Aggressive Social Media Post Detection System Containing Symbolic Images

Kirti Kumari[1], Jyoti Prakash Singh[1], Yogesh K. Dwivedi[2], and Nripendra P. Rana[2(✉)]

[1] National Institute of Technology Patna, Patna, India
{kirti.cse15,jps}@nitp.ac.in
[2] School of Management, Swansea University, Swansea, UK
ykdwivedi@gmail.com, nrananp@gmail.com

Abstract. Social media platforms are an inexpensive communication medium help to reach other users very quickly. The same benefit is also utilized by some mischievous users to post objectionable images and symbols to certain groups of people. This types of posts include cyber-aggression, cyberbullying, offensive content, and hate speech. In this work, we analyze images posted on online social media sites to hurt online users. In this research, we designed a deep learning based system to classify aggressive post from a non-aggressive post containing symbolic images. To show the effectiveness of our model, we created a dataset crawling images from Google search to query aggressive images. The validation shows promising results.

Keywords: Cyber-aggression · Cyberbullying · Online Social Networks · Convolutional Neural Network · Augmentation

1 Introduction

With the emergence of the web-based popular Online Social Networks (OSN) such as Instagram[1], Facebook[2], Vine[3], these are exponentially increasing the user-generated content, that can reach billions of people in mere of a second. These sites make a user find people with common interests, share enormous real-time information, and eases business. In spite of these benefits, there are many detrimental outcomes associated with OSN such as Internet harassment, Cyber-aggression [4], Cyberstalking [12], Cyberbullying [23], and many more. Among them, Cyber-aggression is a growing and serious problem for online users. Cyber-aggression is defined as aggressive or hostile behavior that uses electronic media to cause harm to other people [8,9,13,14]. Cyber-aggression could occur

[1] www.instagram.com.
[2] www.facebook.com.
[3] https://vine.co/.

Published by Springer Nature Switzerland AG 2019
I. O. Pappas et al. (Eds.): I3E 2019, LNCS 11701, pp. 415–424, 2019.
https://doi.org/10.1007/978-3-030-29374-1_34

in various form like, written/verbal aggression (e-mails, instant messaging, chats, verbal post, etc.), visual-based aggression (posting, sending or sharing embarrassing images or video).

Cyber-aggression crosses all physical borderline. The Internet has abundantly opened up the global platform to users who access it on a wide-range of devices. Some users use free to post or send whatever they want on an online platform without bearing in mind how that content can inflict pain and sometimes cause severe psychological and emotional injuries. Online users can hide their identities through the Internet too easily [7]. The social networking site such as ask.fm[4] allow the users to post with hiding their identity. As a result, the experiences of a victim may be unnoticed, and the activities of a bully may remain uncontrolled. Even if bullies are recognized, many individuals are unaware of responding properly to these instances. Cyber-aggression can be continual phenomenon because of the easily available, and access of the Internet make viral and exposing the victims to an entire virtual world. It makes them feel sick and worthless. Although any age-group of social networking user could be affected by Cyber-aggression, teenagers and youngsters are the most affected people. Cyber-aggression on teenagers and youngsters have been shown to cause both mental and psychological issues. Most of the time, kids and teenagers use online social sites only with curiosity irrespective of knowing the potential risks [22]. Recent studies have reported that teenagers make generous use of image and video sharing online sites (e.g., Instagram, Vine) [18]. In particular, visual (image and video) content now accounts for more than 70 % of all web contents[5]. All together, there has been a substantial rise in using image and video content for Cyber-aggression [25] and it has been declared that Cyber-aggression grows bigger and meaner with photos and video [10]. The reality is that cyberbullying is one such issue that only becomes more severe if it is ignored. Therefore, it must be monitored at an earlier stage. The severity of the problem needs immediate attention from a technical point of view because manual detection is not scalable as well as time-consuming. Automated tools need to be developed, which can be helpful in automated monitoring [30] that can minimize the mental and physical health issues on users.

Therefore, this motivated us to develop an automated tool to detect the cases of Cyber-aggression so that users can feel safe and secure and get unconditional support. Identifying the Cyber-aggression on social media is a very challenging task due to several reasons, e.g., various form of post, multi-lingual text, the non-standard writing style of online users, etc. Most of the existing works [2–5,15,20,24,29] have solved Cyber-aggression issues based on text. Some recent works [9,28] tried to solve Cyberbullying issues related to the image-based post. Hosseinmardi et al. [9] built a model to predict Cyberbullying incidents on the Instagram network based on initial user data such as the post of an image with associated text caption, and the number of followers & followings. Singh et al.

[4] https://ask.fm/.

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

[28] built a model to identify Cyberbullying incidents on the Instagram network with visual and text features. To the best of our knowledge, no previous work has been proposed to detect Cyber-aggression on the image-based post, especially on symbolic images. We analyze aggressive post of several social media such as Facebook, Twitter, and Instagram and found that some post only the image part of the post are aggressive that contain direct aggression or indirect aggression in the form of symbolic aggression where bullies target the user to humiliate, insult, and to make fun of or mock them. We mainly considered those type of image where the post is having both types of direct or indirect aggression. The last decade has provided considerable research on the causes and effects of text-based aggression on social media, but there is no research has been done on a symbolic type of aggression related to the image-based post. Due to a scarcity of aggressive image based post, we created a dataset crawling images from Google search to query aggressive images.

The current approach focused on the image content of the detection of Cyber-aggression on OSN. We target to detect Cyber-aggression because it may lead to Cyberbullying events in the near future. Our method is tested on the dataset of 3600 images. We propose a deep Convolutional Neural Network (CNN) for identification of Cyber-aggression on social media. The fundamental idea of CNN is to consider features extraction and classification task as collaboratively trained task. The idea of using deep CNN (many layers of convolutions and pooling) to extract a hierarchical representation of the input sequentially. For generalization purpose, we augmented the image and used a dropout layer in between the two convolutional layers. Our main contributions can be summarized as:

- Creation and labeling of Cyber-aggressive posts containing symbolic images from Google search to query aggressive images.
- A deep Convolutional Neural Network based system to classify images containing symbolic aggression and no-aggression.

The remainder of the paper is organized as follows: Sect. 2 presents related works in Cyber-aggression detection while Sect. 3 presents our proposed framework for Cyber-aggression detection. The finding of the proposed system is presented in Sect. 4. Finally, we conclude the paper and discusses future work directions in Sect. 5.

2 Related Works

Cyber-aggression is widely recognized as a social challenge from the last few years, especially for teenagers and youngsters [26]. Recently, a number of researches have been proposed to address Cyber-aggression over online platforms. In this section, we briefly discuss some of the potential works proposed in this domain.

A number of works [4–6] performed Cyber-aggression classification on English text whereas [2,3,15,20,24,29] performed Cyber-aggression classification on

multi-lingual text. The Cyber-aggression classification performed by [4] on twitter. They found that when user and network-based features are combined with text-based features gave better accuracy. They got overall precision and recall of 0.72 and 0.73 respectively for four classes classification: Bully, Aggression, Spam, and Normal tweets. Chavan and Shylaja [5] detected Cyber-aggression on unknown social media. They used the Term Frequency-Inverse Document Frequency (TF-IDF), and n-gram features, Support Vector Machine (SVM) and logistic regression as a classifier. They reported the best Area Under Curve (AUC) score was 0.87. Chen et al. [6] detected aggressive tweets using Convolutional Neural Network (CNN) based on a sentiment analysis method. They found the accuracy of 0.92. Raiyani et al. [20] used dense system architecture on the multi-lingual text. Their system was suffered from false positive cases, and they removed the words that are not found in the vocabulary. Julian and Krestel [21] used ensemble learning and data augmentation techniques. They augmented training dataset using machine translation of three different languages. Their system is not stable, especially for Hindi dataset for the same domain it was performed well, but for other domain, it fails to classify the tweets with good accuracy. Aroyehun and Gelbukh [2] used various deep learning models such as Long Short Term Memory (LSTM), CNN, and FastText as word representation. Their system was not clearly classified covertly aggressive comments from overtly aggressive comments with significant accuracy. Modha and Majumder [15] used various deep learning models such as LSTM, CNN, Bidirectional LSTM, and FastText as word representation and machine learning classifiers. They used ensemble learning based on majority voting scheme. Samghabadi et al. [24] used ensemble learning based on various machine learning classifiers such as logistic regression, SVM and word n-gram, character n-gram, word embedding, sentiment, etc., as a feature set. Srivastava et al. [29] identified online social aggression on Facebook comment and Wikipedia toxic comments using stacked various LSTM units followed by Convolution layer and Fasttext as word representation. They achieved 0.98 AUC for Wikipedia toxic comment classification. For code-mixed English dataset, they achieved a weighted F1 score of 0.63 for the Facebook domain and 0.59 for the Twitter domain.

Very few researchers [9,28] have begun using visual characteristics to identify Cyberbullying. Hosseinmardi et al. [9] anticipated the Cyberbullying event taking into account visual characteristics and using original user data such as picture, caption, number of followers and followings, but visual characteristics do not help. By integrating textual and visual characteristics, Singh et al. [28] identified Cyberbullying. Their sample of practice is very small and high adverse words in the dataset predominated. Most of the work performed in the Cyber-aggression domain is concentrated on the text in particular. Very few operates with image-based post on the detection of Cyber-aggression. Best of our knowledge, there is no work on symbolic aggression classification.

3 Methodology

To automatically detect Cyber-aggression in OSN, we propose a Convolutional Neural Network (CNN) approach. In the following subsections, we describe the details of the dataset in Subsect. 3.1. The deep CNN based model is described next in Subsect. 3.2.

3.1 Data Collection and Labelling

We analyze the aggressive post containing images used by Internet users of multiple social media sites such as Facebook, Twitter, and Instagram, etc., and discovered that users of these sites usually use aggressive symbolic images to insult, harass, and humiliate other Internet users. We gathered some images from these social media (Facebook, Twitter, and Instagram). Because of the scarcity of marked information to make machine learning classifiers to identify the post contains aggression, we use Google Search to query aggressive images; specifically, we used some keywords such as aggressive images, Cyber-aggressive images, bullying pictures, etc. These images are manually filtered based on the clarity of decidable for a level of aggression, and then finally, we got a total of 3600 images. Three graduate students volunteered to annotate the images. They individually annotated the images into three classes of aggression: high aggression, medium aggression, and no aggression. We considered only those images on which at least two students agreed. The images which are having physical threats are labelled as high aggressive images, images which are having indirect aggression are labelled as medium aggressive images, and the images which do not have any threat are labelled as not aggressive images. The details of our dataset can be seen in Table 1.

Table 1. Description of Cyber-aggressive image dataset

Image class	Number of sample
no_aggression	1566
midium_aggression	1080
high_aggression	954
total images	3600

3.2 Proposed Model

The Convolutional Neural Network (CNN) is a deep neural network architecture that can take the image as an input and extract essential features in their hidden layers to do the classification task. In the proposed model, we used six layers of convolution, followed by three dense layers. We used max-pooling layer after every two convolution layer. We also used dropout between each of the CNN

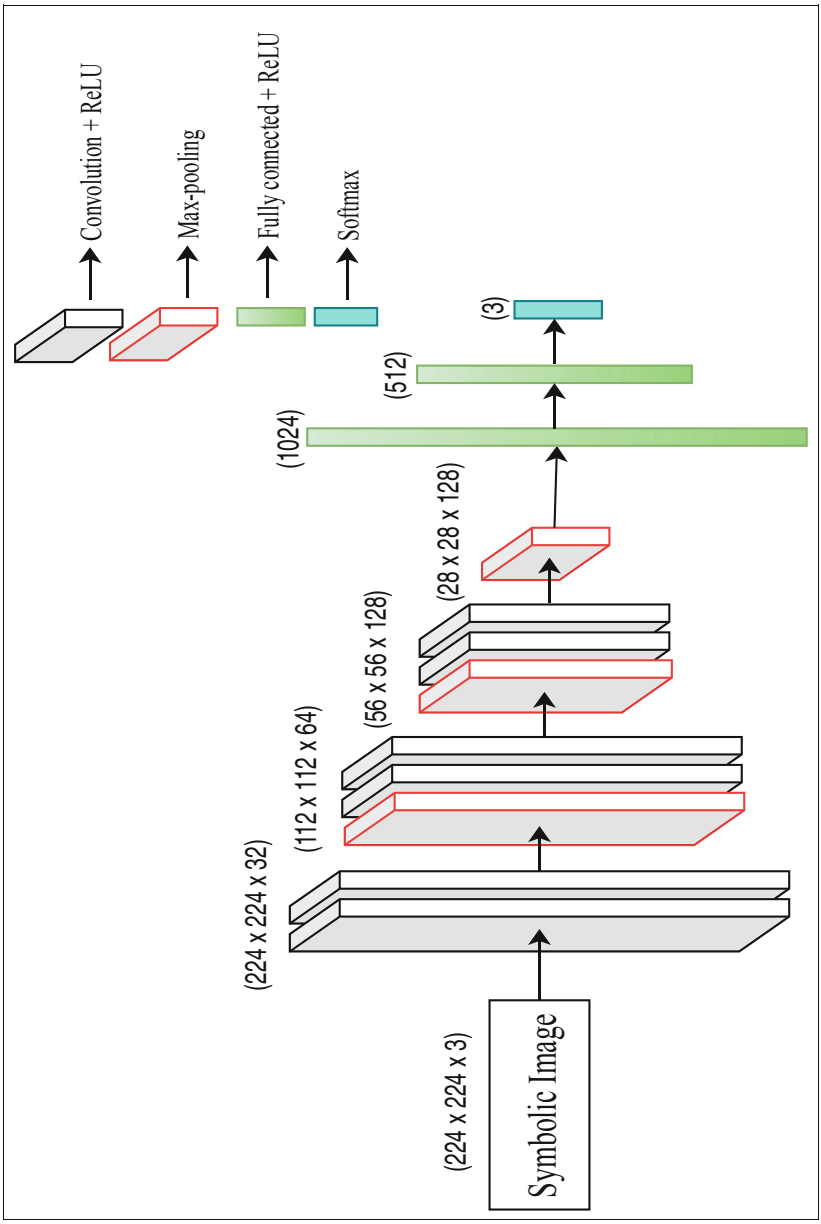

Fig. 1. Overview of the proposed CNN based architecture

as well as dense layers. The overall architecture for the proposed CNN based model can be seen from Fig. 1. We first converted all the images into an equal size, i.e., $224 \times 224 \times 3$. For the normalization of pixel values, the pixel matrix of the image is divided by the maximum pixel value, i.e., 255 and given to the CNN layers. We used 32, 32, 64, 64, 128, and 128 filters each of size 3×3 in the first, second, third, fourth, fifth, and sixth convolution layer respectively. After applying all the convolution operation, we flatten the feature vector and then passes it into the dense layers as can be seen in Fig. 1.

The detailed explanation of the CNN model can be seen in [11]. The normalized matrix is then used by the proposed system to train and test the model. In every cases, out of total data samples, 75% of them were used for training, and the remaining 25% samples were used for testing the performance of the models. In the hidden layers, ReLU activation function and softmax activation function at the output layer used. The model performed best with categorical cross-entropy as a loss function, the learning rate 0.001, batch size 10, and an epoch of 100. Table 2 listed all the hyper-parameters used during our experiments.

Table 2. Hyper-parameters setting for the proposed model

Description	Values
Filter region size	3×3
Feature map	32, 32, 64, 64, 128, 128
Pooling size	2×2
Activation function	ReLU, Softmax
Dropout rate	0.2
Learning rate	0.01
Batch size	10
Epoch	100

4 Results and Discussions

The findings of our current strategy to symbolic image-based Cyber-aggression detection are described in this section. Precision, recall, and weighted F1-score are the performance metrics used. The result of the proposed Convolutional Neural Network on the dataset after image augmentation tabulated in Table 3. The proposed model achieved a precision of 0.86, 0.91, and 0.93 for no-aggression, medium-aggression, and high-aggression class, respectively. The corresponding recall values are 0.95, 0.89, and 0.79. The weighted F1-score for no-aggression, medium-aggression, and high-aggression class are 0.91, 0.90, & 0.86 respectively. We also experimented with VGG-16 [27], which generally perform well for image classification in several scenarios [1,16,17,19]. In VGG-16 experimentation, we

Table 3. Results of detection of Cyber-aggressive images

Approach	Image class	Results		
		Precision	Recall	F1-score
VGG-16	no_aggression	0.67	0.87	0.76
	midium_aggression	0.74	0.63	0.68
	high_aggression	0.76	0.52	0.62
	weighted average	**0.72**	**0.71**	**0.70**
CNN	no_aggression	0.86	0.95	0.91
	midium_aggression	0.91	0.89	0.90
	high_aggression	0.93	0.79	0.86
	weighted average	**0.90**	**0.89**	**0.89**

trained the last two layers of VGG-16, and all other layers marked as non-trainable. The VGG-16 model achieved 0.70 weighted F1-score, whereas our CNN based model, which is a lesser number of layers in compare to VGG-16, got 0.89 weighted F1-score. As shown in Table 3, our model can currently identify 93 out of 100 cases of predicted high-aggression post.

One of our major contributions is the creation of labelled dataset for aggressive symbolic images. There is no such labelled dataset exist which contains symbolic images for Cyber-aggression detection task. Therefore, we collected symbolic images from different online social sites and then labelled them into three classes of aggression. Next contribution is the development of Convolutional Neural Network based model, which is six layers of convolution and performed well for image classification. Our model can able to classify the images with good F1-score of around 90% whereas the VGG-16 model achieved F1-score of 0.70.

5 Conclusion and Future Work

In this article, we presented a deep Convolutional Neural Network based approach to identify aggressive posts containing symbolic images. We used six layers of convolution, followed by three dense layers and got weighted F1-score of 89% for aggressive post-identification. We explored the existing models of VGG-16 [27] to compare the performance of the pre-trained model for this task and found that our CNN based model performed better than VGG-16 which has more number of layers in compare to our model. One of the major limitations of the current research is that it only considers the symbolic images ignoring any textual contents associated with those images that may be more correct informative content for identifying the aggression on symbolic images. In the future, the textual contents can be exploited along with the symbolic images to make this system more robust and more accurate.

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

References

1. Alam, F., Imran, M., Ofli, F.: Image4Act: online social media image processing for disaster response. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 601–604. ACM (2017)
2. Aroyehun, S.T., Gelbukh, A.: Aggression detection in social media: using deep neural networks, data augmentation, and pseudo labeling. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 90–97 (2018)
3. Arroyo-Fernández, I., Forest, D., Torres-Moreno, J.M., Carrasco-Ruiz, M., Legeleux, T., Joannette, K.: Cyberbullying detection task: the EBSI-LIA-UNAM system (ELU) at COLING 2018 TRAC-1. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 140–149 (2018)
4. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A.: Mean birds: detecting aggression and bullying on Twitter. In: Proceedings of the 2017 ACM on Web Science Conference, pp. 13–22. ACM (2017)
5. Chavan, V.S., Shylaja, S.: Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2354–2358. IEEE (2015)
6. Chen, J., Yan, S., Wong, K.C.: Verbal aggression detection on Twitter comments: convolutional neural network for short-text sentiment analysis. Neural Comput. Appl., 1–10 (2018). https://doi.org/10.1007/s00521-018-3442-0
7. Dadvar, M., Trieschnigg, D., de Jong, F.: Experts and machines against bullies: a hybrid approach to detect cyberbullies. In: Sokolova, M., van Beek, P. (eds.) AI 2014. LNCS (LNAI), vol. 8436, pp. 275–281. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06483-3_25
8. Grigg, D.W.: Cyber-aggression: definition and concept of cyberbullying. J. Psychol. Couns. Sch. **20**(2), 143–156 (2010)
9. Hosseinmardi, H., Rafiq, R.I., Han, R., Lv, Q., Mishra, S.: 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), pp. 186–192. IEEE (2016)
10. Kornblum, J.: Cyberbullying grows bigger and meaner with photos, video. Eschool News (2008). https://www.eschoolnews.com/2008/07/15/cyber-bullying-grows-bigger-and-meaner-with-photos-video/
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
12. League, A.D.: Glossary of cyberbullying terms. adl. org (2011). http://www.adl.orgleducationlcurricuIurnconnectionslcyberbullyingjglossary.pdf
13. Machackova, H., Dedkova, L., Sevcikova, A., Cerna, A.: Bystanders' supportive and passive responses to cyberaggression. J. Sch. Violence **17**(1), 99–110 (2018)

14. Modecki, K.L., Barber, B.L., Vernon, L.: Mapping developmental precursors of cyber-aggression: trajectories of risk predict perpetration and victimization. J. Youth Adolesc. **42**(5), 651–661 (2013)
15. Modha, S., Majumder, P., Mandl, T.: Filtering aggression from the multilingual social media feed. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 199–207 (2018)
16. Nguyen, D.T., Ofli, F., Imran, M., Mitra, P.: Damage assessment from social media imagery data during disasters. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 569–576. ACM (2017)
17. Nguyen, D.T., Alam, F., Ofli, F., Imran, M.: Automatic image filtering on social networks using deep learning and perceptual hashing during crises. arXiv preprint arXiv:1704.02602 (2017)
18. Pater, J.A., Miller, A.D., Mynatt, E.D.: 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, pp. 2305–2314. ACM (2015)
19. Paul, R.: Classifying cooking object's state using a tuned VGG convolutional neural network. arXiv preprint arXiv:1805.09391 (2018)
20. Raiyani, K., Gonçalves, T., Quaresma, P., Nogueira, V.B.: 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), pp. 28–41 (2018)
21. Risch, J., Krestel, R.: Aggression identification using deep learning and data augmentation. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), pp. 150–158 (2018)
22. Rybnicek, M., Poisel, R., Tjoa, S.: Facebook watchdog: a research agenda for detecting online grooming and bullying activities. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2854–2859. IEEE (2013)
23. Salawu, S., He, Y., Lumsden, J.: Approaches to automated detection of cyberbullying: a survey. IEEE Trans. Affect. Comput. (2017). https://doi.org/10.1109/TAFFC.2017.2761757
24. Samghabadi, N.S., Mave, D., Kar, S., Solorio, T.: RiTual-uh at TRAC 2018 shared task: aggression identification. arXiv preprint arXiv:1807.11712 (2018)
25. Seiler, S.J., Navarro, J.N.: Bullying on the pixel playground: investigating risk factors of cyberbullying at the intersection of children's online-offline social lives. Cyberpsychol.: J. Psychosoc. Res. Cyberspace **8**(4) (2014). http://dx.doi.org/10.5817/CP2014-4-6
26. Servance, R.L.: Cyberbullying, cyber-harassment, and the conflict between schools and the first amendment. Wis. Law Rev. **6**, 1213–1244 (2003)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
28. Singh, V.K., Ghosh, S., Jose, C.: Toward multimodal cyberbullying detection. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pp. 2090–2099. ACM (2017)
29. Srivastava, S., Khurana, P., Tewari, V.: 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 (2018)
30. Van Royen, K., Poels, K., Daelemans, W., Vandebosch, H.: Automatic monitoring of cyberbullying on social networking sites: from technological feasibility to desirability. Telematics Inform. **32**(1), 89–97 (2015)