Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09631-w

Identification of cyberbullying: A deep learning based multimodal approach

Sayanta Paul[1] · **Sriparna Saha**[1] · **Mohammed Hasanuzzaman**[2]

Received: 4 April 2020 / Revised: 3 July 2020 / Accepted: 13 August 2020 /
Published online: 10 September 2020

Abstract
Cyberbullying can be delineated as a purposive and recurrent act, which is aggressive in nature, done via different social media platforms such as Facebook, Twitter, Instagram and others. While existing approaches for detecting cyberbullying concentrate on unimodal approaches, e.g., text or visual based methods, we proposed a deep learning based early identification framework which is a multimodal (textual and visual) approach (inspired by the informal nature of social media data) and performed a broad analysis on vine dataset. Early identification framework predicts a post or a media session as bully or non-bully as early as possible as we have processed information for each of the modalities (both independently and fusion-based) chronologically. Our multimodal feature-fusion based experimental analysis achieved 0.75 F-measure using ResidualBiLSTM-RCNN architecture, which clearly reflects the effectiveness of our proposed framework. All the codes of this study are made publicly available on paper's companion repository.

Keywords Cyberbullying · Multimodal information fusion · Deep learning

1 Introduction

With the exponential growth of digitization, e.g., use of different forms of social media platforms, where people can share and express their insights and feelings freely and publicly with others, which may come in different modalities, e.g., text, audios, videos, gestures and

https://github.com/sayantapaul/Multimodal-Cyberbullying-Identification

✉ Sayanta Paul
1811cs16@iitp.ac.in

Sriparna Saha
sriparna@iitp.ac.in

Mohammed Hasanuzzaman
Mohammed.Hasanuzzaman@cit.ie

[1] Indian Institute of Technology Patna, Bihta, India

[2] Cork Institute of Technology Cork, Cork, Ireland

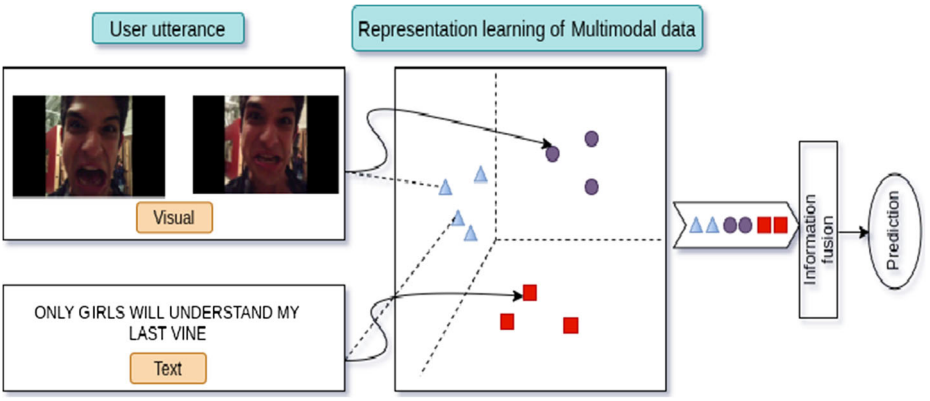


Fig. 1 Utterances from both textual and visual content of a post are fused to capture meaningful multimodal information

so on - misuse of this medium to promote offensive and hateful language, which may include harassment of regular users in the form of stalking, affects business of online companies, and may even have severe real-life consequences. Due to this unrestricted nature of viability of internet, all these activities can be appeared as an assortment of tech-empowered exercises, e.g., photo and video sharing, blogging, business networks, comments & reviews and many others which introduce continuous harassment and stalking which are commonly referred as cyberbullying [28]. Broadly cyberbullying can come up of different forms such as racism (e.g., facial features, skin colour), sexism (e.g., male, female), physical appearance (e.g., ugly, fat), intelligence (e.g., ass, stupid) and so on. Sometimes this act of cyberbullying is anonymous[1], i.e., quite hard to trace, which has intense and devastating effects. Therefore detecting cyberbullying at its initial stage is a crucial step to prevent this act and also to avoid any fatal incidents caused by it. In recent years, researchers have focused on developing different machine learning and deep learning based methods for solving the cyberbullying detection problem, most likely, through machine learning.

The report from National Crime Prevention Council[2] reveals that in USA, around 43% youth have been bullied. Therefore early identification of cyberbullying is necessary and essential as well in order to avoid the fatal consequences caused by cyberbullying. In this paper, we have proposed an automatic early identification framework of cyberbullying that effectively detects the signs of such brutal activities considering a multimodal approach. Existence of multiple modalities, e.g., textual, visual, audio and other channels is referred to as multimodality [22]. Due to the presence of heterogeneous sources of information, dependency upon a single modality lacks to infer enriched information, which often limits the unimodal approaches. For example, only text based information inference makes use of characters, words, phrases, which are inadequate for extracting meaningful knowledge from the content. But an ideal union of textual and video data may help in discovering events like cyberbullying in a more generic and enriched way, e.g., videos provide us visual information, which in association with text modality, can better identify cyberbullying. Figure 1 shows the shared multimodal representation of our proposed framework.

[1] https://cyberbullying.org/

[2] https://www.ncpc.org/wp-content/uploads/2017/11/

In this work, we make use of a dataset of *Vine*, a manually annotated dataset that comprises of video, audio and text modalities. Here, in all our experiments, we have considered text and video modalities and ignored audio as most of the videos are without proper sound or having completely irrelevant audio with the respective video. Different multimodal deep learning based frameworks have been proposed to effectively identify cyberbullying. As the part of information fusion, we have empirically explored various information fusion techniques and zeroed to two state-of-the-art fusion techniques, namely, feature level fusion or early fusion and decision level fusion or late fusion. In feature level or early fusion technique, features extracted from different modalities are fused as general feature vectors and the combined features are considered for further analysis. On the other hand, in decision level or late fusion method, features of each modality are examined and classified independently and the results are then fused as decision vector in order to get the final decision. After all the experiments, i.e., text-based early detection, visual-based early detection, feature fusion and decision fusion as well, it can be seen in the subsequent sections that feature-based fusion technique achieves the overall best performance among all other frameworks.

The main contributions of this paper are as follows:

- ***Multimodal identification framework:*** As discussed earlier, the consequences of cyberbullying are adverse, even it may be a cause of suicide, in many cases. Detecting it in it's primitive stage may pay off its certain outcomes. Therefore, we treated the texts in their chronological order, i.e., in the order each comment is posted against an video. In our dataset, each of the posts consists of 15 comments. In order to early identify, we first have considered 5 comments for each of the posts, then 10, and finally 15, which is the highest number of comments. Similarly, we have processed the frames, extracted from each of the videos, chronologically.
- ***Post-level embedding:*** In order to get enriched semantic information to capture the underlying contexts of cyberbullying, we have created our own text embedding with the help of Transformer Encoder of Universal Sentence Encoder(USE). Principle Component Analysis algorithm has also been used to reduce dimensions.
- ***Evaluations:*** We empirically explored and validated all aforementioned methodologies and frameworks over *Vine* dataset. No previous research activities in the literature exist for identifying cyberbullying over Vine using deep learning based multimodal approach. The details of the experiments and the obtained results are reported in the subsequent sections.

The organization of this paper is as follows: A brief survey of previous works upon the cyberbullying detection task has been exhibited in the following Section 2. Section 3 describes the details of the dataset used throughout this work. The proposed frameworks have been explained in Section 4. Experimental evaluation along with the obtained results have been presented in Section 5. A discussion on our approach is elaborated in Section 6. Subsequently, the conclusion of this work and it's future scopes are elucidated in Section 7.

2 Related works

Identifying cyberbullying over social media has been an increasingly trending issue over the past few years. A great number of research activities have been published, trying to address this problem in social networks, and in various forms. In literature, several works can be found towards providing an effective solution for identifying cyberbullying via text-based,

image-based or video-based, sometimes incorporating multimodality as well, but to the best of our knowledge, our paper is the foundation to study cyberbullying from a video-based mobile social media, specifically Vine, using deep learning based frameworks which also include multimodality information.

2.1 Existing unimodal approaches

Natural Language processing (NLP) and other language technologies have shown their potential performance for solving problems like detection of hate-speech, fake news, harassment, cyberbullying and abusive language. It can also be found that, reinforcement learning have shown its potential effectiveness in developing predictive models, e.g., trading in financial markets such as stock and forex [16]. However, many significant architectures and other approaches have been introduced for the objective task, e.g., Karthik et al. [7] proposed detection of cyberbullying (specifically sexual, racism content) over YouTube dataset using Naive bayes and SVM classifiers; their system achieved 80.20%, 68.30% accuracy, respectively, on sexuality and racism contents. Reynolds et al. [27] presented a social media optimization technique over Formspring dataset which indicates a particular post is either of bully or non-bully. This system achieved 78.5% accuracy. In 2015, Djuric et al. [8] came up with paragraph-to-vector based distributed representations of comments over Yahoo social media that can easily detect hate-speech achieving a sound accuracy of 80%. Recently in 2017, Badjatiya et al. [2] introduced precise cyberbullying identification, i.e., Racism, Sexism and others using CNN and LSTM architecture over Twitter data which achieved 93% F1-score. Kumari et al. [14] designed a deep learning based system to classify aggressive post from a non-aggressive post containing symbolic images collected from Google search to query aggressive images. Their CNN with six convolution layers achieved a F1 score of 90%.

2.2 Existing multimodal approaches

Due to the presence of the ambiguity in textual content, it has often become very difficult task to infer any meaningful information from text-based medium. To extract any significant information properly, we therefore need to have some resource of input, from which we can gather more rich information. Along with the rapid development of social media, along with the text, visual modalities are also available. Many researchers, now-a-days, are moving towards multimodal approaches, where information from different modalities can be acquired to solve any real-world problems such as cyberbullying identification. Poria et al. [21] proposed a multimodal framework in which they have showed how information sourcing from different modalities can be deployed to get more fine grained decision. Another multimodal information fusion approach has been there in the literature which is one of the key motivations behind this work, proposed by Cambria et al. [4] where they have fused multi-source information for the continuous interpretation of semantics and sentics. Recently, Kumari et al. [13] proposed a CNN-based unified representation of textual and visual content together to eliminate the need for separate learning modules for image and text for detecting cyberbullying. Authors have collected data from several widely used online social networks, e.g., Facebook, Instagram and Twitter. They used three layers of text and three layers of a colour image to represent the input that computed a recall of 74% of the bullying class with one layer of Convolutional Neural Network. In 2020, Kumari et al. [12] developed a genetic algorithm based mutlimodal cyberbullying detection framework where authors used a pre-trained VGG-16 network and convolutional neural network to extract the

features from images and text, respectively, over the same dataset as introduced in [13] with an improved F1 score of 78%. Some multimodal approaches for depression detection are developed in [24].

3 Data description

The dataset used in this paper, *Vine*, is provided by Rahat et al. [25, 26]. It is a short-term video hosting network on which users shared six-second-long, looping video clips. It was founded in June 2012; American micro-blogging website Twitter acquired it in October 2012, before its launch on January 24, 2013. The reason that we have considered this social media to accomplish our task is, there is no such work in the literature that has explored cyberbullying issue over *Vine* using deep learning frameworks, a multimodal approach.

The data have been collected using snowball sampling method [10] as follows: select a random user u_s as a seed and collect all the users that u_s is following. Using the aforementioned method, initially, 59,560 users' information have been collected: user_id, user_name, full name, location, profile description, no. of videos posted with post_id, no. of followers, id of the followers. In total, 625K media sessions were collected. But the main objective was to detect cyberbullying in media sessions for which sufficient no. of comments are required, therefore those media sessions that have at least 15 comments are selected. Rahat et al. [25, 26] have also conducted a profanity test [11], where a media session is said to be profane if it's at least one word in comment is profane. Depending upon % of comments with profanity, 969 media sessions have been sampled as the final dataset, each of which belongs to a distinct user. The % of profanity for the media session has been shown in Fig. 2.

In the final dataset, among 969 media sessions, there are certain number of media sessions missing. This may happen as the authors [25, 26] have downloaded these videos sometime after they conducted the survey initially, during the survey they actually had the

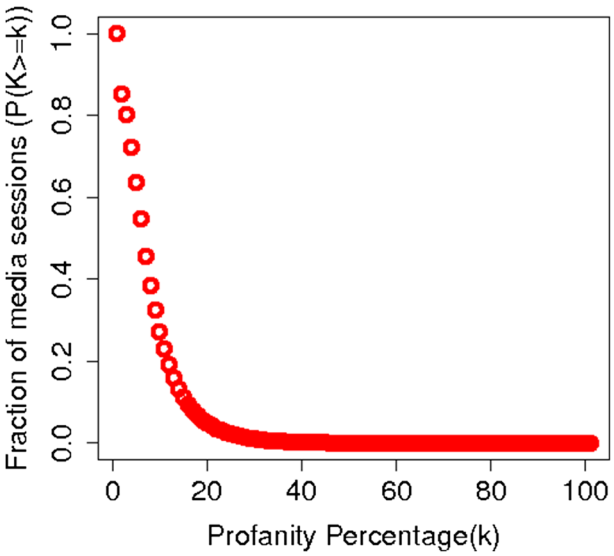


Fig. 2 Profanity percentage vs. fraction of media sessions

Table 1 Statistics of the data

Video	Session
Total number of videos	969
Number of bullying videos	304
Number of non-bullying videos	665
Number of empty/missing video files	130
Number of bullying videos within missing files	35
Number of non-bullying videos within missing files	95

urls to the videos. After the survey, they decided that it would be a good idea to also download the videos as well. It may be the case that within that time frame, some of these videos got deleted by the respective users. After performing a thorough exploratory data analysis, data statistics are presented in Table 1.

We have also explored the distribution of *likes* and *comments* of the bullying and non-bullying media sessions, as indicated in the following Fig. 3a & b, respectively.

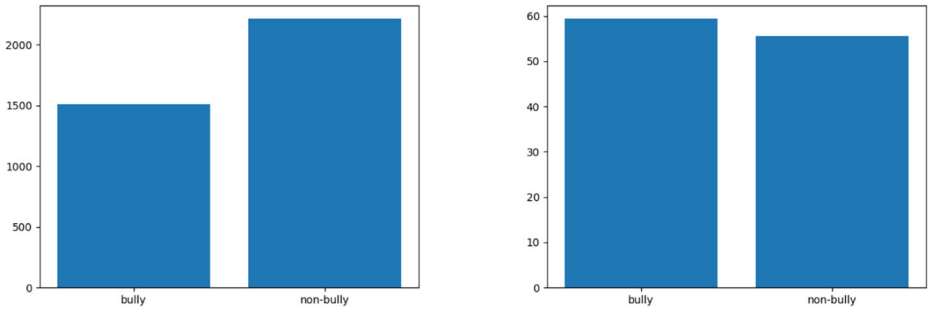
4 Proposed methodologies

In this section, we have detailed about the sentence embedding generated, different architectures that have been used throughout the experiments and the information fusion techniques [22], i.e., feature fusion and decision fusion. We have also presented how the input flows through the network to obtain desired output.

4.1 Post embedding

Building any deep learning model in natural language processing, generation of good text embedding plays an important role. To generate good text embedding with rich semantic information, we have used Universal Sentence Encoder.[3] The Universal Sentence Encoder encodes text into high dimensional vectors that can be used for text classification, semantic similarity determination, clustering and other natural language tasks. Instead of training embedding from scratch on the comments, which has the risk of over-fitting, we have instead used pre-trained embedding. Universal Sentence Encoder offers two models, Transformer Encoder and Deep Averaging Network (DAN), we opted to go with the Transformer Encoder [5], as this architecture constructs sentence embedding using the encoding subgraph of transformer architecture, which computes a context-aware representation of words in a given sentence that considers both the order and identity of all other words in the sentence. It takes input a lower cased PTB-tokenized[4] string and outputs a 512 dimensional vector as sentence embedding. As mentioned in the previous section, our final dataset consists of 269 bullying media sessions and 570 non-bullying media sessions, i.e., videos along with their comments, with which our post embedding has been framed. The generated embedding can be seen in Fig. 4. After inducing the embedding, we have also explored

[3] https://tfhub.dev/google/universal-sentence-encoder/1
[4] https://nlp.stanford.edu/software/tokenizer.shtml



(a) *likes* (b) *comments*

Fig. 3 Average no. of *likes & comments* over all media sessions

the quality of the generated embedding, i.e., we plot the class-specific embeddings, e.g., for non-bully and bully classes, respectively. The sparseness present in the obtained Fig. 4 clearly shows us non-bully textual content is more than that of the bully content as the intensity of bully content is comparatively less.

The embedding generated is a vector of 512 dimension. Figure 4 plots the word-representations in graphical form after reduced from 512D to 2D using PCA decomposition algorithm [31]. The effectiveness of using Universal Sentence Encoder for creating our own text embedding is shown in Fig. 5.

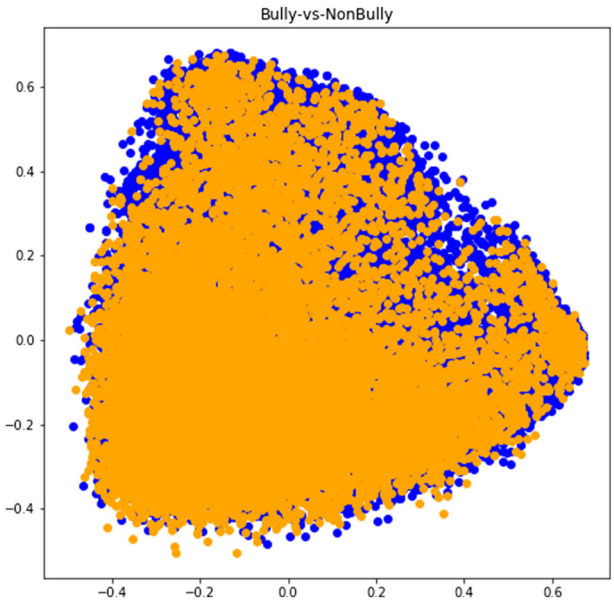


Fig. 4 Post embedding generated using Universal Sentence Encoder(USE) which clearly shows the density of non-bully class over bully class

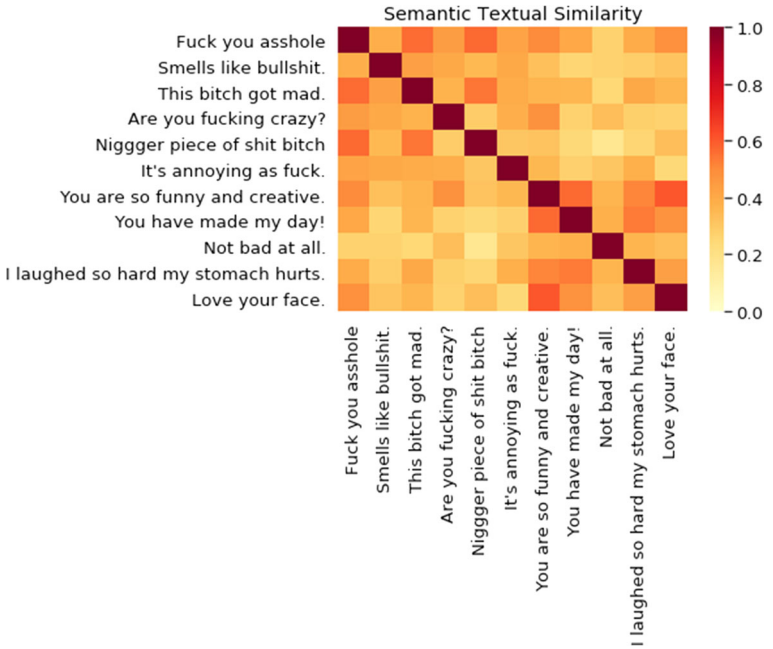


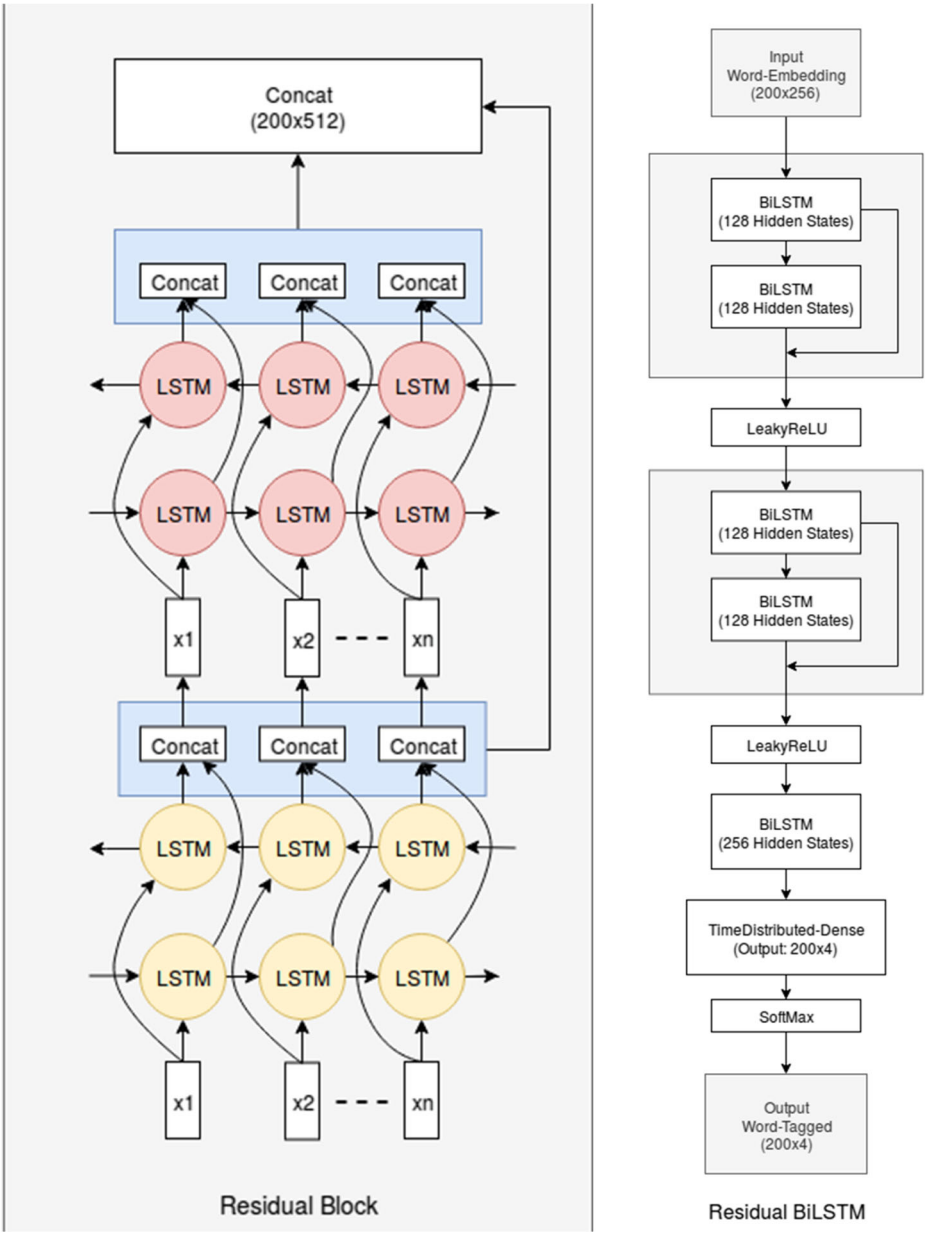
Fig. 5 Captured semantic similarity on the textual content of our dataset, selected from both the classes via Universal Sentence Encoder

4.2 Early detection of bullying based on textual modality

We processed the texts in their chronological order, i.e., in the order each comment is posted against an video. In our dataset, each of the posts consists of 15 comments. In order to early identify, we first have considered 5 comments for each of the posts, then 10, and finally 15, which is the highest number of comments. We have proposed Residual-BiLSTM (Res-BiLSTM) architecture for text classification. Our proposed architecture consists for two residual blocks [23] which can be increased or decreased as required. Figure 6a shows the internal architecture of the residual-BiLSTM block. The weights of the model are initialized using Xavier's initialization technique [9], as this weight initialization technique tries to make the variance of the output to be equal to the variance of its input. The reason we opted to go with residual based architecture is that, as exhibited by ResNet [30] in image classification task, residual blocks have the ability to incorporate both high-level features, achieved from various layers, as well as low-level features. Using all these features can significantly improve the network's performance and also the network will not get over-fitted during training as excessive residual blocks can simply act as identity blocks, thus avoiding over-fitting. Based on these reasons, we have opted to use the residual-based architecture. The residual block, as shown in Fig. 6a, works as follows:

$$\hat{\boldsymbol{h}}_t^{(l)} = f_h^l(h_t^{(l-1)}, h_{(t-1)}^l) + x_{(l-n)} \tag{1}$$

The above $\hat{\boldsymbol{h}}$ function is being learned for the layers with residual connections as it requires the input to be in similar dimension as the output of h_t. Here, $\hat{\boldsymbol{h}}$ for layer l is updated with



(a) Residual-BiLSTM Block (b) Proposed Model

Fig. 6 Internal architecture of Residual block & ResBiLSTM

residual value $x_{(l-n)}$ and x_i represents the input to layer $i + 1$. Residual connections are to be added after each n layers.

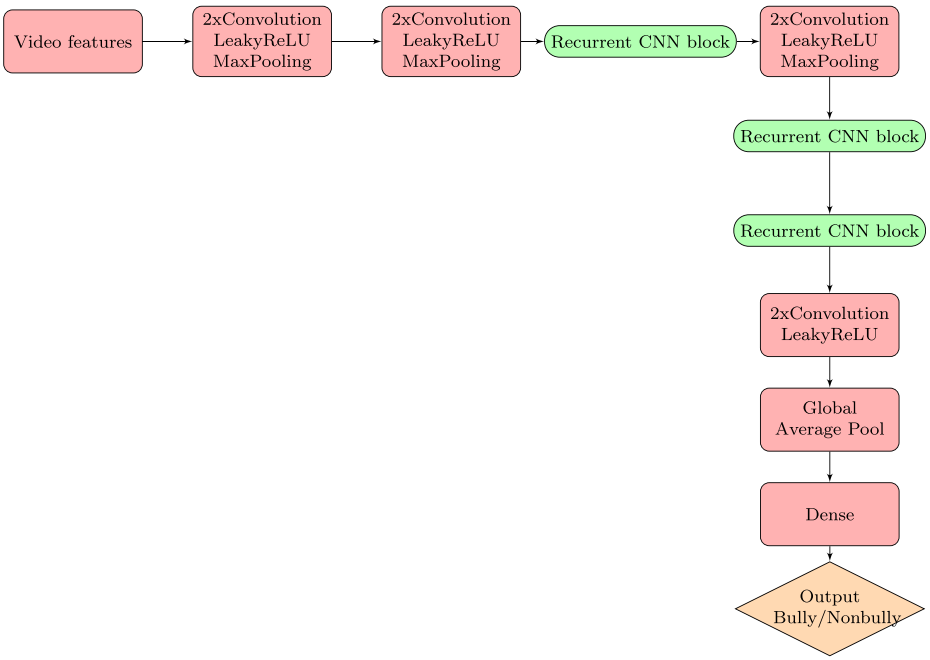


Fig. 7 Proposed Aarchitecture

We propose a framework, as indicated in Fig. 6, for early detection [18] using a common observation, i.e., if a video displays a bullying action, a few initial comments will also be bully-related and may also result in further "replies/comments" which will be either encouraging or discouraging bullying and this pattern will gradually fade as the number of "comments" will increase. Therefore, we classify the post-based solely on few initial comments and observe the results.

4.3 Early detection of bullying based on visual modality

As in the case of text, we have also processed the image-frames extracted from the videos, in the order they have been generated. Here, Recurrent-CNN (RCNN) [20] has been used for visual classification. The architecture of the image classifier illustrates how we process the frames sequentially. Figure 7 shows the exact architecture of the model. Recurrent-CNN (RCNN) takes input features of 30-frames of the video stacked on top of another (Fig. 8). 30-frames of each video are extracted by selecting a frame at a regular interval. We use InceptionResNetV2[5] model due to its exceptional performance for the task of image-classification to extract the features of each frame.

4.4 Information fusion

Let us consider there are n different data modalities $\{f_i \leq i \leq n\}$ (text-visual modalities in our case) and output decisions about cyberbullying incident C at time instant t in terms of

[5] https://ai.googleblog.com/2016/08/improving-inception-and-image.html

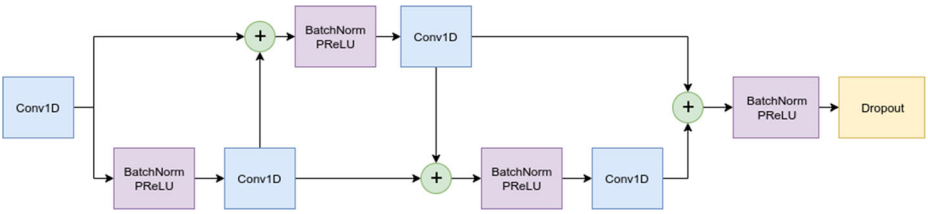


Fig. 8 Recurrent-CNN block

n probability values are $p_1(t)$, $p_2(t)$, .., $p_n(t)$. Here, $p_i(t) = p(C|f_{i,t})$, which represents the probability that cyberbullying event C has occurred based on modality f_i at the time t. To learn the intrinsic correlation between different modalities of features, we have used recurrent attention network [19]. This mechanism helps to capture modality independent features including both fine-grained local and contextual information by using the attention mechanism. The multimodal information has been fused using two techniques as follows.

4.4.1 Feature-level fusion

In this information fusion strategy, features from each independent modality are extracted, which are then combined (weighted average) together in order to get a single set of combined features [22]. We have extracted text features using Residual-BiLSTM [29] and visual features are extracted using Recurrent Convolution Neural Network [20]. Both set of features are then fused using attention fusion. In order to get the final decision, we have deployed different state of the art deep learning based frameworks, e.g., BiLSTM [1], Recurrent-CNN [20], integration framework of previous two architectures. But our proposed framework performs the best amongst all the other frameworks, as reported in the subsequent section.

Let $T_i = \{t_{i,1}, t_{i,2}, \ldots, t_{i,n}\}$ and $V_i = \{v_{i,1}, v_{i,2}, \ldots, v_{i,n}\}$ be the feature vector of comments (text modality) and images (visual modality), respectively, for a particular post. The fused vectors $F_i = \{f_{i,1}, f_{i,2}, ..., f_{i,n}\}$ are obtained by combining these two feature vectors.

Figure 10 shows a generalized architecture for feature-level fusion. Embedding of comments and features extracted from video-frames are input parallel to the classifier. We experimented with various architectures for text-feature extractor and video-feature extractor. BiLSTM with 1024 hidden-size and Residual-BiLSTM both have been used as text-feature extractors, however, Residual-BiLSTM leading slightly over BiLSTM. As video-feature extractor, we have used two recurrent-CNNs. These two features, text and video, are then fused together with attention fusion (Fig. 9). A final classifier is built on top of this attention fused combined features.

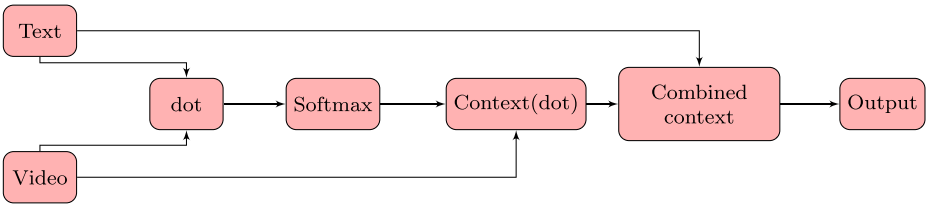


Fig. 9 Framework for attention fusion

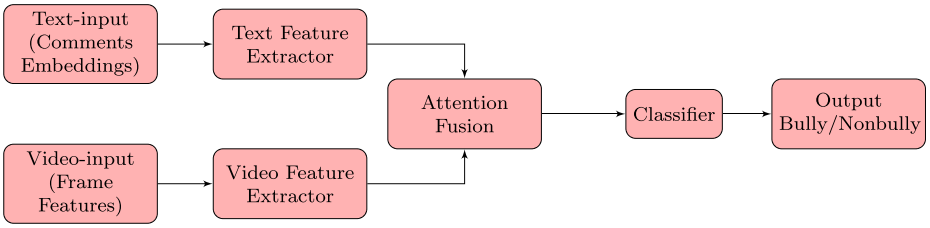


Fig. 10 Generalized architecture for Feature-level fusion

4.4.2 Decision-level fusion

Figure 11 shows a generalized architecture for decision-level fusion [22]. This follows the same process as feature-level fusion except that the text-feature extractor and video-feature extractors are replaced by binary-classifiers (bully v/s non-bully). In this strategy, the individual classifier provides the local decisions, say $x_1, ..., x_n$, that are obtained based on individual features $f_1, ..., f_n$. The local decisions are then combined as a decision vector $\bar{x} = (x_1, ..., x_n)$, where $0 \leq x_i \leq 1, 1 \leq i \leq n$. The independent decisions of these two classifiers are then fused together with attention fusion technique (Please refer to Fig. 9).

5 Experimental evaluation

We have reported the performance of the proposed deep learning frameworks on the validation set using only text modality, then only visual modality and eventually move forward to different fusion methods, e.g., feature-level fusion and decision-level fusion of both text and visual modalities. As in the training data, instances of bully class are very less compared to those of the non-bully data, we have used precision, recall and F-measure [3] as the evaluation metrics to measure the performance of the deep learning frameworks. Here, we have also shown the hyper-parameter settings for training our deep learning frameworks.

5.1 Text-based early detection

Here, we have processed the texts, i.e., the comments in their chronological order and presented the results in the following table. Based on the obtained experimental results, it can be clearly observed that with the increase in the number of comments, inference of cyberbullying is effective (Table 2). Therefore, we hypothesize that we can use initial comments to decide whether the post will lead to bullying interaction in near-future and also if the post

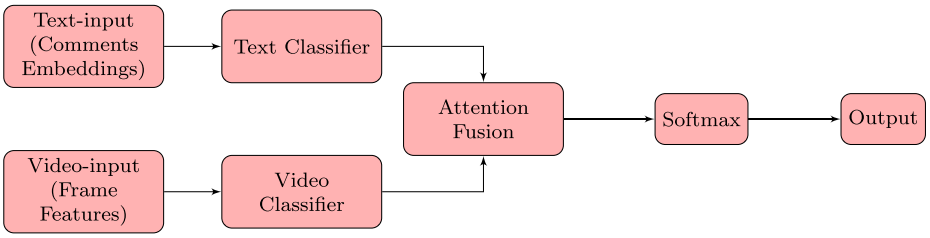


Fig. 11 Generalized architecture for Decision-level fusion

itself is bullying or not. From the obtained results in the following table we can claim that as the number of comments increases, existence of bullying instances increases, which is the fundamental reason of enhanced F-measure score for the bully class. On the other hand, by the increment of bullying instances, non-bully instances are relatively less exist in the comments, which caused the decrease in metric scores in case of Non-bully class.

5.2 Visual-based early detection

In this scenario, we have processed the image frames generated from the videos using Recurrent-CNN architecture [20], which is the state-of-the-art architecture for image processing and presented the results in Table 3.

5.3 Information-fusion based early detection

In this section, we have shown the experimental results of different information fusion techniques that we performed as follows:

5.3.1 Feature-level fusion

Here, we have tried different deep learning classifiers to achieve feature-level information fusion but the underlying architecture of this fusion is the same as shown earlier. For any classifier that produces a real-valued output, we derive the relationship between the best achievable F1 score and the decision-making threshold that achieves this optimum [15] (Table 3).

As our fundamental motivation is to identify cyberbullying as early as possible, here also we have processed both the modality features chronologically, precisely, text modality input has been given as chunk of 5 comments into the network and for visual modality, the input has been provided in the order of generating frames from the videos and their corresponding features. Different state-of-the-art architectures have been explored for this purpose, e.g., LSTMs, BiLSTMs [29], BiLSTM-RecurrentCNN but our proposed framework, i.e., integration network of ResBiLSTM-RCNN performs relatively well than the others, as reported in Table 4.

Table 2 Comments in chronological order; Classifier: **ResBiLSTM**

Comments			
5 comments	Precision	Recall	F-measure‡
Non-bully	0.76	0.81	0.79
Bully	0.38	0.31	0.34
10 comments	Precision	Recall	F-measure‡
Non-bully	0.85	0.87	0.86
Bully	0.69	0.66	0.61
15 comments	Precision	Recall	F-measure‡
Non-bully	0.77	0.74	0.76
Bully	0.71	0.69	0.71

‡ indicates all the results obtained above are statistically significant

Table 3 Results obtained by Recurrent-CNN classifier with no. of frames: 30

Image frames Class	Precision	Recall	F-measure‡
Non-bully	0.65	0.88	0.76
Bully	0.42	0.15	0.23

Figure 12 shows the resulting graph of F1-score calculated at different thresholds ranging from 0.0 to 1.0 for BiLSTM-RecurrentCNN. Results of different frameworks have been reported in Table 4.

5.3.2 Decision-level fusion

The decisions generated by deployed classifiers are fused here in order to get a single decision. To be precise, the decision given by text classifier and the decision produced by image classifier are fused using attention layer. Both the decisions coming from each of the modalities, are integrated using Recurrent Convolution Neural Network [20]. The obtained results have been reported in Table 5.

5.4 Training details

Out of 969 media sessions, we have trained all our models on 776 media sessions and validated on 194 sessions. All the models are trained on 0.0002 learning rate which is periodically reduced by a factor of 10 when validation loss further reduces. Batch size is kept as 16. Since the model is a binary-classifier, we use binary-cross-entropy as our loss function. Adam is used for optimization. The maximum number of epochs is 50. We follow this aforementioned set of hyper-parameters to train the architectures throughout the experiments. However, these sets of parameters varied through models a bit. The optimised sets of hyper-parameters used in all the conducted experiments have been shown in the Table 6.

Table 4 Results of different feature-level fusion methods

Performance of feature-level fusion Classifier	Class	Precision	Recall	F-measure‡
BiLSTM [1]	Non-bully	0.76	0.80	0.78
	Bully	0.50	0.45	0.47
Recurrent-CNN [20]	Non-bully	0.76	0.68	0.72
	Bully	0.57	0.67	0.62
BiLSTM-RecurrentCNN	Non-bully	0.78	0.89	0.83
	Bully	0.72	0.55	0.69
ResBiLSTM-RCNN⋆	Non-bully	0.87	0.87	0.87
	Bully	0.75	0.75	**0.75**

⋆ Our proposed framework

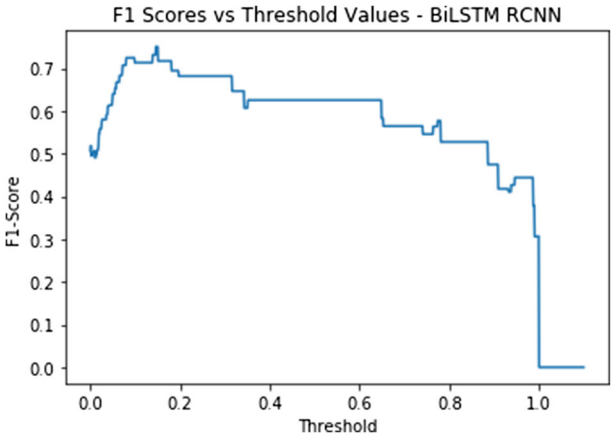


Fig. 12 F1-Score vs Thresholds

5.5 Key observations

All the reported results above are statistically significant as we have performed statistical t-test [6] at 5% significance level. Thus, to ensure that no ambiguity was introduced during training, the experiments were conducted for 5 times. In each time, the available data sample S is randomly divided into a training set R of a specified size (e.g., typically two thirds of the data) and a test set T. The deep learning frameworks are trained on R and tested on T. Let $P_X^{(i)}$ and $P_Y^{(i)}$ be the observed proportion of test examples misclassified by classifier X and Y, respectively, during ith time. If we assume that the 5 differences, $P^{(i)} = P_X^{(i)} - P_Y^{(i)}$ were drawn independently from a normal distribution, then we can apply Student's t-test, by computing the statistics:

$$t = \frac{\bar{p}.\sqrt{n}}{\sqrt{\frac{\Sigma_{i=1}^{n}(p^{(i)}-\bar{p})^2}{n-1}}} \tag{2}$$

where $\bar{p} = \frac{1}{n}\Sigma_{i=1}^{n} p^{(i)}$. Note that, $P_X^{(i)}$ and $P_Y^{(i)}$ are independent as our data is supposed to be normalised because S is randomly divided into a training set R and a test set T.

In the data description section, we have seen that the instances of bully class are very less in comparison to total number of non-bully instances. Therefore, unimodal analysis was not enough. Obtained results clearly show us that only text based or only visual based analysis was not effective enough, precisely, while in text-based early detection framework, initial F-measure was only 0.34, considering 5 comments. Again, for visual-based early detection, our score is only 0.23, which is very less compared to text-based analysis. These motivate us to move towards information fusion methodologies. In information fusion scenario, we have

Table 5 Results of decision-level fusion; Classifier used: **Recurrent-CNN**

Performance of decision-level fusion			
Class	Precision	Recall	F-measure‡
Non-bully	0.68	0.92	0.78
Bully	0.45	0.35	0.41

Table 6 Set of hyper-parameters

Optimised set of hyper-parameter obtained in our final models configuration	
Model	Hyperparameters
BiLSTM	Activation=LeakyReLU, Dropout probability=0.2, No. of hidden units=512, BatchNormalization=No
Recurrent-CNN	Activation=PReLU, Dropout probability=0.1, BatchNormalization=Yes
BiLSTM-RecurrentCNN	Activation=LeakyReLU, PReLU, No. of hidden units=512, BatchNormalization=Yes
ResBiLSTM-RCNN	Activation=LeakyReLU, Dropout probability=0.1, BatchNormalization=Yes

Table 7 F-measure for 5 runs of models on Vine dataset

Performances of different models			
BiLSTM	RCNN	BiLSTM-RCNN	ResBiLSTM-RCNN⋆
0.47	0.62	0.69	0.75
0.46	0.62	0.69	0.75
0.47	0.62	0.68	0.74
0.46	0.62	0.69	0.75
0.46	0.62	0.68	0.75

⋆ Our proposed framework

explored both feature based and decision based techniques but feature based information fusion performed very well. The reason behind the decision fusion not performing well is that it simply integrates the decisions generated by unimodal classifiers. Now, we have seen that unimodal classifiers did not produce any potential output, which is why feature fusion works very well rather than decision fusion.

5.5.1 Statistical significance test

As mentioned in the earlier subsection (refer Section 5.5), results of statistical t-test of 5 different runs of our proposed model on the dataset have been shown in Table 7.

The p-values [17] after conducting t-test on the results of our proposed framework with respect to other models are shown in Table 8. The threshold value of p is 0.05. Results show that for majority of the cases the p-value is less than threshold signifying that the performance improvements attained by our model are statistically significant.

5.5.2 Error analysis

We manually checked the misclassified instances to perform a thorough error analysis. The proposed models are unable to classify certain media sessions which contain comments that are likely to appear in other categories, e.g., "This bitch got mad". In the quoted example, the word 'bitch' is an abusive word and frequently used for different bullying comments, which leads it to be categorized as bully but originally the sentence belongs to non-bully. Table 9 contains some instances of misclassification and the reasons for the same. The possible reasons behind these misclassifications can be as follows:

1. Presence of abusive or swear words in the posts or comments. Our proposed framework could not be able to understand underlying sarcasm of post or comments. For example, "Oh God! You are such an asshole!" contains an abusive word *asshole*, apparently which is the reason behind occurrence of a misclassification error.

Table 8 p-values obtained after comparing our proposed framework with other state-of-the-art models for vine dataset

Model	p-value
BiLSTM	2.64e-13
RCNN	3.96e-12
BiLSTM-RCNN	4.77e-08

Table 9 Misclassified instances and reasons behind misclassification

Misclassified instances			
Instance	Predicted	Original	Possible reason
This bitch got mad.	bully	non-bully	The word 'bitch' indicates some hate-speech
Are you fucking crazy?	bully	non-bully	Due to presence of the words, e.g., 'fucking' and 'crazy'
Your damn post drove me feeling lost.	non-bully	bully	Absence of any abusive words.

2. Also, it can be observed that each of the corpora contains non-standard English words which our proposed model is unable to interpret properly to build vector representations. Consider the following example: "Tht fckin sht shuld die!(Standard English: That fucking shit should die!)" which our model is unable to understand which leads to a misclassification.

6 Discussion

Analysis of textual content enables a model to understand meaningful information underlying words, phrases and hidden inter-dependencies. On the other hand, analysis of visual content helps to identify emotions or sentiments from expressions in a more efficient way. This is regardless to mention that each modality endures its own shortcomings, as can be noted from experimental results, shown in Section 5. When we fused both the modalities of information together, it can help to create a more fine-grained prediction. Our fundamental challenge to integrate multimodal information is to develop fusion methodology that can combine effective information from different sources.

We, here, developed two different fusion strategies, i.e., feature-level fusion and decision-level fusion. Both the techniques were empirically experimented over our multimodal dataset. As our dataset suffers from class imbalance problem, feature-level fusion performs potentially significant over decision-level fusion (please refer to Table 4). However, our multimodal model suffers from certain demerits, as errors from different classifiers tend to be correlated and basically, our proposed model is feature dependent. Apart from this, further error analysis has been reported in Section 5.

7 Conclusions and future works

This work forges ahead the state of the art in cyberbullying identification far off unimodal approaches to propose a deep learning based multimodal method for fusing heterogeneous textual and visual features in order to identify cyberbullying in more fine-grained way. Both text and visual information have been considered and processed individually and these are further fused for accomplishing the task. This paper contributes the following towards effectively identifying cyberbullying at its primitive stage in order to avoid it's adverse effects: (a) developing a comment-level embedding for this task, (b) modelling a deep learning based early detection framework for cyberbullying detection, (c) empirically evaluating the performance of the frameworks over *Vine* dataset.

It can be seen that feature-level information fusion shows significantly potential performance improvements over the decision-level fusion method. The reason behind this would

be number of training data, especially bullying data is very less, which may not contribute much when independent decisions were being made. The experimental results show that the performance of proposed methodologies are reasonably good. As the further extension of this work, we are planning to expand the area of social networking sites by providing with our early detection framework as an automatic cyberbullying detection tool.

Acknowledgments We would like to express our gratitude to Rahat et al. [25, 26] for sharing their labeled multimodal dataset, *Vine*. Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research.

References

1. Agrawal S, Awekar A (2018) Deep learning for detecting cyberbullying across multiple social media platforms. In: European Conference on Information Retrieval, Springer, pp 141–153
2. Badjatiya P et al (2017) Deep learning for hate speech detection in tweets. In: Proceedings of the 26th international conference on world wide Web companion, pp 759–760
3. Basu T, Murthy CA (2012) A feature selection method for improved document classification. In: International conference on advanced data mining and applications, Springer, pp 296–305
4. Cambria E et al (2013) Sentic blending: Scalable multimodal fusion for the continuous inter- pretation of semantics and sentics. In: 2013 IEEE symposium on computational intelli- gence for human-like intelligence (CIHLI), IEEE, pp 108–117
5. Cer D et al (2018) Universal sentence encoder. In: arXiv:1803.11175
6. Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 10.7:1895–1923
7. Dinakar K, Reichart R, Lieberman H (2011) Modeling the detection of textual cyberbullying. In: Fifth international AAAI conference on weblogs and social media
8. Djuric N et al (2015) Hate speech detection with comment embeddings. In: Proceedings of the 24th international conference on world wide Web, pp 29–30
9. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feed- forward neural networks. In: Proceedings of the Thirteenth international conference on artificial intelligence and statistics, pp 249–256
10. Goodman LA (1961) Snowball sampling. In: The annals of mathematical statistics, pp 148–170
11. Hosseinmardi H et al (2015) Analyzing labeled cyberbullying incidents on the instagram social network. In: International conference on social informatics, Springer, pp 49–66
12. Kumari K, Singh JP (2020) Identification of cyberbullying on multi-modal social media posts using genetic algorithm. In: Transactions on Emerging Telecommunications Technologies, pp e3907
13. Kumari K et al (2019) Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach. In: Soft Computing, pp 1–12
14. Kumari K et al (2019) Aggressive social media post detection system containing symbolic images. In: Conference on e-Business, e-Services and e-Society, Springer, pp 415–424
15. Lipton ZC, Elkan C, Narayanaswamy B (2014) Thresholding classifiers to maximize f1 score. In:
16. Meng TL, Khushi M (2019) Reinforcement learning in financial markets. Data 4.3:110
17. Nuzzo R (2014) Scientific method: statistical errors. Nature News 506.7487:150
18. Paul Sayanta, Jandhyala SK, Basu T (2018) Early detection of signs of anorexia and depression over social media using effective machine learning frame- works . In: CLEF (Working notes)
19. Peng Y, Qi J, Yuan Y (2018) Modality-specific cross-modal similarity mea- surement with recurrent attention network. IEEE Trans Image Process 27.11:5585–5599
20. Pinheiro P, Collobert R (2014) Recurrent convolutional neural networks for scene labeling. In: International conference on machine learning, pp 82–90
21. Poria S et al (2015) Towards an intelligent framework for multimodal affective data analysis. Neural Netw 63:104–116
22. Poria S et al (2017) A review of affective computing: From unimodal analysis to multi- modal fusion. Inform Fusion 37:98–125

23. Prakash A et al (2016) Neural paraphrase generation with stacked residual lstm networks. In: arXiv:1610.03098
24. Qureshi SA et al (2019) Multitask representation learning for multimodal estimation of depression level. IEEE Intel Syst 34.5:45–52
25. Rafiq RI et al (2015) Careful what you share in six seconds: Detecting cyberbullying instances in Vine. In: 2015 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), IEEE, pp 617–622
26. Rafiq RI et al (2016) Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network. Soc Netw Anal Mining 6.1:88
27. Reynolds K, Kontostathis A, Edwards L (2011) Using machine learning to detect cyberbullying. In: 2011 10th international conference on machine learning and applica- tions and workshops, Vol 2, IEEE, pp 241–244
28. Smith PK et al (2008) Cyberbullying: Its nature and impact in secondary school pupils. J Child Psychol Psychiatry 49.4:376–385
29. Sun S, Xie Z (2017) Bilstm-based models for metaphor detection. In: National CCF conference on natural language processing and chinese computing, Springer, pp 431–442
30. Targ S, Almeida D, Lyman K (2016) Resnet in resnet: Generalizing residual architectures. In: arXiv:1603.08029
31. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemo- Metrics and Intelligent Laboratory Systems 2.1-3:37–52

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH ("Springer Nature").

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users ("Users"), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use ("Terms"). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com