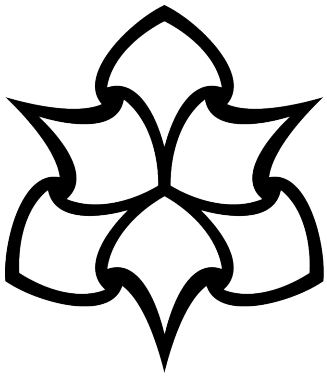


# Multimodal Cyberbullying Detection Using Capsule Network with Dynamic Routing and Deep Convolutional Neural Network

Akshi Kumar[1*], Nitin Sachdeva[2]
[1,2]Department of Computer Science & Engineering, Delhi Technological University, Delhi, India
*akshikumar@dce.ac.in, nits.usit@gmail.com

**Abstract:** Cyberbullying is the use of information technology networks by individuals' to humiliate, tease, embarrass, taunt, defame and disparage a target without any face-to-face contact. Social media is the "virtual playground" used by bullies with the upsurge of social networking sites such as Facebook, Instagram, YouTube, Twitter etc. It is critical to implement models and systems for automatic detection and resolution of bullying content available online as the ramifications can lead to a societal epidemic. This paper presents a deep neural model for cyberbullying detection in three different modalities of social data, namely, textual, visual and info-graphic (text embedded along with an image). The all-in-one architecture, CapsNet-ConvNet, consists of a Capsule network (CapsNet) deep neural network with dynamic routing for predicting the textual bullying content and a convolution neural network (ConvNet) for predicting the visual bullying content. The info-graphic content is discretized by separating text from the image using Google Lens of Google Photos App. The perceptron based decision-level late fusion strategy for multimodal learning is used to dynamically combine the predictions of discrete modalities and output the final category as bullying or non-bullying type. Experimental evaluation is done on a mix-modal dataset which contains 10000 comments and posts scrapped from YouTube, Instagram and Twitter. The proposed model achieves a superlative performance with the AUC-ROC of 0.98.

**Keywords:** Cyberbullying, Multimodal, Deep learning, Convolution neural network, Capsule network

## 1. Introduction

Bullying is an adverse societal issue which is rising at an alarming rate. In general, a bullying behavior can be categorized on the basis of type of behavior (verbal, social and physical), the environment (in person and online), its mode (direct and indirect), the visibility (overt and covert), the damage caused (physical and psychological) and the context in terms of place of occurrence (home, workplace and school)[1]. Cyberbullying is typically a social behavior bullying within the online setting done covertly by direct or indirect means causing short-term and long-term psychological harm. The increasing availability of reasonable data services and social media presence has given some uninhibited effects where online users have discovered wrong & unlawful ways to harm and humiliate individuals through hateful comments on online platforms or apps. The persistence, audience size and damage speed makes cyberbullying even more damaging than face-to-face bullying causing serious mental health and wellbeing issues to victims and making them feel totally overwhelmed. Cyberbullying can result in increased distress for the victims along with low self-esteem, increased anger, frustration, depression, social withdrawal and in some cases, developing violent or suicidal traits [1-3].

Technology allows the bullies to be anonymous, hard to trace and insulated from confrontation. To the targets of cyberbullying, it feels invasive and never-ending. With the amount of emotional and psychological distress caused to victims it is urgently required to find appropriate provisions which can detect and prevent it. Effective prevention relies on the timely and satisfactory detection of potentially toxic posts [4]. The information overload on the chaotic and complex social media portals necessitates advanced automatic systems to identify potential risks pro-actively. Researchers worldwide have been trying to develop new ways to detect cyber bullying,

[1] https://bullyingnoway.gov.au/WhatIsBullying/Pages/Types-of-bullying.aspx

manage it and reduce its prevalence on social media. Advanced analytical methods and computational models for efficient processing, analysis and modeling for detecting such bitter, taunting, abusive or negative content in images, memes or text messages are imperative. More recently, as memes, online videos and other image-based, inter-textual content have become customary in social feeds; typo-graphic and info-graphic visual content (Fig.1) have become a considerable element of social data [5, 6].

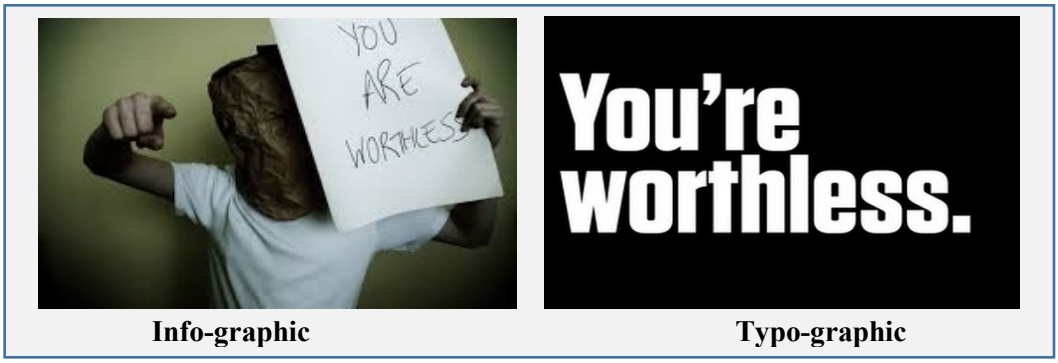



**Fig.1.** Types of visual content

Cyberbullying through varied content modalities is very common. Social media specificity, topic dependence and variety in hand-crafted features currently define the bottlenecks in detecting online bullying posts [2]. Deep learning methods are proving useful and obtaining state-of-the-art results for various natural language tasks with end-to-end training and representation learning capabilities [7]. Pertinent studies report the use of deep learning models like CNN, RNN and semantic image features for bullying content detection by analyzing textual, image based and user features [2, 8].But most of the research on online cyber-aggression, harassment detection and toxicity has been limited to text-based analytics. Few related studies have also re-counted analysis of images to determine bullying content but the domain of visual text which combines both text and image has been least explored in literature. The combination can be observed in two variants: typo-graphic (artistic way of text representation) or info-graphic (text embedded along with an image). This paper presents a deep neural model for bullying content prediction, where the content, c ε {text, image, info-graphic}.The primary contribution of the work is:

- The all-in-one hybrid deep architecture, CapsNet-ConvNet, consists of a Capsule network (CapsNet) with dynamic routing [9, 10] for predicting the textual bullying content and convolution neural network (ConvNet) [11] for predicting the visual bullying content.
- The info-graphic content is discretized by separating text from the image using Google Lens of Google Photos App[2].
- The processing of textual and visual components is carried out using the hybrid architecture and late-fusion decision layer is then used to output the final prediction.
- The performance of CapsNet-ConvNet is validated on 10000 comments and posts (text, image, and info-graphic) prepared using three social media sites YouTube, Instagram and Twitter. The results of textual processing module are compared against state-of-the-art *toxic comment classification challenge* dataset[3] from Kaggle competition whereas the results of the visual processing module are compared with baseline machine learning classifiers.

[2] https://photos.google.com/
[3] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

This unifying model thus considers modalities of content and processes each modality type using a deep neural learning techniques for an efficient decision support for cyberbullying detection. The paper organization is as follows: Section 2 discusses the related work followed by the description of the proposed CapsNet-ConvNet model for cyberbullying detection in multi-modal online content in section 3. Section 4 presents the results and finally the conclusion is given in section 5.

## 2. Related Work

Cyber space is 'Virtual Society' where we can get huge information and also share information. Most importantly we express ourselves- our understanding, our opinions and our views. Given the amount of increasing users and reach of cyberspace, cyberbullying is a definitely is a serious social concern. Recent literature accounts the use of machine learning methods for classifying hate speech, aggression, comment toxicity and bullying content on social forums. Dinakar et al. [12] constructed a common sense knowledge base that analyzed knowledge about bullying situations and the messages on Formspring website. Hinduja et al. [13] explored the relationship between cyberbullying and suicide among adolescents. The characteristic profile of wrongdoers and victims and conceivable strategies for its preventions were introduced in [14]. Till now the majority of the work is devoted to text analysis [15-27].

The use of deep learning models has also been reported [8]. Agrawal et al. [28] experimented with four deep neural models for cyberbullying detection which included CNN, LSTM, BLSTM, and BLSTM with attention on multiple social networks such as Twitter, Formspring and Wikipedia. Huang et al. in [29] concatenated social network features with textual features for improved performance of cyberbullying detection. Authors analyzed features such as number of friends, network embeddedness, and relationship centrality with the textual features and observed significant improvement in performance. Some rule based classifications for the identification of bullies have also been done in [30] using FormSpring data set. A new sentence-level filtering model was established by Xu et al. [31] which semantically eliminates bullying words in texts by utilizing grammatical relations among words on the YouTube dataset. Works have been done where pictures are utilized for the discovery of cyberbullying utilizing deep learning models like CNN, RNN or where semantic image features are utilized for identifying bullying [32-35].

The pertinent literature reports automated models of mono-modal (primarily text only) cyberbullying detection in social media. Few recent studies also report multimodal cyberbullying detection in online platforms. A framework was proposed by Kansara et al. in [36] with Bag-of-Visual-Word (BoVW) model, local binary pattern (LBP) and SVM classification for image analysis and Bag-of-Word (BoW) model with Naïve Bayesian classification for text analysis. Singh et al. [37] used some visual features along with textual features to improve the accuracy results. Yang et al. [38] classified the posts with images using deep multimodal fusion techniques, including simple concatenation, bilinear transformation, gated summation, and attention mechanism. The work in this paper aims to build a three-in-one modality model based on deep learning which not only predicts cyberbullying in textual or visual content but also mix-modal info-graphic content. The details of the proposed CapsNet-ConvNet model is given next.

## 3. The proposed CapsNet-ConvNet model

The proposed deep neural model comprehends the complexities of natural language and deals with different modalities of data in online social media content where the representations of data in

different forms, such as text and image, is learned as real-valued vectors. In addition to text, we examine the image as well as utilize info-graphic property of the image (information which is the content/text embedded on that picture) to predict bullying content. The proposed CapsNet-ConvNet model consists of four modules, namely, modality discretization module, textual processing module, visual processing module, and prediction module (fig.2).

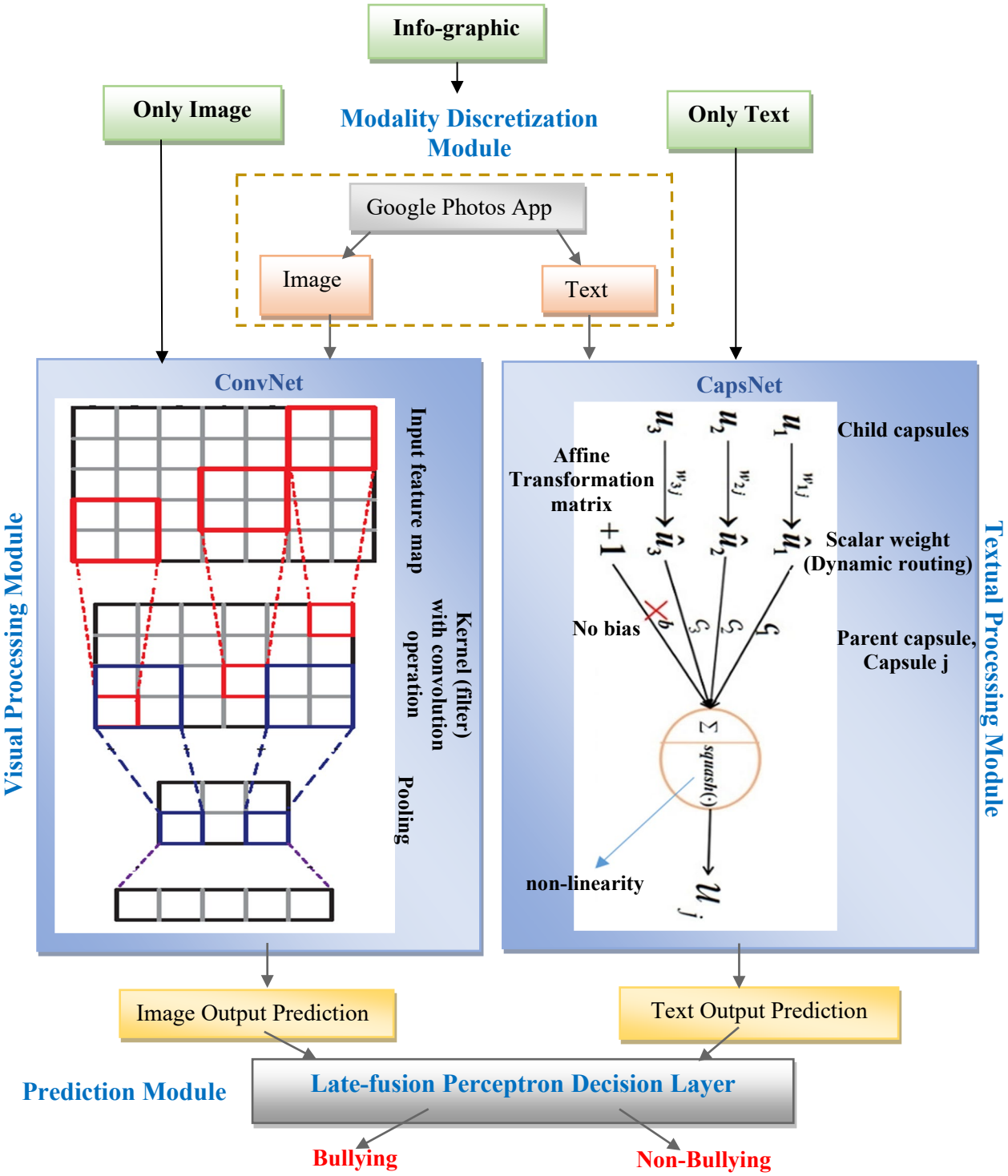



**Fig. 2.** The proposed CapsNet-ConvNet model

The details of each module are as follows:

### 3.1. Modality Discretization Module

Depending on the input modality, that is text only or image only, the content is forwarded to the respective processing modules. If the input is an info-graphic post/ comment, which is the image with text embedded on it, the CapsNet-ConvNet model utilizes a Google Photos App to extract text from an image. This visual analysis tool separates the text from the image and sent to the respective textual processing and visual processing modules for analysis. The Google Lens feature has the ability to recognize texts in the images recorded utilizing the Optical Character Recognition (OCR). The Google Cloud's Vision API offers powerful pre-trained machine learning models which can detect and extract text from images. There are *two* annotation features that support OCR, namely TEXT_DETECTION that detects and extracts text from any image and DOCUMENT_TEXT_DETECTION which extracts text from an image, but the response is optimized for dense text and documents.

### 3.2. Textual Processing Module: *CapsNet with Dynamic Routing*

Capsule Network (CapsNet) belongs to the class of deep neural networks comprising of set of capsules [39]. These are further composed of groups of neurons arranged in a layer and do the actual internal computations in order to predict instantiation parameters of any feature, such as orientation, color etc. at any given location. Pertinent literature reports the use of many routing techniques for text classification such as dynamic, attention based, clustering, static, where dynamic routing reported major applicability. In this research, the workflow of CapsNet [3] involves the following steps:

- The embedding layer of a neural network converts an input from a sparse representation into a distributed or dense representation. In this research, we use the state-of-the-art pre-trained ELMo 5.5B word embeddings model [40] to generate the word vectors. We preferred ELMo over the conventional embedding models such as Word2Vec or GloVe, as ELMo offers contextualized word representations, which essentially means that the representation for each word depends on the entire context in which it is used. The same word can have two different vector representations based on different contexts. ELMo creates vectors on-the-go by passing words through the deep learning model rather than having a dictionary of words and their corresponding vectors, as is the case with traditional word embedding models. Also, ELMo representations are purely character-based, which allows the network to form representations for words that are not seen in training. All this motivated us to use the ELMo 5.5B model for implementing the embedding layer.
- Encoding layer, thereafter, reshapes the word vector matrix into feature vectors of single dimension, where this encoding layer is executed as capsule network. This network comprises of convolution, primary caps and class caps layers. Here, the scalar outputs of each convolution layer are fed as input to primary caps layer that generates capsules.
- It must be noted that the output of a capsule is a vector that exhibits the object's existence. Whereas, the vector's orientation represents the object's properties. The vector is passed to all the possible parents in the network.
- These capsules work towards detecting the parts of the object under consideration in order to associate the random parts of the object to the whole.

- For accomplishing this, CapsNet uses non-linear-dynamic routing algorithm in order to capture the capsules part-whole relationship dynamics. Thus, ensuring that the output of the capsule is sent to the possible and relevant parent.
- Lower level capsule vectors are multiplied with weight matrices in order to encode spatial and other relationships between features of lower and higher level using equation 1.

$$u_{j|i} = W_{ij} u_i \qquad (1)$$

Where 'i' is low level capsule, 'j' is high level capsule and $W_{ij}$ is the translation matrix

- Lower level capsule knows which upper level capsule accommodates its results in an efficient way and therefore adjust it coupling coefficient. Thus previous step output is multiplied with coupling coefficients using equation 2.

$$s_j = \sum_i c_{ij}\, u_{j|i} \qquad (2)$$

Where $c_{ij}$ is coupling coefficient and $u_{j|i}$ is output vector from equation 1.

- Post this, squashing is applied for normalizing the length of each capsule's output vector in the range of [0, 1] using equation 3.

$$v_j = \frac{||s_j||^2}{1+||s_j||^2} \frac{s_j}{||s_j||} \qquad (3)$$

### 3.3. Visual Processing Module: *ConvNet*

To analyze visual bullying content the model uses a convolution neural network (ConvNet) [11]. A ConvNet is a deep neural architecture which works using multiple copies of the same neuron in different places. It has the power of self-tuning & learning skills by generalizing from the training data. The visual processing is shown is fig.3.

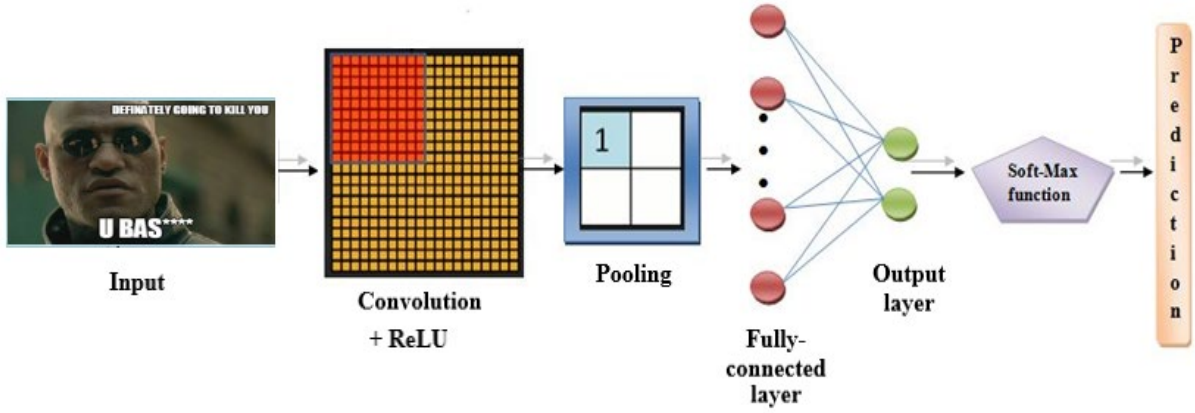



**Fig.3.** Visual Processing module

A ConvNet convolves learned features with input data and uses 2D convolutional layers. It usually consists of several convolutional networks with filters (kernels) in combination with non-linear and pooling layers [41]. The image is passed through convolution layers such that the output of the primary layer becomes the input for the subsequent layer. Convolution is a linear operation but images are non-linear. Therefore, non-linearity is added post every convolution operation using an activation function such as ReLU, Leaky_ReLU, tanh or sigmoid. Each non-linear layer is followed by a pooling layer which reduces the spatial size of the image and performs a downsampling operation. Pooling operation thus helps to progressively reduce the size of the input representation and control overfitting too. We can either use max, average or sum pooling. A fully connected layer is then attached to this series of convolution, non-linear and pooling layers which outputs the information from the convolutional networks. The working of a typical ConvNet is shown in fig. 4.

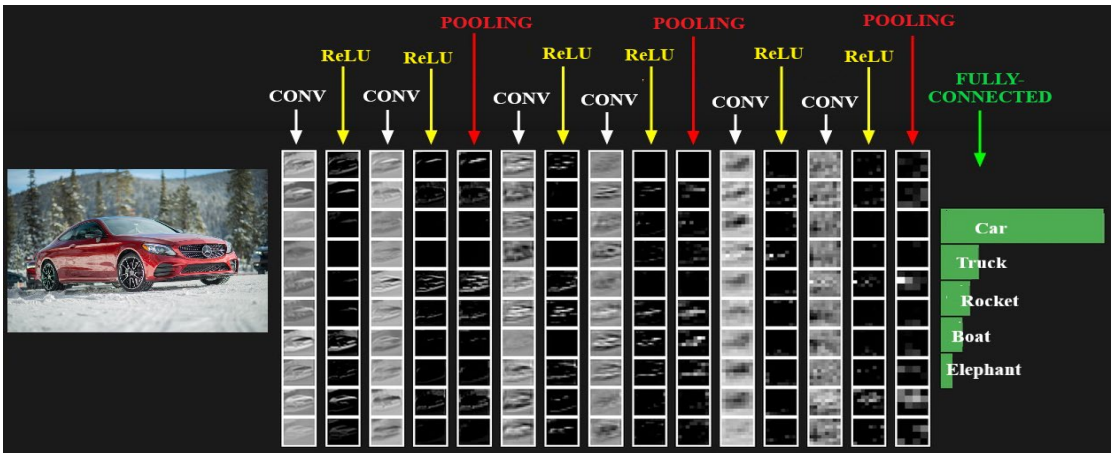



**Fig. 4.** Working of a typical ConvNet

In this work, the visual processing module has 3 convolutional layers followed by 3 max-pooling layers to extract the features of images, a flatten layer which takes the output from the previous max-pooling layer and convert it to a 1D array such that it can be feed into the dense layers and finally, 2 dense layers. The dense layer is a standard layer of neurons where the actual learning is done by adjusting the weights. Here we have 2 such dense layers and since this is a binary classification only 1 neuron there is in the output layer. The details of the layers are as follows:

- **Convolution layer:** The convolution layer transforms the input image to extract the features. This is done by convolving the image with a filter (kernel) which is specialized to extract certain features. Mathematically, the convolution operation (a.k.a. scalar product) is the summation of the element-wise product of two matrices (filter-sized patch of the input and filter) which results in a single value.
- **Activation layer and Pooling layer:** The activation (ReLU) layer is intended to introduce non-linearity to the system and produces a rectified feature map which is inserted into the pooling layer where a max-pooling operation is applied to each convolution $c_{max} = \max(c)$. The max-pooling operation extracts the 'k' most important features for each convolution. The output of the final convolution layer, that is, the pooled feature map is a representation of our original image.
- **Fully Connected layer:** A fully connected neural network is a feed forward network that will have the feature vector of n dimension obtained after concatenating every $c_i$ obtained by the

application of n filters. Now we train the network using back-propagation algorithm. Gradients are back propagated and when we reach at the convergence we finally stop the algorithm. A softmax function is used to classify the post as bullying (+1) or non-bullying (-1).

### 3.4. Prediction Module

The final prediction is done using an additional decision layer implementing multimodal classification fusion. Typically, there are two strategies to multimodal fusion: model-free and model-level. Model-free fusion can be further classified into early fusion (feature-level) and late fusion (decision-level). In early fusion, the different types of input features are firstly concatenated and then fed into a classifier, whereas in late fusion, the predictions of different classifiers trained for distinct input types are combined to provide us with the final output. Model-level fusion combines the advantages of both of these strategies by concatenating high-level feature representations from different classifiers. In this work, late fusion strategy for multimodal learning is used, that is, the bullying content prediction of mono-modalities (text and image separately) is done by the respective classification models. Late fusion allows the use of different models on different modalities, thus allowing more flexibility. It is easier to handle a missing modality as the predictions are made separately. The class probabilities are thus fused together to join information from the two modalities to perform a final prediction task. We use a multi-layer perceptron (MLP) neural network with sigmoid activation function to implement the decision-level fusion.

MLP can be thought of as a linear classifier, which means that it can segregate two different entities or classes from each other using a straight line. The input to a perceptron is usually a feature vector x, which is multiplied to a weight w and then finally added to a bias b as given in equation 4.

$$y = w * x + b \tag{4}$$

A perceptron is a shallow neural network and thus incapable of solving classification problems in which the number of classes is more than two. It takes in a number of inputs and generates an output by forging a linear coalition by utilizing the weights of its inputs. It also, sometimes, passes the output through a non-linear activation function. This can be shown through the following equation 5:

$$y = \varphi(\sum_{i=1}^{n} \quad w_i x_i + b) \tag{5}$$

where, w stands for the weight vector, x stands for the input vector, b stands for the bias, and phi represents the non-linear activation function.

## 4. Results and Discussions

The dataset prepared for experiments contains 10000 comments and posts (text, image, and info-graphic) prepared using three social media sites YouTube, Instagram and Twitter. The modalities within the dataset were 60% textual, 20% visual and 20% info-graphic (Fig.5).

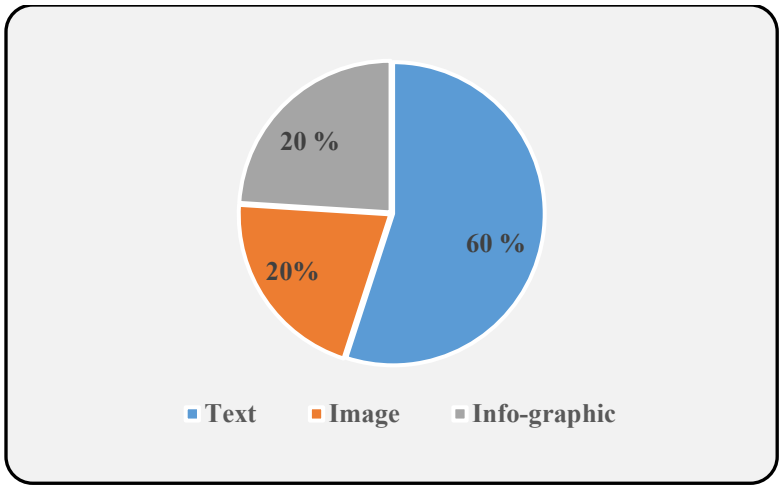



**Fig. 5.**Modality Distribution in Dataset

Table 1 below shows the actual distribution of data in numbers.

**Table 1.**Categorization of data used for training

| **Type of modality** | **Number of instances** | |
|---|---|---|
| | **Bullying** | **Non-bullying** |
| **Image only** | 1260 | 740 |
| **Text only** | 3000 | 3000 |
| **Info-graphic** | 1440 | 560 |

We performed 10-fold cross validation and calculated the AUC-ROC curve. We used the Scikit-learn library and Keras deep learning library with Theano backend. We used the Scikit-learn library and Keras deep learning library with Theano backend. Since two different models were used based on the input-type, the tuning of each model's respective hyperparameters was performed. The choice of model hyperparameters considered are given in table 2.

**Table 2.** Model hyperparameters

| | Hyperparameter | Value |
|---|---|---|
| Word embeddings | ELMo 5.5B | |
| CapsNet | No. of convolution layers | 1 |
| | Batch size | 64 |
| | No. of filters | 32 |
| | Activation | ReLU |
| | Learning rate | 0.001 |
| | Kernel_size | 3 |
| | No. of capsule in PrimaryCaps layer | 8 |
| | No. of nodes in PrimaryCaps layer | 64 |
| | Routing Time | 2 |
| | Dimension of each capsule | 8 |
| | Length of PrimaryCaps | 2 |
| | Length of Digi Caps | 2 |
| | Optimizer | Adam |
| ConvNet | Number of filters | 150 of each size |
| | Filter size | 2, 3, 4 and 5 |

| | | |
|---|---|---|
| | Drop out | 0.5 |
| | Epochs | 5 |
| | Non-linearity function | ReLU |

The proposed model achieves a performance of 0.98 (Fig.6). Confusion matrices for all type of modalities are shown in fig.7.

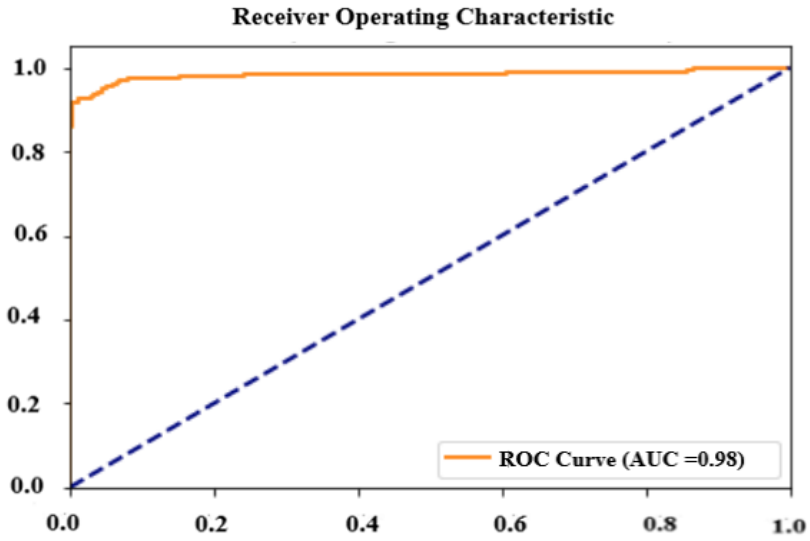
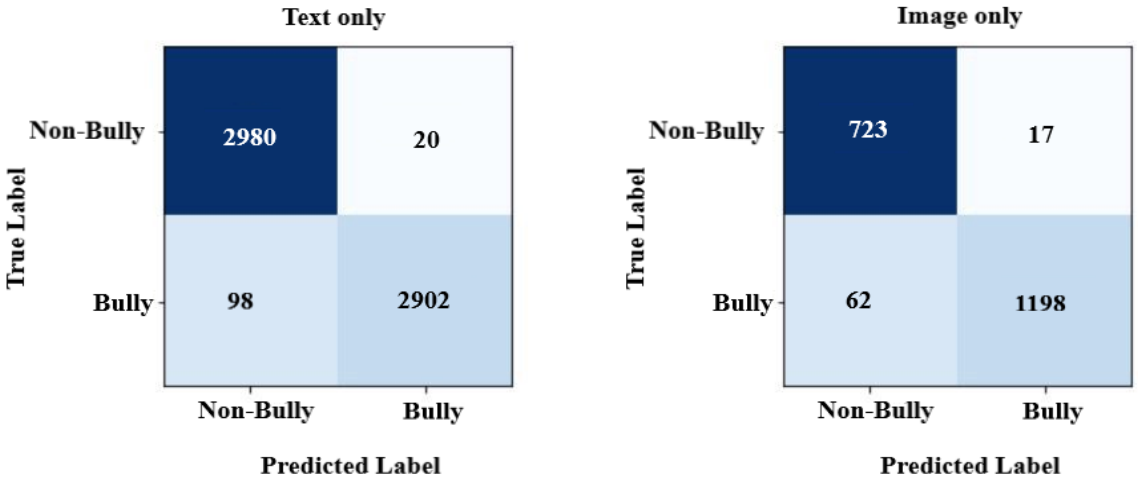



**Fig.6.** Performance of CapsNet-ConvNet Model



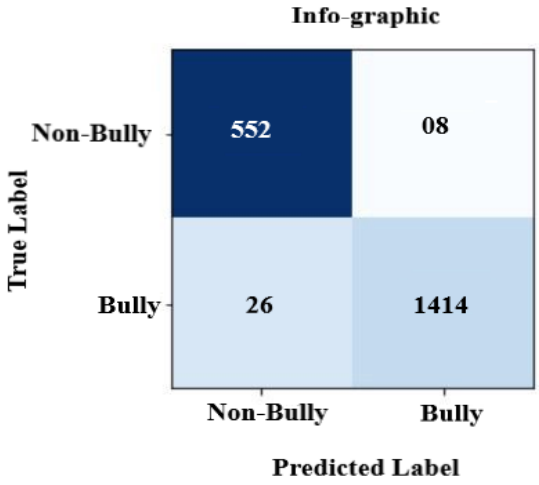



**Fig.7.** Confusion matrices for all modalities

As we proposed training a CapsNet model for textual content, it was imperative to evaluate the robustness of this module. In our previous work [3], we proposed a multi-input integrative learning model based on deep neural networks (MIIL-DNN) which combined information from three sub-networks to detect & classify bully content in real-time code-mix data. It took three inputs, namely English language features, Hindi language features (transliterated Hindi converted to Hindi language) and typographic features, which are learned separately using sub-networks (Capsule network for English, Bi-LSTM for Hindi and MLP for typographic). The CapsNet sub-network for English was trained using GloVe pre-trained embeddings and the results were compared with the existing state-of-the-art *Toxic Comment Classification Challenge* dataset[4] from a Kaggle competition. The dataset contained 159571 Wikipedia manually labelled comments categorized as: *toxic; severe toxic; obscene; threat; insult* and *identity hate*. All these categories accounted for cyberbullying whereas any comment with value = 0 in all fields indicated non-cyberbullying i.e., non-toxic comments. As per www.kaggle.com, the first place solution reported a performance of 0.9885 using a Bi-GRU with the pseudo-labeling technique. The performance of the best single model of the competition was around 0.9869 and a single layer RNN-Capsule Network with GRU cell performed at 0.9857. In 2019, Srivastava et al. [42] used capsule network with focal loss and achieved an ROC-AUC of 0.9846 on the Kaggle toxic comment dataset. The performance of the proposed CapsNet was comparable at 0.9841. In this work as we used ELMo, the results further improved and achieved a ROC-AUC of 0.9924.

Three machine learning classifiers, namely, K- nearest neighbor (K-NN) and Naïve Bayesian (NB) and support vector machine (SVM) were compared with deep neural ConvNet image classifier. The Bag-of-Visual words (BoVW) approach [43] was used to extract the features and train the three machine learning classifiers. It was observed that the ConvNet outperformed the other classifiers. Comparative analysis of the image classification algorithms is given in table 3.

**Table 3.** Comparative Analysis of different classifiers used for Image Modality

| Classifier | Precision | Recall | Accuracy |
|---|---|---|---|
| K-NN | 71.3 | 71.8 | 65.8 |

[4] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

| | | | |
|---|---|---|---|
| NB | 67 | 69.2 | 64.4 |
| SVM | 76.86 | 79.17 | 73.2 |
| **ConvNet** | **98.6** | **95.08** | **97.05** |

We also implemented reversed the hybrid by using a ConvNet for the textual processing module and the CapsNet for the visual processing module and it was observed that the original set-up achieved superlative results. The ROC-AUC curve for this variation is shown in fig.9.

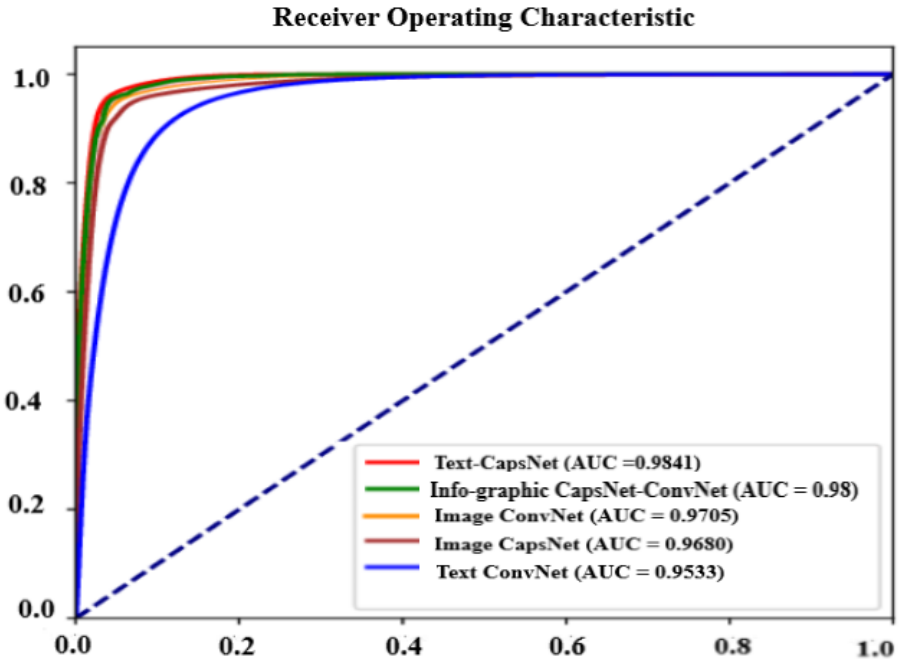



**Fig.9.** Performance results of model reversal

## 5. Conclusion

Social media and the internet have unlocked innovative modes to communication, empowerment and oppression. Meaningful engagement has transformed into a detrimental avenue where individuals are often vulnerable targets to online ridiculing. Cyberbullying is a rising predicament linked to the use of social media due to the increased consequential mental health risks associated. Predictive models to detect this cyberbullying in online content is imperative and this research proffered a prototype model for the same. The uniqueness of the proposed hybrid deep learning model, CapsNet-ConvNet is that it deals with different modalities of content, namely, textual, visual (image) and info-graphic (text with image). The results have been evaluated and compared with various baselines and it is observed the proposed model gives superlative performance accuracy.

The limitations of the model arise from the characteristics of real-time social data which are inherently ‘high-dimensional’, ‘imbalanced or skewed’, ‘heterogeneous’, and ‘cross-lingual’. The growing use of micro-text (wordplay, creative spellings, slangs) and emblematic markers (punctuations and emoticons) further increase the complexity of real-time cyberbullying detection. Other content modalities such as audio, GIFs, videos are open to research too. Also recently, the transformer-based methods, namely, BERT, ELECTRA, XLNet, RoBERTa and DistilBERT have been used within the NLP landscape, outperforming the state-of-the-art on several tasks and therefore their use for cyberbullying detection can be probed.

## References


1. Campbell, M. A. (2005). Cyber bullying: An old problem in a new guise? *Journal of Psychologists and Counsellors in Schools*, *15*(1), 68-76.
2. Kumar, A., & Sachdeva, N. (2019). Cyberbullying detection on social multimedia using soft computing techniques: a meta-analysis. *Multimedia Tools and Applications*, *78*(17), 23973-24010.
3. Kumar, A., & Sachdeva, N. (2020). Multi-input Integrative Learning using Deep Neural Networks and Transfer Learning for Cyberbullying Detection in Real-time Code-Mix Data. *Multimedia Systems,* https://doi.org/10.1007/s00530-020-00672-7
4. Sangwan, S.R., Bhatia, M.P.S. (2020). D-BullyRumbler: a safety rumble strip to resolve online denigration bullying using a hybrid filter-wrapper approach. *Multimedia Systems,* https://doi.org/10.1007/s00530-020-00661-w.
5. Kumar, A. Srinivasan, K., Cheng, W.H & Zomaya, A.Y. (2020). Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing & Management,* Elsevier, vol. 57, no.1, 102141.
6. Kumar, A. (2020). Using cognition to resolve duplicacy issues in socially connected healthcare for smart cities. *Computer Communications* 152 (2020): 272-281. https://doi.org/10.1016/j.comcom.2020.01.041
7. Young T., Hazarika D., Poria S. & Cambria, E. (2017). Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence magazine,* vol. 13, no. 3, pp. 55-75.
8. Dadvar, M., & Eckert, K. (2018). Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study. *arXiv preprint arXiv:1812.08046*.
9. Zhao, W., Ye, J., Yang, M., Lei, Z., Zhang, S. and Zhao, Z., 2018. Investigating capsule networks with dynamic routing for text classification. *arXiv preprint arXiv:1804.00538*.
10. Kim, J., Jang, S., Park, E. and Choi, S., 2019. Text classification using capsules. *Neurocomputing*
11. Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, *29*(9), 2352-2449.
12. Dinakar, K., Jones, B., Havasi, C., Lieberman, H. and Picard, R., 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(3), p.18.
13. Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. *Archives of suicide research*, *14*(3), 206-221.
14. Kokkinos, C. M., Antoniadou, N., & Markos, A. (2014). Cyber-bullying: An investigation of the psychological profile of university student participants. *Journal of Applied Developmental Psychology*, *35*(3), 204-214.
15. Dadvar, M., Jong, F. D., Ordelman, R., & Trieschnigg, D. (2012). Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.
16. Dadvar, M., Trieschnigg, D., Ordelman, R. and de Jong, F. (2013). Improving cyberbullying detection with user context. In *European Conference on Information Retrieval* (pp. 693-696). Springer, Berlin, Heidelberg.
17. Nahar, V., Unankard, S., Li, X., & Pang, C. (2012, April). Sentiment analysis for effective detection of cyber bullying. In *Asia-Pacific Web Conference* (pp. 767-774). Springer, Berlin, Heidelberg.
18. Nahar, V., Al-Maskari, S., Li, X., & Pang, C. (2014, July). Semi-supervised learning for cyberbullying detection in social networks. In *Australasian Database Conference* (pp. 160-171). Springer, Cham.
19. Reynolds, K., Kontostathis, A., & Edwards, L. (2011, December). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops* (Vol. 2, pp. 241-244). IEEE.
20. Michal, P., Pawel, D., Tatsuaki, M., Fumito, M., Rafal, R., Kenji, A., & Yoshio, M. (2010). In the service of online order: Tackling cyber-bullying with machine learning and affect analysis. *International Journal of Computational Linguistics Research*, *1*(3), 135-154.
21. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A. and Edwards, L. (2009). Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, *2*, pp.1-7.
22. Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W. and Hoste, V., (2015). Automatic detection and prevention of cyberbullying. In *International Conference on Human and Social Analytics (HUSO 2015)* (pp. 13-18). IARIA.
23. Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W. and Hoste, V., (2015). Detection and fine-grained classification of cyberbullying events. In *Proceedings of the international conference recent advances in natural language processing* (pp. 672-680).

24. Al-garadi, M.A., Varathan, K.D. and Ravana, S.D. (2016). Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, *63*, pp.433-443.
25. Xu, J.M., Jun, K.S., Zhu, X. and Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 656-666). Association for Computational Linguistics.
26. Zhao, R., Zhou, A. and Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking* (p. 43). ACM.
27. Raisi, E. and Huang, B., 2017, July. Cyberbullying detection with weakly supervised machine learning. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 409-416). ACM.
28. Agrawal, S., & Awekar, A. (2018, March). Deep learning for detecting cyberbullying across multiple social media platforms. In *European Conference on Information Retrieval* (pp. 141-153). Springer, Cham.
29. Huang, Q., Singh, V. K., & Atrey, P. K. (2014, November). Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia* (pp. 3-6).
30. Reynolds, K., Kontostathis, A., & Edwards, L. (2011, December). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops* (Vol. 2, pp. 241-244). IEEE.
31. Xu, Z., & Zhu, S. (2010, July). Filtering offensive language in online communities using grammatical relations. In *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference* (pp. 1-10).
32. Cheng, L., Guo, R., Silva, Y., Hall, D., & Liu, H. (2019, May). Hierarchical attention networks for cyberbullying detection on the instagram social network. In *Proceedings of the 2019 SIAM International Conference on Data Mining* (pp. 235-243). Society for Industrial and Applied Mathematics.
33. Al-Hashedi, M., Soon, L. K., & Goh, H. N. (2019, November). Cyberbullying Detection Using Deep Learning and Word Embeddings: An Empirical Study. In *Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems* (pp. 17-21).
34. Founta, A. M., Chatzakou, D., Kourtellis, N., Blackburn, J., Vakali, A., & Leontiadis, I. (2019, June). A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science* (pp. 105-114).
35. Mahlangu, T., & Tu, C. (2019, November). Deep Learning Cyberbullying Detection Using Stacked Embbedings Approach. In *2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)* (pp. 45-49). IEEE.
36. Kansara, K.B. and Shekokar, N.M., 2015. A framework for cyberbullying detection in social network. *International Journal of Current Engineering and Technology*, *5*(1), pp.494-498.
37. Singh, V. K., Ghosh, S., & Jose, C. (2017, May). Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2090-2099).
38. Yang, F., Peng, X., Ghosh, G., Shilon, R., Ma, H., Moore, E., & Predovic, G. (2019, August). Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 11-18).
39. Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. In *Advances in neural information processing systems* (pp. 3856-3866).
40. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. *arXiv preprint* arXiv:1802.05365. 2018 Feb 15.
41. Caldeira, M., Martins, P., Costa, R. L. C., & Furtado, P. (2020). Image Classification Benchmark (ICB). *Expert Systems with Applications*, *142*, 112998.
42. Srivastava, S., Khurana, P., & Tewari, V. (2018, August). Identifying aggression and toxicity in comments using capsule network. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)* (pp. 98-105).
43. Alkhawlani, M., Elmogy, M., & Elbakry, H. (2015). Content-based image retrieval using local features descriptors and bag-of-visual words. *Int J Adv Comput Sci Appl*, *6*(9), 212-219.