVM, Naive Bayes and Random forest	English tweets	5-fold cross validation	0.61 (Recall)
Zhao and Mao [26]	Text	Semantic-Enhanced Marginalized Denoising Autoencoder	comments of Twitter and MySpace	Holdout validation	0.72 - 0.77 (wgt. F1)
Chen et al. [20]	Text	CNN	English tweets	Holdout validation	0.92 (Accuracy)
Chatzakou et al. [4]	Text	Bag of words, user and network-based features	English tweets	10-fold cross validation	0.73 (Recall)
Kumari and Singh [31]	Text	CNN model with One-hot, GloVe and FastText embeddings	Hindi-English code-mixed tweets	Holdout validation	0.52 - 0.78 (wgt. F1)

Raiyani et al. [9]	Text	Dense architecture and One-hot encoding	Hindi-English code-mixed comments of Facebook and Twitter	Holdout validation	0.48 - 0.60 (wgt. F1)
Julian and Krestel [13]	Text	Ensemble learning with word *n-gram*, character *n-gram*, Term Frequency-Inverse Document Frequency (TF-IDF) features	Hindi-English code-mixed comments of Facebook and Twitter	10-fold cross-validation	0.38 - 0.63 (wgt. F1)
Modha and Majumder [10]	Text	LSTM, CNN and FastText embedding	Hindi-English code-mixed comments of Facebook and Twitter	Holdout validation	0.50 - 0.62 (wgt. F1)
Samghabadi et al. [11]	Text	Logistic Regression and GloVe embedding, word *n-gram*, character *n-gram*, TF-IDF feature	Hindi-English code-mixed comments of Facebook and Twitter	Holdout validation	0.50 - 0.62 (wgt. F1)
Sadiq et al. [30]	Text	MLP with TF-IDF features	English comments of Twitter	10-fold cross validation	0.92 (Accuracy)
Hosseinmardi et al. [15]	Image and text	Logistic Regression and visual features	Multi-modal posts of Instagram	5-fold cross-validation	0.68 (Recall)
Singh et al. [16]	Image and text	Bagging Classifier and visual features	Multi-modal posts of Instagram	Holdout validation	0.81 (Accuracy)
Kumari et al. [18]	Image	CNN	symbolic aggressive images	Holdout validation	0.89 (wgt. F1)
Kumari et al. [17]	Image and text	Unified representation and CNN	Multi-modal posts	Holdout validation	0.69 (wgt. F1)
Kumari and Singh [14]	Image and text	VGG-16, CNN and Genetic algorithm	Multi-modal posts	Holdout validation	0.78 (wgt. F1)

Bilingual and multilingual comments that have words from two or more languages in one post are very common in countries where English is not the native language. Recently the research community has started presenting solutions for bilingual and multilingual posts also. Kumari and Singh [31] introduced a four-layered CNN model with various

embedding techniques to detect hate speech in multilingual text comments. Models to classify social media multilingual posts into three classes: (i) Overtly Aggressive (OAG), (ii) Covertly Aggressive (CAG) and (iii) Non-aggressive (NAG) were proposed by [12, 10, 9, 11, 13] on a dataset released by organizers of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC - 1) at COLING - 2018. The results reported had a weighted F1-Score between 0.36 to 0.64.

Visual features have been the main focus of researchers [15, 16, 18] in recent times to identify cyberbullying and cyber-aggression from social media posts. The contribution of visual features and other user features such as the number of followers and followees was evaluated by Hosseinmardi et al. [15]. They, however, reported that visual characteristics do not help much in identifying cyberbullying posts. Singh et al. [16] on the other hand found visual features to be useful and utilized both textual and visual characteristics of comments to identify cyberbullying posts. Kumari et al. [18] considered only the image features for the identification of aggressive posts having symbolic images using a six-layered CNN model. They reported that their model outperformed the pre-trained VGG-16 [41] network with a weighted F1-Score of 0.89. Kumari et al. [17] proposed a model for cyberbullying detection for comments containing text and images together. They developed a unified representation for text and images to address the multi-modality of the post and got a weighted F1-Score of 0.69. Kumari and Singh [14] further extended their previous model [17] to optimize the combined features of text and images using the Genetic Algorithm (GA). They reported a performance improvement of about 4% with the optimized features of images and text compared to combined features of images and text without optimization.

Particle Swarm Optimization (PSO) based techniques have been used in several domains to optimize the parameters of neural network models [32, 33, 34, 35, 36]. Armaghani et al. [32] used the PSO algorithm for the optimization of the biases and weights of the Artificial Neural Network (ANN) to determine the ultimate bearing strength of rock-socketed piles. Similarly, Armaghani et al. [33] and Mohamad et al. [36] employed a PSO-based technique to optimize the biases and weights of their model for evaluating the uniaxial compressive strength of rocks. Hasanipanah et al. [35] used the PSO algorithm to develop a precise power equation to predict the trajectory of flyrock during blasting. Abad et al. [34] on the other hand used PSO to predict the durability of limestone aggregates. Population-based optimization techniques have hardly been used in multi-modal settings for cyber-aggression or cyberbullying detection, and that motivated us to utilize the BPSO algorithm for feature optimization in this setting. The summary of the related works with respect to the modality of data, type of dataset, methods used and results reported by them are placed in Table 1, where wgt. F1 refers to the weighted F1-Score. Interpretability, compatibility of contents, hyper-parameter tuning and development of the hybrid model are among several challenges for the fusion of multi-modal contents. Some researchers in their works [45, 46, 47] have discussed and addressed these issues for sentiment analysis using fuzzy logic and sentic blending techniques.

In recent years, a significant number of studies have been performed on automated detection of cyberbullying and cyber-aggression. Most of these studies have considered the textual component and overlooked the visual component of social media posts. However, the visual contents in the social media posts are on the increase, motivating us to

include these contents for cyber-aggression detection. So, we utilized visual characteristics in addition to textual characteristics in our research for detection of cyber-aggression on the one hand and optimization technique to get the important features from images and text on the other hand, as discussed in the next section.

3. Methodology

This section presents a model for automatic cyber-aggression detection of multi-modal social media posts. In the following subsection, we describe details of collection, labelling and description of our datasets. The proposed model is described next in Subsection 3.2.

3.1. Data Collection, Labelling and Description

Symbolic images are used by many people to annoy, threaten and humiliate other online users with the help of social media sites such as Twitter, Facebook and Instagram. We collected some of these images from Facebook, Twitter and Instagram. To find such images, we also used Google search with query terms like *cyber-aggressive images*, *aggressive images* and *bullying images* to increase the number of required images in the collection. Keeping in mind the level of aggression, the images were manually filtered and thus we finally got a total of 3,600 images. Among these collected images, some were associated with comments and some were without associated comments. The comment was composed and added to each of those images that did not carry it by a group of undergraduate students. The purpose of taking help from these students was to incorporate the thoughts of tender minds in online harassment activities. Considering both image and associated comment, each post was then manually labelled as either (i) Non-aggressive, (ii) Medium-aggressive or (iii) High-aggressive by three independent expert annotators. To do this job the annotators were instructed to label the post indicating physical threat as a 'High-aggressive post', the post having indirect aggression or aggression other than a physical threat as a 'Medium-aggressive post', and the post with no aggression as a 'Non-aggressive post'. After labelling independently, a majority voting scheme among the annotators was applied to assign the final label to each post. The count of different classes of posts in the created dataset is given in Table 2. In this dataset, 1,804 posts (50%) are labelled as Non-aggressive, 1,327 posts (37%) are labelled as Medium-aggressive and the remaining 469 posts (13%) are labelled as High-aggressive, as shown in Figure 5. The examples of our dataset for each class are shown in Figures 2, 3 and 4.

3.2. Proposed Model

The proposed model is a combination of VGG-16 network, three-layered CNN and BPSO algorithm. The schematic diagram of the model is shown in Figure 6. The proposed system comprises of two parallel neural architectures: (i) a VGG-16 network for processing image contents, (ii) a three-layered CNN for processing the text contents and (iii) a BPSO algorithm for optimizing hybrid features to classify the posts into (i) Non-aggressive, (ii) Medium-aggressive and (iii) High-aggressive classes. The different components of the model are described in the following subsections.

Table 2: Description of the dataset

Class of post	Number of posts
Non-aggressive	1,804
Medium-aggressive	1,327
High-aggressive	469
Total posts	**3,600**

Figure 2: High-aggressive post

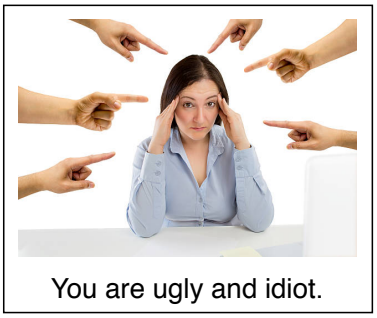


Figure 3: Medium-aggressive post

Figure 4: Non-aggressive post

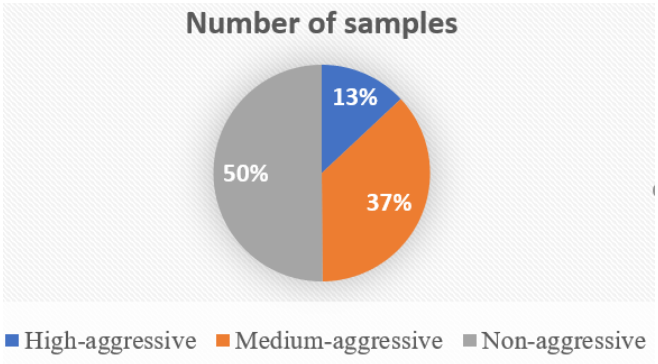


Figure 5: The proportion distribution of Non-aggressive, Medium-aggressive and High-aggressive posts

3.2.1. Text Feature Extraction

We used a three-layered CNN model for extracting features from the text. The reason behind selecting the CNN model is that the comment part of the post in our dataset is shorter with an average length of 25 words only, and this model performs well for the shorter length text [20]. For model development, text comment length was fixed to 30 words only. If text comment length was less than 30 words, padding with zero was done to make the length of the comment equal to 30 words. On the other hand, if the text comment length was more than 30 words then it was truncated after 30 words. The text comment was embedded into a 100 size vector using GloVe [37] embedding.

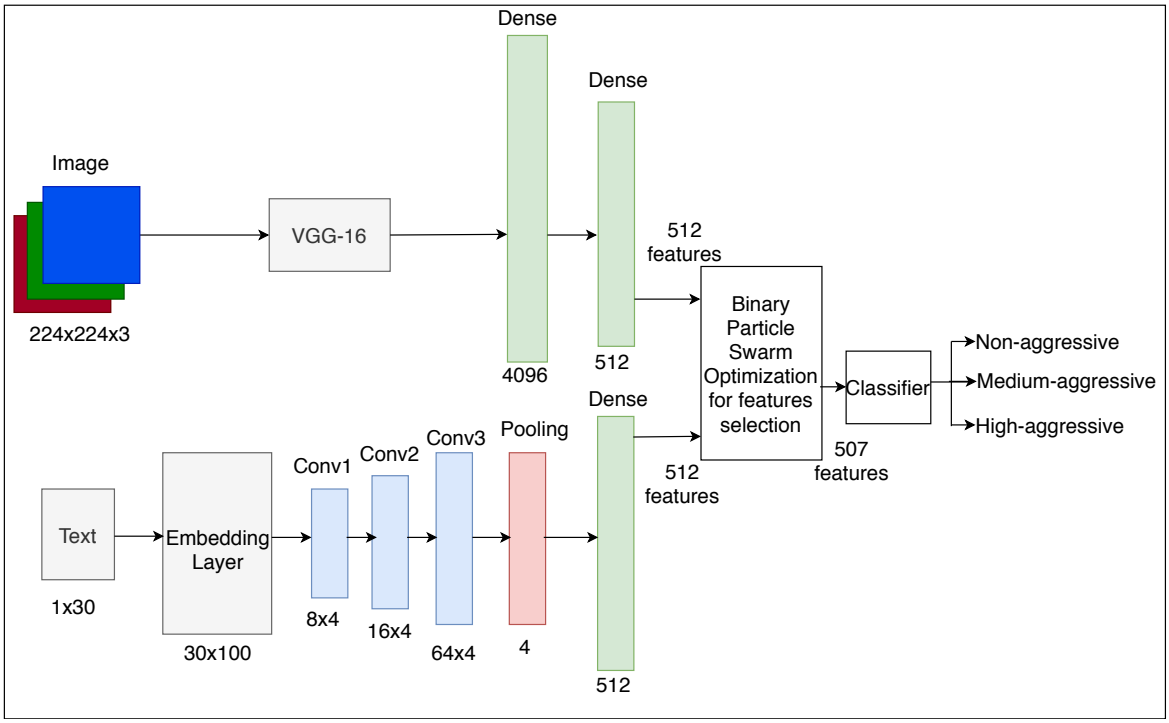


Figure 6: Overview of the proposed multi-modal architecture

A max-pooling layer of size four was used after three convolutional layers. The number of convolutional filters in each of the first, second and third layers was 8, 16 and 64, respectively with a filter size of 4 in each case. The data after max-pooling was flattened and the flattened layer was further reduced to size 512. We used Rectified Linear Unit (ReLU) and Softmax activation function in each convolutional and output (dense) layer, respectively. The output of the ReLU activation function for input (i), represented as ($g(i)$), can be calculated according to Equation 1. Over the predicted output classes, the non-normalized output is mapped by Softmax function into a normalized probability distribution. The output which is given by Softmax function is equivalent to a probability distribution and it gives the probability of the input being in a particular class. This function is mathematically represented as Equation 2, where z is a vector of the inputs to the output layer (if there are n output units, then we have n elements in z). And k gives index to the output units, hence $k = 1, 2, 3, ..., n$. The output of the dense layer of size 512 was stored as features of the text.

$$g(i) = \begin{cases} 0, & \text{if } i \leq 0 \\ i, & \text{otherwise} \end{cases} \tag{1}$$

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{n} e^{z_k}} \tag{2}$$

3.2.2. Image Feature Extraction

We experimentally evaluated several pre-trained deep neural network models for feature extraction from images and found that VGG-16 outperformed the others. The evaluation results are shown in Table 4. VGG-16 [41] model was used for extracting features from images. VGG-16 is made of 16 layers, out of which 13 are convolutional layers followed by three dense (fully connected) layers. Each input image is resized to $224 \times 224 \times 3$ for processing by the first layer of VGG-16. Each image passes through a stack of convolutional and max-pooling layers. After 13 convolutional layers, an image is flattened to a linear vector. Then the flattened image vector is passed through two dense layers of size 4096 and 512, respectively. ReLU activation function is used at the last dense layer. The output of the last dense layer having a size of 512 is taken as features of images.

3.2.3. Features Optimization and Classification

BPSO is a modified version of PSO algorithm applicable to discrete binary optimization problems. The PSO algorithm, proposed by Kennedy and Eberhart [38, 39], is a population-based optimization technique, which operates with a group of particles called a swarm. It utilises a number of particles (candidate solutions) flying in the searching space to get the best solution and is inspired by the social behaviour of bird flocking. While roaming in the search space, the particles think of their own best solution (*PBest*) together with the best solution (*GBest*) achieved so far by the swarm to move to the next location. The equations for updating velocity and position of the particles in PSO are given by Equation 3 and Equation 4, respectively.

$$v_m^{n+1} = \omega\, v_m^n + k_1\, r\, (PBest_m - x_m^n) + k_2\, r\, (GBest - x_m^n) \tag{3}$$

$$p_m^{n+1} = p_m^n + v_m^{n+1} \tag{4}$$

In the above equations, p_m^n denotes the position of m^{th} particle at iteration n and v_m^n is velocity of m^{th} particle at iteration n. $PBest_m$ is the best personal position obtained by the m^{th} particle and *GBest* is the global best position obtained by the whole swarm. *k1* and *k2* are acceleration parameters. r is a random number in the interval [0, 1]. ω is inertia weight which controls the balance between exploration phase and exploitation phase.

The important steps of PSO can be given as follows:

(i) Initialization of the parameters.

(ii) Fitness value evaluation of the swarm.

(iii) Updating the personal best (*PBest*) and global best (*GBest*).

(iv)(a) Update velocity od each particle $v_m^{n+1} = \omega\, v_m^n + k_1\, r\, (PBest_m - x_m^n) + k_2\, r\, (GBest - x_m^n)$

(b) Update position of each particle $p_m^{n+1} = p_m^n + v_m^{n+1}$

(v) Verifying the termination criteria;

if fulfilled then stop and return the best particle (solution);

otherwise go to Step (ii).

Many real-life problems like scheduling, routing, dimensionality reduction and feature selection have discrete binary search spaces where the binary version of PSO (BPSO) [40] is applicable. In addition to this, problems that have continuous real search space can be transformed into binary problems by changing the variables into binary variables. In BPSO, a solution is expressed by zeros ("0") and ones ("1"). Thus, *PBest* and *GBest* are constricted to binary values: "0" and "1". The two versions of PSO, continuous and binary, are identified by these two different components: (a) a new transfer function and (b) a different position updating procedure. To map a continuous search space to a binary search space, a transfer function is used and an updating method is designed to swap particles' positions between "0" and "1". As a transfer function part, a Logistic function is used in Equation 5 which is applied to convert all real values of velocities to probability values in the interval [0.0, 1.0].

$$\sigma(v_m^n) = \frac{1}{(1 + e^{-v_m^n})} \tag{5}$$

According to the probability of their velocities, positions could be updated as per Equation 6 after the conversion of velocities to the probability values as given below:

$$x_m^{n+1} = \begin{cases} 0, & \text{if } r < v_m^{n+1} \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

For our problem, we had a hybrid feature of size 1,024 obtained by concatenating 512 text features and 512 image features. BPSO algorithm is applied to get optimal features from this hybrid feature set as all the features may not be equally effective for classification. Each swarm is represented by a 1,024-dimensional vector having zero or one at each location of the vector randomly. A zero at a location represents that feature value is not considered whereas a one represents that it is considered. We used 50 swarm particles and iterated for 500 iterations. The BPSO algorithm selected the most relevant features after dropping irrelevant and redundant features to give 507 features out of 1,024 features. The feature optimization process is shown through a flowchart in Figure 7. After getting the most prominent features, we used Random Forest classifier to classify the posts into three classes: (i) Non-aggressive, (ii) Medium-aggressive and (iii) High-aggressive.

We have utilized VGG-16 and CNN for feature extraction, BPSO algorithm for features optimization and a machine-learning classifier for classification. Deep neural network models are well-known methods for automated feature extraction [10]. For classification, we have evaluated with several machine-learning classifiers (Naive Bayes, Decision Tree, KNN, Gradient Boosting and Random Forest) as discussed in the next section. The neural network model was built using Keras[5] library with TensorFlow at the back end. The CNN and VGG-16 networks were trained for 200 epochs with a batch size of 50, with *Adam* as an optimizer function and *Categorical cross-entropy* as a loss

[5]https://keras.io/

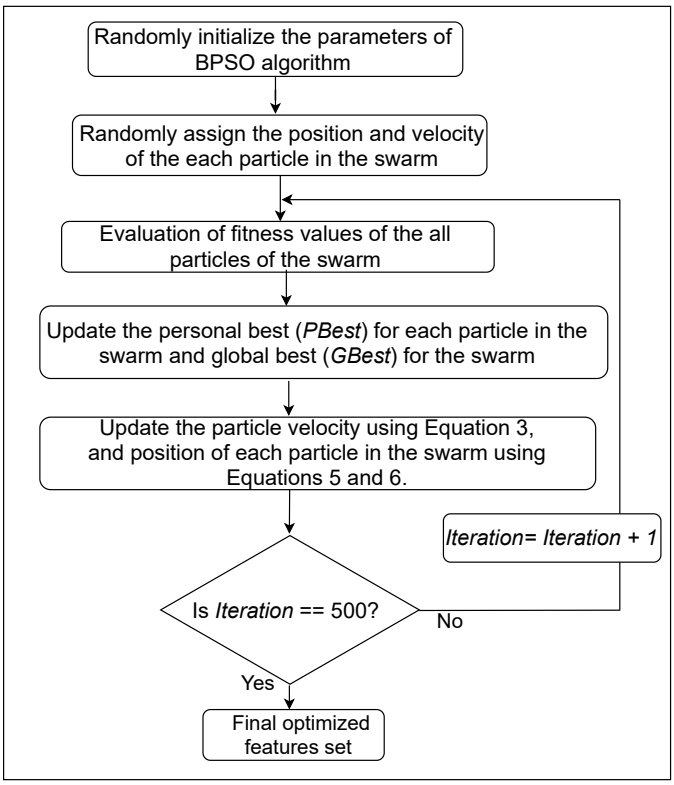


Figure 7: Flowchart of the proposed BPSO algorithm

function for the model. The dataset was divided into train and test samples in the ratio of 3:1 for all the experiments. The classifiers were implemented using Scikit-learn tool[6] in Python. The BPSO algorithm was developed in Python language using Pyswarms package library[7]. We have tried values of hyperparameters manually for our proposed models and found that those specific values which are mentioned in Table 3 are performing well with our dataset.

4. Results

Results obtained from the proposed model are presented in this section. The results are grouped into three subsections (i) Selection of pre-trained deep neural network model, (ii) Selection of the size of features from image and text and (iii) BPSO feature selection and classification for better presentation. To evaluate the performance of the current system, we have used three different performance metrics: Precision (P), Recall (R) and F1-Score. These performance metrics for High-aggressive class are defined below. The metrics for the other classes can be defined similarly.

$$P_{High-aggressive} = \frac{Number\ of\ accurately\ predicted\ High-aggressive\ post}{Total\ number\ of\ predicted\ High-aggressive\ post} \tag{7}$$

$$R_{High-aggressive} = \frac{Number\ of\ accurately\ predicted\ High-aggressive\ post}{Total\ number\ of\ actual\ High-aggressive\ post} \tag{8}$$

$$F1\text{-}Score_{High-aggressive} = 2 \times \frac{(P_{High-aggressive} \times R_{High-aggressive})}{(P_{High-aggressive} + R_{High-aggressive})} \tag{9}$$

4.1. Selection of Pre-trained Deep Neural Network Model

The first experiment was done to decide the more suitable pre-trained deep neural network model for extracting features from the image. We evaluated five different existing models (i) VGG-16 [41], (ii) VGG-19 [41], (iii) ResNet-50 [42], (iv) Inception [43] and (v) Xception [44] for extracting features from symbolic images for aggression identification. We found that VGG-16 is outperforming the other models as its performance in terms of F1-Score are 76%, 69% and 60% for Non-aggressive, Medium-aggressive and High-aggressive, respectively as shown in bold in Table 4. Hence, for our further experiments, we have used VGG-16 for extracting features from images.

4.2. Selection of Size of Features from Image and Text

Our second design parameter was to decide the size of the features of text and images separately. So, we experimented with different feature sizes of images and text to classify the posts. We tried different combinations of feature sizes such as the size of 256 for image and 128 for text; 256 for image and 256 for text; 256 for image and 512 for text; 256 for image and 1,024 for text and 512 for image and 512 for text. We found that the feature of size 512 for

[6]https://scikit-learn.org/stable/

[7]https://pypi.org/

Table 3: Description of the hyperparameters used in different models

Model	**Hyperparameters**	**Value**
CNN	Maximum length of comment	30
	Embedding dimension	100
	Number of convolutional layers	3
	Number of pooling layers	1
	Number of dense layers	1
	Number of filters	8, 16, 64
	Filter size	4
	Pooling size	4
	Dropout rate	0.2
	Learning rate	0.01
	Activation function	*ReLU*, *Softmax*
	Loss function	*Categorical cross-entropy*
	Optimizer	Adam
	Epoch	200
	Batch size	50
VGG-16	Image size	$224 \times 224 \times 3$
	Activation function	*ReLU*, *Softmax*
	Loss function	*Categorical cross-entropy*
	Optimizer	Adam
	Epoch	200
	Batch size	50
BPSO	Total number of features used (Dimensionality)	1024
	Number of optimized features	507
	Number of swarm particles used	50
	Value of inertia weight (ω)	0.9
	Value of the acceleration parameters (*k1* and *k2*)	0.5
	Iteration	500

Table 4: Results of detection of cyber-aggressive posts using pre-trained deep neural network models

Method	**Class**	**Results**		
		Precision	**Recall**	**F1-Score**
VGG-16 + CNN	Non-aggressive	0.73	0.79	**0.76**
	Medium-aggressive	0.69	0.68	**0.69**
	High-aggressive	0.71	0.52	**0.60**
	Weighted-average	0.71	0.71	**0.71**
VGG-19 + CNN	Non-aggressive	0.72	0.82	0.76
	Medium-aggressive	0.71	0.60	0.65
	High-aggressive	0.62	0.61	0.62
	Weighted-average	0.71	0.71	0.70
ResNet-50 + CNN	Non-aggressive	0.59	0.91	0.72
	Medium-aggressive	0.76	0.26	0.39
	High-aggressive	0.52	0.50	0.51
	Weighted-average	0.65	0.60	0.56
Inception + CNN	Non-aggressive	0.72	0.73	0.72
	Medium-aggressive	0.61	0.66	0.63
	High-aggressive	0.65	0.45	0.53
	Weighted-average	0.67	0.67	0.66
Xception + CNN	Non-aggressive	0.77	0.70	0.73
	Medium-aggressive	0.64	0.76	0.70
	High-aggressive	0.62	0.46	0.53
	Weighted-average	0.70	0.69	0.69

Table 5: Results of multi-modal data with a different combination of features

Image features size	**Text features size**	**Class**	**Results**		
			Precision	**Recall**	**F1-Score**
256	128	Non-aggressive	0.73	0.79	0.76
		Medium-aggressive	0.69	0.68	0.69
		High-aggressive	0.71	0.52	0.60
		Weighted-average	0.71	0.71	0.71
256	256	Non-aggressive	0.74	0.77	0.75
		Medium-aggressive	0.67	0.66	0.66
		High-aggressive	0.66	0.59	0.62
		Weighted-average	0.70	0.70	0.70
256	512	Non-aggressive	0.70	0.85	0.77
		Medium-aggressive	0.73	0.62	0.67
		High-aggressive	0.76	0.52	0.62
		Weighted-average	0.72	0.71	0.71
256	1,024	Non-aggressive	0.73	0.81	0.77
		Medium-aggressive	0.72	0.58	0.64
		High-aggressive	0.59	0.70	0.64
		Weighted-average	0.71	0.71	0.70
512	**512**	Non-aggressive	0.73	0.77	**0.75**
		Medium-aggressive	0.70	0.67	**0.69**
		High-aggressive	0.66	0.64	**0.65**
		Weighted-average	0.71	0.71	**0.71**

both image and text is performing better than other features combinations with F1-Score of 75%, 69% and 65% for Non-aggressive, Medium-aggressive and High-aggressive classes, respectively, as shown in bold in Table 5. Hence, for further evaluation, we have used the feature size of 512 for both images as well as for text.

4.3. Feature Selection and Classification

Our next design parameter was to select the optimum number of features from images and text. BPSO algorithm was used to find optimized features from the hybrid feature set of 1,024 (512 for images and 512 for text). Performance of the system is evaluated using five different classifiers: (i) Naive Bayes, (ii) Decision Tree, (iii) K-Nearest Neighbours (KNN), (iv) Gradient Boosting and (v) Random Forest, with hybrid features as well as with optimized

Table 6: Results of detection of cyber-aggressive posts using BPSO feature selection with different classifiers

Classifier	**Class**	**Results**		
		Precision	**Recall**	**F1-Score**
Naive Bayes	Non-aggressive	0.76	0.71	0.73
	Medium-aggressive	0.77	0.49	0.60
	High-aggressive	0.33	0.79	0.46
	Weighted-average	0.71	0.64	0.65
Decision Tree	Non-aggressive	0.74	0.69	0.71
	Medium-aggressive	0.63	0.65	0.64
	High-aggressive	0.56	0.67	0.61
	Weighted-average	0.68	0.67	0.67
KNN	Non-aggressive	0.75	0.80	0.77
	Medium-aggressive	0.69	0.67	0.68
	High-aggressive	0.68	0.57	0.62
	Weighted-average	0.72	0.72	0.72
Gradient Boosting	Non-aggressive	0.76	0.78	0.77
	Medium-aggressive	0.69	0.67	0.68
	High-aggressive	0.64	0.67	0.66
	Weighted-average	0.72	0.72	0.72
Random Forest	Non-aggressive	0.77	0.83	*0.80*
	Medium-aggressive	0.73	0.66	*0.69*
	High-aggressive	0.69	0.69	*0.69*
	Weighted-average	**0.74**	**0.75**	**0.74**

features. The obtained weighted F1-Score of the classifiers with hybrid features and with optimized features is presented by a bar-graph in Figure 8. As can be seen from Figure 8, the performance of classifiers is always better with optimized features compared to hybrid features. The detailed results of each class with different classifiers with optimized features are listed in Table 6. It can be seen from Table 6 that the Random Forest classifier outperforms the other classifiers in terms of a weighted average of Precision, Recall and F1-Score. The weighted average of Precision, Recall and F1-Score obtained for the Random Forest classifier are 74%, 75% and 74%, respectively, as shown in bold in Table 6.

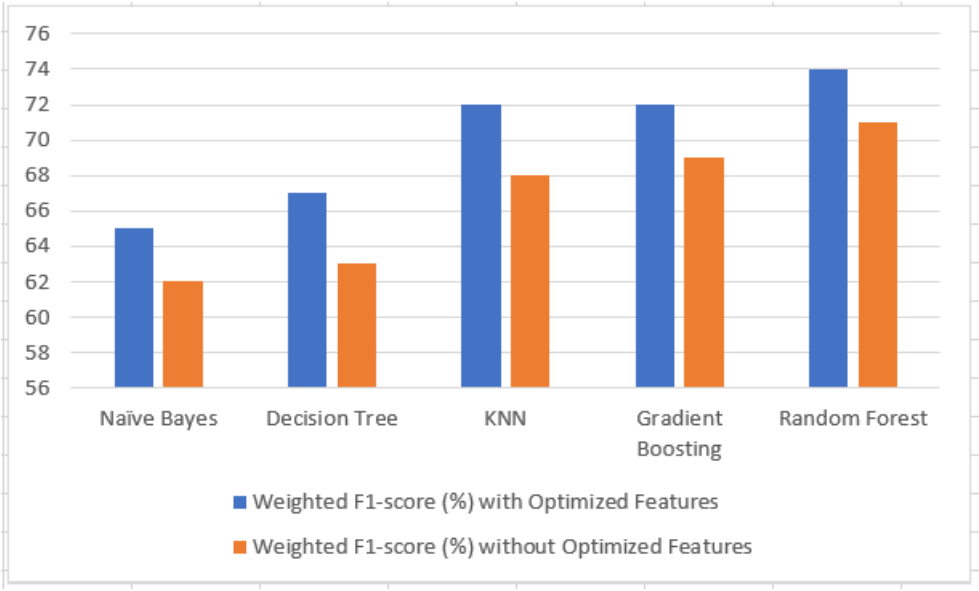


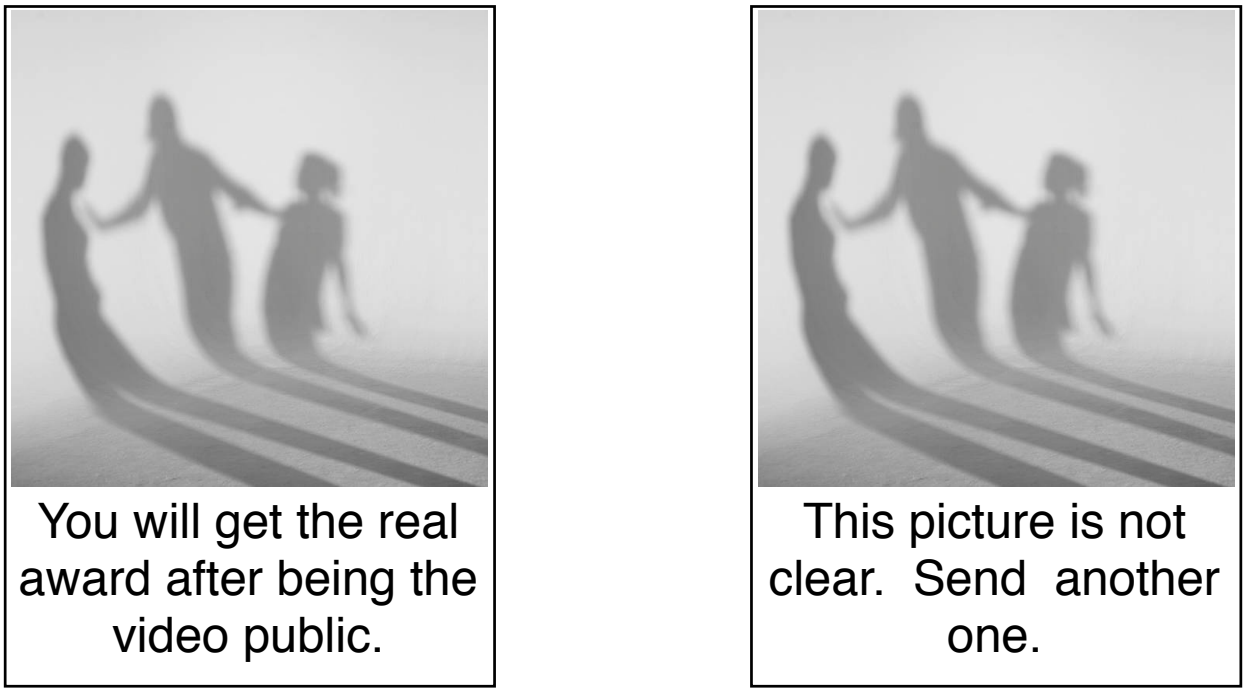
Figure 8: Performance comparison of the model with different classifiers with optimized features and without optimized features

5. Discussion and Implication

The major finding of current research is that BPSO is able to optimize feature space and this improves the classification performance with features obtained from the deep neural network model. Another major finding is that VGG-16 is a better model for extracting features from symbolic images. The next finding of this research is that a 512 sized image feature vector obtained by using CNN and 512 sized text feature vector jointly performed better to distinguish Non-aggressive, Medium-aggressive and High-aggressive posts. It is also found that the Random Forest classifier is outperforming the other classifiers with an optimized feature set with an F1-Score of 80%, 69% and 69% for Non-aggressive, Medium-aggressive and High-aggressive classes, respectively, which is shown in italic in Table 6.

We have found that the BPSO-based optimization algorithm significantly reduces the features space from 1,024 to 507 by eliminating the redundant features obtained from hybrid features of text and image. Sometimes the conclusion drawn from separate features of text and image may not indicate the sense of a post correctly and may contradict the conclusion drawn from combined features of text and image as shown in Figures 9a and 9b. In Figure 9a, separately both image and text features are non-aggressive but together they seem highly aggressive. Whereas, the same image with other non-aggressive text makes the post non-aggressive as shown in Figure 9b. It is evident from the examples that separate features of image and text are not fairly representative and may be contradictory also when compared with the combined features of text and image. In such a scenario, the optimization of hybrid features, especially in a multi-modal setting, may improve the performance of the model.

The current research enhances the field of feature optimization in the multi-modal setting. The results are in



(a) Aggressive post

(b) Non-aggressive post

Figure 9: Examples of posts having a symbolic image and text

line with the results of Kumari and Singh [14], where the genetic algorithm was used to optimize hybrid features of images and text for cyberbullying detection with a different dataset to achieve an improvement of about 4% with optimized features over unoptimized features. Our model also improved the results by 3% with optimized features over unoptimized features. A direct comparison of these two results is not possible as [14] was a two-class classification and we have a three-class classification with a different dataset.

The current system can be integrated with social media where people use text and images together to interact with other online users. As the feature extraction was done independently, if one part (image or text) is absent in a post then the features of that part will be filled by zeros and the features of the other part can classify the post. Hence, the proposed model can also work with the posts containing only text or only images.

The main limitation of our current system is that we have only considered the visual and textual information of the posts to detect cyber-aggression. However, the other contents such as audio, video, animated Graphics Interchange Formats (GIFs), memes and URLs are not considered for identifying such posts. The other limitation is that we have tested the system only with English comments, but it will be interesting to see the performance of the system for other languages as well as for multi-lingual comments. The current system also overlooked the users' behaviour after the occurrence of incidents and ignored the socio-demographic features such as age, gender, persons involved in such activities and their roles. This work is focused on the identification of aggression of social media posts and without considering the moderation task, which may also be effective to address the cyber-aggression issues.

6. Conclusion

Social media is affecting our society badly through hate speech, cyber-aggression and cyberbullying. To control and minimize the spread of online aggressive comments, the current research presents a hybrid model to detect aggressive posts containing images and text on social media. The proposed system used VGG-16 and CNN for the extraction of the features from image and text, respectively. The model also extracts the optimized feature set from hybrid features of images and text using the Binary Particle Swarm Optimization algorithm. The proposed model achieves a weighted F1-Score of 74% with a Random Forest classifier applied on the optimized feature set.

The present research can be further extended to include other modalities such as audio, video, animated GIFs, memes and URLs to detect aggressive posts. As future work, code-mixed comments including posts in more than one languages can be explored. The future system can also be extended by considering behavioural factors such as reactions of victims and variations in hours spent on social media by them; socio-demographic features such as age, gender, persons involved in the episode and their roles to add contrast in the aggression level of posts. Finally, active learning and unsupervised learning technique may be explored to reduce the labelling effort.

Acknowledgement

The first author would like to acknowledge the Ministry of Electronics and Information Technology (MeitY), Government of India, for financial support during the research work through Visvesvaraya Ph.D. Scheme for Electronics and IT.

References

[1] M. Lytras, A. Visvizi, L. Daniela, A. Sarirete, P. Ordonez De Pablos, Social networks research for sustainable smart education, Sustainability 10 (9) (2018) 2974.

[2] A. Visvizi, M. D. Lytras, E. Damiani, H. Mathkour, Policymaking for smart cities: Innovation and social inclusive economic growth for sustainability, Journal of Science and Technology Policy Management 9 (2) (2018) 126–133.

[3] Z. Zhang, B. B. Gupta, Social media security and trustworthiness: overview and new direction, Future Generation Computer Systems 86 (2018) 914–925.

[4] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, A. Vakali, Mean birds: Detecting aggression and bullying on Twitter, in: Proceedings of the 2017 ACM on web science conference, ACM, 2017, pp. 13–22.

[5] S. Salawu, Y. He, J. Lumsden, Approaches to automated detection of Cyberbullying: A survey, IEEE Transactions on Affective Computing 11 (1) (2020) 3–24.

[6] J. A. Pater, A. D. Miller, E. D. Mynatt, This digital life: A neighborhood-based study of adolescents' lives online, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, ACM, 2015, pp. 2305–2314.

[7] J. Kornblum, Cyberbullying grows bigger and meaner with photos, video, USA Today, dated July 17, 2008, Retrieved from https://cybercrimes.wordpress.com/2008/07/17/cyberbullying-growsbigger-and-meaner-with-photos-video/.

[8] K. Van Royen, K. Poels, W. Daelemans, H. Vandebosch, Automatic monitoring of Cyberbullying on social networking sites: From technological feasibility to desirability, Telematics and Informatics 32 (1) (2015) 89–97.

[9] K. Raiyani, T. Gonçalves, P. Quaresma, V. B. Nogueira, Fully connected neural network with advance preprocessor to identify aggression over Facebook and Twitter, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 28–41.

[10] S. Modha, P. Majumder, T. Mandl, Filtering aggression from the multilingual social media feed, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 199–207.

[11] N. S. Samghabadi, D. Mave, S. Kar, T. Solorio, Ritual-uh at TRAC 2018 shared task: Aggression identification, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 12–18.

[12] I. Arroyo-Fernández, D. Forest, J.-M. Torres-Moreno, M. Carrasco-Ruiz, T. Legeleux, K. Joannette, Cyberbullying detection task: the ebsilia-unam system (elu) at COLING'18 TRAC-1, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 140–149.

[13] J. Risch, R. Krestel, Aggression identification using deep learning and data augmentation, in: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), 2018, pp. 150–158.

[14] K. Kumari, J. P. Singh, Identification of Cyberbullying on multi-modal social media posts using genetic algorithm, Transactions on Emerging Telecommunications Technologies (2020) e3907doi:10.1002/ett.3907.

[15] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv, S. Mishra, Prediction of Cyberbullying incidents in a media-based social network, in: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2016, pp. 186–192.

[16] V. K. Singh, S. Ghosh, C. Jose, Toward multi-modal Cyberbullying detection, in: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, ACM, 2017, pp. 2090–2099.

[17] K. Kumari, J. P. Singh, Y. K. Dwivedi, N. P. Rana, Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach, Soft Computingdoi:10.1007/s00500-019-04550-x.

[18] K. Kumari, J. P. Singh, Y. K. Dwivedi, N. P. Rana, Aggressive social media post detection system containing symbolic images, in: Conference on e-Business, e-Services and e-Society, Springer, 2019, pp. 415–424.

[19] K. Kumari, J. P. Singh, AI_ML_NIT_Patna @ TRAC - 2: Deep Learning Approach for Multi-lingual Aggression Identification, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (May 2020), 2020, pp. 113–119.

[20] J. Chen, S. Yan, K.-C. Wong, Verbal aggression detection on Twitter comments: Convolutional Neural Network for short-text sentiment analysis, Neural Computing and Applications (2018) 1–10.

[21] P. Burnap, M. L. Williams, Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & Internet 7 (2) (2015) 223–242.

[22] M. A. Al-garadi, K. D. Varathan, S. D. Ravana, Cybercrime detection in online communications: The experimental case of Cyberbullying detection in the Twitter network, Computers in Human Behavior 63 (2016) 433–443.

[23] H. Chen, S. Mckeever, S. J. Delany, Harnessing the power of text mining for the detection of abusive content in social media, in: Advances in Computational Intelligence Systems, Springer, 2017, pp. 187–205.

[24] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for Hate Speech detection on Twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

[25] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated Hate Speech detection and the problem of offensive language, in: Eleventh International AAAI Conference on Web and Social Media, 2017, pp. 512–515.

[26] R. Zhao, K. Mao, Cyberbullying detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder, IEEE Transactions on Affective Computing 8 (3) (2017) 328–339.

[27] S. D. Bhattacharjee, W. J. Tolone, V. S. Paranjape, Identifying malicious social media contents using multi-view context-aware active learning, Future Generation Computer Systems 100 (2019) 365–379.

[28] F. R. Gallo, G. I. Simari, M. V. Martinez, M. A. Falappa, Predicting user reactions to Twitter feed content based on personality type and social cues, Future Generation Computer Systemsdoi:10.1016/j.future.2019.10.044.

[29] A. Visvizi, J. Jussila, M. D. Lytras, M. Ijäs, Tweeting and mining OECD-related microcontent in the post-truth era: a cloud-based APP, Computers in Human Behavior (2019) 105958doi:10.1016/j.chb.2019.03.022.

[30] S. Sadiq, A. Mehmood, S. Ullah, M. Ahmad, G. S. Choi, B.-W. On, Aggression detection through deep neural model on Twitter, Future Generation Computer Systems (2020) doi:10.1016/j.future.2020.07.050.

[31] K. Kumari, J. P. Singh, AI_ML_NIT Patna at HASOC 2019: Deep learning approach for identification of abusive content, in: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019), 2019, pp. 328–335.

[32] D. J. Armaghani, R. S. N. S. B. Raja, K. Faizi, A. S. A. Rashid, et al., Developing a hybrid PSO-ANN model for estimating the ultimate bearing capacity of rock-socketed piles, Neural Computing and Applications 28 (2) (2017) 391–405.

[33] D. J. Armaghani, E. T. Mohamad, M. S. Narayanasamy, N. Narita, S. Yagiz, Development of hybrid intelligent models for predicting TBM penetration rate in hard rock condition, Tunnelling and Underground Space Technology 63 (2017) 29–43.

[34] S. V. A. N. K. Abad, M. Yilmaz, D. J. Armaghani, A. Tugrul, Prediction of the durability of limestone aggregates using computational techniques, Neural Computing and Applications 29 (2) (2018) 423–433.

[35] M. Hasanipanah, D. J. Armaghani, H. B. Amnieh, M. Z. A. Majid, M. M. Tahir, Application of PSO to develop a powerful equation for prediction of flyrock due to blasting, Neural Computing and Applications 28 (1) (2017) 1043–1050.

[36] E. T. Mohamad, D. J. Armaghani, E. Momeni, A. H. Yazdavar, M. Ebrahimi, Rock strength estimation: a PSO-based bp approach, Neural Computing and Applications 30 (5) (2018) 1635–1646.

[37] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[38] R. Eberhart, J. Kennedy, A new optimizer using particle swarm theory, in: MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, IEEE, 1995, pp. 39–43.

[39] J. Kennedy, R. Eberhart, Particle swarm optimization, in: Proceedings of ICNN'95-International Conference on Neural Networks, Vol. 4, IEEE, 1995, pp. 1942–1948.

[40] J. Kennedy, R. C. Eberhart, A discrete binary version of the particle swarm algorithm, in: 1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation, Vol. 5, IEEE, 1997, pp. 4104–4108.

[41] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[42] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[43] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167.

[44] F. Chollet, Xception: Deep learning with depth wise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[45] E. Cambria, N. Howard, J. Hsu, A. Hussain, Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics, in: Proceedings of the IEEE symposium on computational intelligence for human-like intelligence (CIHLI), 2013, pp. 108–117.

[46] E. Cambria, Affective Computing and Sentiment Analysis, IEEE Intelligent Systems 31 (2) (2016) 102–107.

[47] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognition Letters 125 (2019) 264–270.