GRADUATE CERTIFICATE IN INTELLIGENT REASONING SYSTEMS PRACTICE MODULE REPORT

Abuse Image Detection

Institute of Systems Science, National University of Singapore, Singapore 119615

ABSTRACT

Offensive Image is pervasive in social media. Individuals frequently take advantage of the perceived anonymity of computer-mediated communication, using this to engage in behavior that many of them would not consider in real life. Online communities, social media platforms, and technology companies have been investing heavily in ways to cope with offensive language in the form of text or images to prevent abusive behavior in social media.

The proliferation of, social media platforms resulted in a remarkable increase in user-generated content. These platforms have empowered users to create, share and exchange content for interacting and communicating with each other. However, these have also opened new avenues to cyberbullies and haters who can spread their negativity to a larger audience, often anonymously. Due to the pervasiveness and severity of this behavior, many automated approaches that employ natural language processing (NLP), machine learning and deep learning techniques have been proposed in the past. This survey offers an extensive overview of the state-of-the-art approaches proposed by the research community to identify offensive content. Based on our comprehensive literature survey, a categorization of different approaches and features employed by the researchers in the detection process are presented. This survey also incorporates the major challenges that require considerable research efforts in this domain. Finally, future research directions with an aim of developing robust abusive content detection system for social media are also discussed.

"The Online Criminal Harms Act will empower the Singapore government to issue directions to individuals, entities, online service providers and app stores requiring them to remove or block access to potentially criminal content."

1. INTRODUCTION

A dramatic change has occurred in the dynamics of child sexual abuse and exploitation as a result of the technological revolution in how images can be made and shared. Since the marketing of cell phones, many people in parts of the world carry easily useable cameras and videos, which have made it trivially easy to create and share images directly, bypassing the need to have a third party develop and view them.

Not surprisingly, child sexual abusers began to take images of their victims for personal gratification and to share with others for status and commercial gain (Casanova et al., 2000; Jenkins, 2001). Law enforcement in turn started to uncover troves of these images on the devices of abusers and on sharing websites (Wolak et al., 2011). As Internet surveillance technology developed and reporting requirements and enforcement efforts ramped up, the general sense among law enforcement has been that child sexual image availability was increasing exponentially (Bursztein et al., 2019; Ibrahim, 2022; WeProtect Global Alliance, 2022).

In their efforts to clearly distinguish this child image contraband from legal pornography, law enforcement and advocacy groups moved away from using the term "child pornography," the term encoded in many long-standing criminal statutes, preferring terms like "child sexual abuse images," (CSAI) or "child sexual abuse material" (CSAM) (Martellozzo, 2019). These terms were thought to better characterize what were deemed to be images made by adult sexual abusers of their crime victims.

Yet, dynamics have continued to evolve as photo sharing has become easier and more common. Youth began to use the technology to share sexual images of themselves, their friends, and their intimate partners across many contexts. These included courtship, intimate play, and humor, but also in contexts of bullying, aggression, and partner abuse (Gordon-Messer et al., 2013; Lenhart et al., 2010). Law enforcement has come to refer to these as "youth produced images" (Wolak et al., 2012). As social norms have changed, these youth produced images became as numerous as the adult made child images of original concern. Indeed, an inventory of the International Child Sexual Exploitation Image Database, a law enforcement investigation tool, found that from 2010 onward self-produced youth images comprised 40 % or more of all images in the archive (Quayle et al., 2018).

Youth produced images have created a variety of complexities. On the one hand, even if voluntarily self-produced without coercion, most images do qualify as criminal contraband under existing child pornography statutes, if they are images of juveniles engaged in sexual acts or with genitals exposed for purposes of arousal. But many juvenile justice advocates have been concerned about criminalizing young people for non-malicious sexual behavior occurring alone,

or voluntarily with peers or intimate partners (Barroso et al., 2023; Ojeda et al., 2022; Strasburger et al., 2019). Laws have been proposed to exempt certain classes of these images (O'Connor et al., 2017).

At the same time, some of these youth produced images are abusive and harmful for reasons beyond the mere sexual depiction of a juvenile. Which types of youth produced images are considered sexual abuse or exploitation (Madigan et al., 2018; Walker & Sleath, 2017)? As would be the case with adult abuser produced images, several malicious contexts qualify as abusive even though not all are images of a sex crime (Krieger, 2017; Strasburger et al., 2019). For example, some youth take sexual images non-consensually of their peers — sleeping, intoxicated, or surreptitiously. Sometimes youth take images with the intent to intimidate, humiliate, shock or extort other youth (Harper et al., 2021). These are non-consensual episodes that have similarities to CSAI, although the offenders are juveniles (Wolak et al., 2012).

In a different malicious youth produced scenario, youth make and share voluntarily images of themselves with friends or intimate partners, but then later find these images passed on to others or posted without their consent. This type of episode has been labeled with terms like aggravated or non-consensual sexting (Strasburger et al., 2019) or revenge pornography.

Yet another scenario involves youth producing sexual images of themselves and willingly sharing them under prohibited conditions. In some cases, the images are produced voluntarily but are exchanged for money or other valuables – a form of prohibited commercial sex. In other scenarios, the sharing is with adults either voluntarily or as a result of manipulation, both reflecting an abusive context of unequal power. This is parallel to the element of conventional sexual abuse called statutory sex offenses which are crimes even with voluntary victim participation when juveniles of a certain age participate in relationships with adults.

The dynamics of image abuse of children are varied. They are not just images taken by adult abusers. They are also not just images misused by romantic partners. They include misuse by other youth, both taking and sharing. They include misuse by adults, who receive voluntarily produced images from youth, sometimes for money.

There are several innovations proposed in this paper that need to be considered by this developing field. First, the field should move away from the term CSAI, to image based

2. OVERVIEW

In this project, I have created and refined machine learning and Deep Learning models to detect Abuse images in community help portal.

The goal of this project is to child abuse images detection with a focus on implicit abuse, to develop a model using NLP and Computer vision techniques to accurately detect Abusive and Non- Abuse images. Acquired data from Nude Net Classifier dataset v1 to detect abuse images.

3. PROBLEM STATEMENT

Offensive Image is pervasive in social media. Individuals frequently take advantage of the perceived anonymity of computer-mediated communication, using this to engage in behavior that many of them would not consider in real life. Online communities, social media platforms, and technology companies have been investing heavily in ways to cope with offensive language in the form of text or images to prevent abusive behavior in social media. A specific popular form of online harassment is the use of abusive language. One abusive or toxic statement is being sent every 30 seconds across the globe. The use of abusive language on social media contributes to mental or emotional stress, with one in ten people developing such issues. These abusive Tweets and comments detection and deletion in social media is more important. Because human brains reply quickly to pictures and color in contrast to other types of information, an image is an almost invincible draw on social media. Of course, to get the concentration you want, you must share images that matter to your target audience. Although this may lead to mass data abuse, images need to be detected and deletion in social media is more evitable.

4. DATASET

we need work to prevent the spread of illegal child sexual abuse material (referred to as CSAM). Child safety organizations and governments rightly expect — and in many cases require — us to take action to remove it from our systems. Which is why, when we find CSAM on our platforms, we remove it, report it and often take the step to suspend the account.

Category	Count
Safe Image	38411
Sexy Image	38005
Nude Image	38000

Fig. 1. Abuse Images Dataset.

5. PREPROCESSING

Preprocessing of data is carried out before the model is built and the training process is executed. Following are the steps carried out during preprocessing

• Initially the images are divided into training and validation sets. • The images are resized to 224 x 224 pixels. • All three channels were used during the training process as these are color images. • The images are normalized by dividing every pixel in every image by 255. • To ensure the mean is zero a value of 0.5 is subtracted.

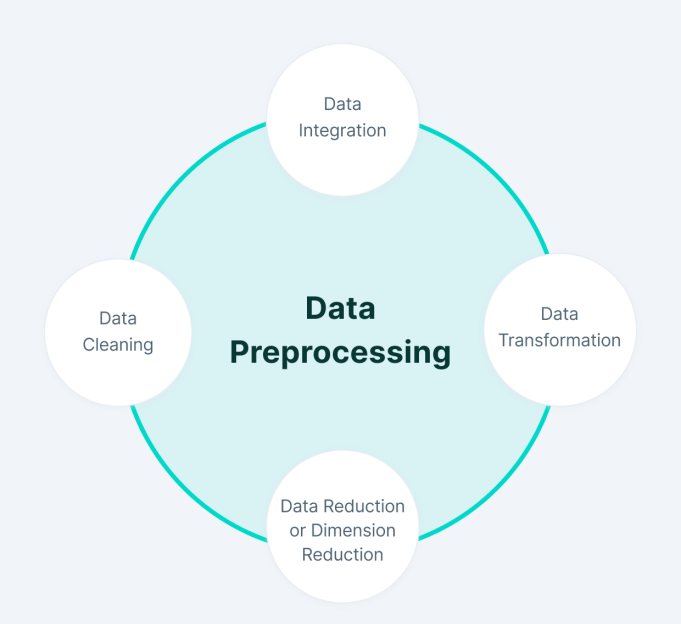

Fig. 2. Images Cleaning Pipeline

6. EXPERIMENTAL DESIGN

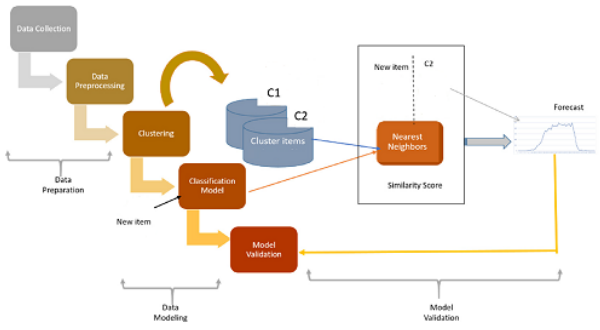

Fig. 3. Experimental design.

7. FEATURE EXTRACTION IN IMAGE PROCESSING

Image processing is one of the best and most interesting domain. In this domain basically you will start playing with your images in order to understand them. So here we use

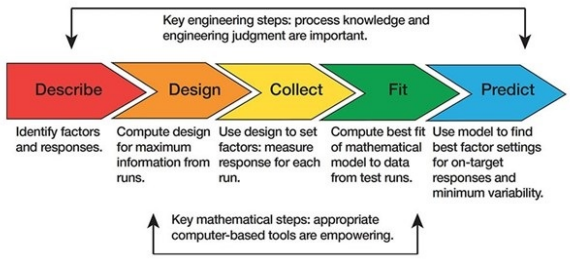

Fig. 4. Key Engineering Steps.

many many techniques which includes feature extraction as well and algorithms to detect features such as shaped, edges, or motion in a digital image or video to process them. Smaller numbers that are closer to zero helps to represent black, and the larger numbers which are closer to 255 denote white. So this is the concept of pixels and how the machine sees the images without eyes through the numbers.

But, for the case of a colored image, we have three Matrices or the channels

Red, Green and Blue. So in these three matrices, each of the matrix has values between 0-255 which represents the intensity of the color of that pixel.

If you have a colored image like the dog image we have in the above image on the left. so being a human you have eyes so you can see and can say it is a dog-colored image. But how a computer can understand it is the colored or black and white image?

So you can see we also have three matrices that represent the channel of RGB – (for the three color channels – Red, Green, and Blue) On the right, we have three matrices. These three channels are superimposed and used to form a colored image. So this is how a computer can differentiate between the images.

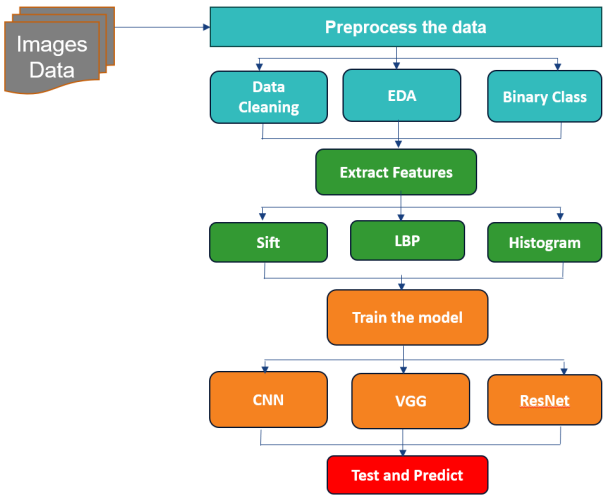

Fig. 5. Abuse Image Proposed system.

8. MODEL IMPLEMENTATION

A Convolutional Neural Network, also known as CNN or ConvNet, is a class of neural networks that specializes in processing data that has a grid-like topology, such as an image. A digital image is a binary representation of visual data. It contains a series of pixels arranged in a grid-like fashion that contains pixel values to denote how bright and what color each pixel should be.

VGG stands for Visual Geometry Group; it is a standard deep Convolutional Neural Network (CNN) architecture with multiple layers. The "deep" refers to the number of layers with VGG-16 or VGG-19 consisting of 16 and 19 convolutional layers. The VGG architecture is the basis of ground-breaking object recognition models. Developed as a deep neural network, the VGGNet also surpasses baselines on many tasks and datasets beyond ImageNet. Moreover, it is now still one of the most popular image recognition architectures.

ResNet (Residual Network) is a convolutional neural network that democratized the concepts of residual learning and skip connections. This enables to train much deeper models.

This is ResNet v1.5, which differs from the original model: in the bottleneck blocks which require downsampling, v1 has stride = 2 in the first 1x1 convolution, whereas v1.5 has stride = 2 in the 3x3 convolution. This difference makes ResNet50 v1.5 slightly more accurate (0.5

- Standard CNN architecture was initially created and trained. We have created 4 convolutional layers with 4 max pooling layers in between.
- Filters were increased from 64 to 512 in each of the convolutional layers.
- Also, dropout was used along with flattening layer before using the fully connected layer.
- Altogether CNN has 2 fully connected layers. The number of nodes in the last fully connected layer were setup as 10 along with SoftMax activation function.
- Relu activation function was used for all other layers. Xavier initialization was used in each of the layers.

9. REFINEMENT

To get the initial result simple CNN architecture was built and evaluated. This resulted in a decent loss. The public score for the initial simple CNN architecture(initial unoptimized model) was 2.67118. After this to further improve the loss, transfer learning was applied to VGG16 along with investigating 2 types of architectures for fully connected layer. Model Architecture1 showed good results and was improved further by using the below techniques

- Drop out layer was added to account for overfitting.
- Xavier initialization was used instead of random initialization of weights
- Zero mean was ensured by subtracting 0.5 during Preprocessing.
- Training was carried out with 400 epochs and with a batch size of 16
- To further improve the loss metric ,VGG16 along with Model Architecture1 was
- selected and fine-tuning was applied. SGD optimiser was used with very slow
- learning rate of 1e-4. A momentum of 0.9 was applied with SGD
- learning rate of 1e-4. A momentum of 0.9 was applied with SGD

10. EXPERIMENTAL RESULTS

Model	Accuracy
VGG Model #1	0.835
VGG Model #2	0.8208
resnet101 [Stratergy-1 : freezing all layer's parameters]	0.85822
Resnet101 [Stratergy-2 : freezing previous layers and training only last	0.91019

Fig. 6. Abuse Images Model Accuracy.

- Among Deep Learning methods, ResNet-101 with stratery-2 i.e., retrain last few layers gave best accuracy.
- Feature Extraction and Image Augmentation techniques helped in improving overall accuracy.

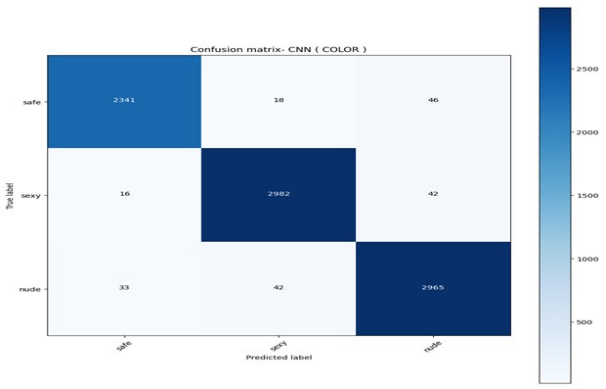

Fig. 7. CNN Confusion Matrix.

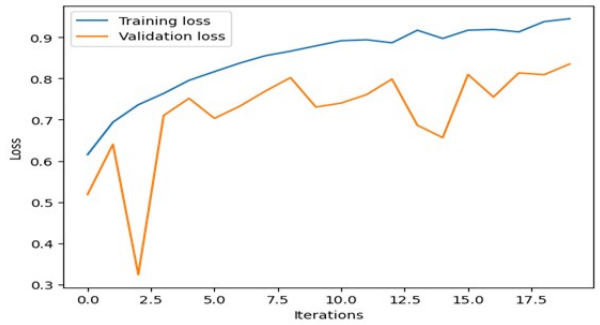

Fig. 8. CNN ROC.

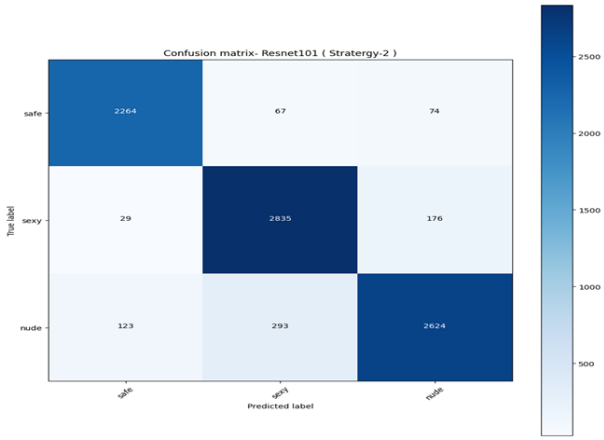

Fig. 9. ResNet-101 (Strategy-2) Confusion Matrix.

11. MAJOR CHALLENGES/ISSUES INVOLVED

From the analysis illustrated above, abusive language detection in social media is a challenging task due to the unstructured, subjective, informal, and often misspelled nature of the textual content. This section covers the major challenges discovered in this task and the possible directions for future research.

- Lack of benchmark data-sets: The main challenge for the abusive content detection is the dearth of benchmark data-sets in this field. Presently, the researchers collect data from different social media platforms and annotate it using one of two methods—their own labeling effort or by using crowd sourcing services such as Crowd Flower and Amazon's Mechanical Turk. Due to unavailability of standard data-sets, comparing different techniques is very difficult.
- Subjectivity involved: Another challenge is the myriad of forms and the lack of a clear, common definition of abusive behavior. Moreover, the notion of abusive or offensive is very subjective, and its degree varies considerably among people, making the labeling process more difficult.

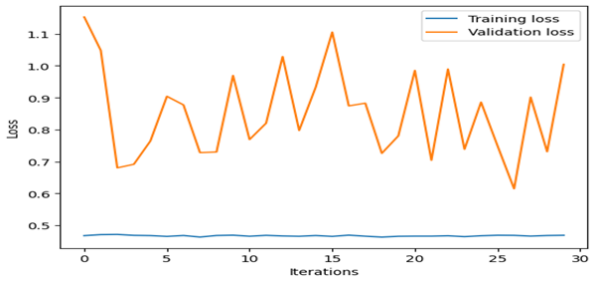

Fig. 10. ResNet-101 (Strategy-2) ROC.

- Sarcasm Detection: abusive comments may involve harassment without the use of profane words, satire, or irony, which are mainly difficult for machines to handle. For instance, the sentence, "You are as intelligent as Einstein" does not contain any profanity but may ironically be used to insult someone.
- Obfuscation: Simple keyword-based techniques fail in practice because of the intentional obfuscation of their text to evade keyword filters by the comment's. Strategies used by inventive users like symbol 280 Simrat Kaur et al. / Procedia Computer Science 189 (2021) 274–281 Author name / Procedia Computer Science 00 (2019) 000–000 7 substitution or false segmentations that still preserve the original semantics such as a$$hole or sh*t make these approaches ineffective [39].
- Context Sensitivity: Another challenge is to incorporate the context of the comment, especially in the case of threaded conversations. The potential for different interpretations of a word or sentence, if it is considered out of context, may impact the decision of the classifier. Based on the survey of numerous research articles, we give the following possible pertinent suggestions for future research: Most of the previous research in this field has performed binary classification only. So, there is a need to explore fine grained categories associated with abusive content such as insults, hate speech, threats etc. This fine-grained categorization will provide insight into different forms of inappropriate content and the degree to which they are alarming. For example,

12. CONCLUSION

This paper presented a comprehensive overview of the field by incorporating research articles spanning time of a decade and proposed taxonomies based on features and methods. The researchers have successfully applied methods from the machine-learning field, with Bag-of-Words (Bow) and N-grams being the most frequently used features in classification. Incorporation of more complex features i.e. from users'

profile, activity statistics and social graph structure have also been shown to be effective in classification. Because of the high dimensionality and sparsity issues of the previous models, several recent works have employed distributed word representations also known as word embeddings. More recently, deep learning-based architectures have also shown encouraging results in this domain. To sum up, deeper linguistic features, analysis of demographic influences and clear annotation guidelines are required to differentiate between different types of abuse accurately. Despite the availability of large amounts of work, it remains difficult to judge the effectiveness and performance of various features and classifiers, mainly due to the use of different datasets by each researcher. To make comparative evaluation possible, clear annotation guidelines and a benchmark dataset are must. From the analysis, it is evident that abusive content detection is yet an open area of interest for research fraternity, requiring more intelligent techniques to tackle the major challenges involved and thus making the online interaction a safer place for its users.

13. AUTHOR

Thota Siva Krishna , e0943696@u.nus.edu

14. REFERENCES

[1] Sourander, A., Klomek, A. B., Ikonen, M., Lindroos, J., Luntamo, T., Koskelainen M., & Helenius, H, (2010) "Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study", Archives of general psychiatry, 67(7), 720-728.

[2] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, L. Edwards (2009) "Detection of harassment on Web 2.0" In Proceedings of Content Analysis in the WEB 2.0 (CAW 2.0) Workshop at WWW, Madrid, April 20-24, 2009.

[3] K. Dinakar, R. Reichart, H. Lieberman (2011) "Modelling the Detection of Textual Cyberbullying." In ICWSM 2011, Spain, July 17-21.

[4] Chen,Y., Zhu, S., Zhou, Y., Xu, H. (2011) "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety." In: Symposium on Usable Privacy and Security. Pittsburgh, USA.

[5] Sood, S. O.; Churchill, E. F.; and Antin, J. (2012) "Automatic identification of personal insults on social news sites." Journal of the American Society for Information Science and Technology, 270-285.

[6] Munezero, M., Montero, C.S., Kakkonen, T., Sutinen, E., Mozgovoy, M. and Klyuev, V. (2014) "Automatic detection of antisocial behaviour in texts." Informatica, 38(1), p.3.

[7] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali (2017) "Mean Birds: Detecting Aggression and Simrat Kaur et al. / Procedia Computer Science 189 (2021) 274–281 281 8 Author name / Procedia Computer Science 00 (2019) 000–000 Bullying on Twitter." In Proceedings of the 2017 ACM on Web Science Conference (pp. 13-22). ACM.

[8] Chen, H., Mckeever, S., Delany, S. J. (2017) "Harnessing the Power of Text Mining for the Detection of Abusive Content in Social Media." In Advances in Computational Intelligence Systems (pp. 187-205). Springer International Publishing.

[9] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber (2017) "Automated Hate Speech Detection and the Problem of Offensive Language." In Proc. of ICWSM.

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

[11] Xu, Z., Zhu, S. (2010) "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).

[12] Justo, R., Corcoran, T., Lukin, S. M., Walker, M., Torres, M. I. (2014) "Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web." Knowledge-Based Systems, 69, 124-133.

[13] Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. (2015) "Antisocial behavior in online discussion communities." In Proceedings of ICWSM. Menlo Park, California: AAAI Press.

[14] Zhao, R., Zhou, A., 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.

[15] E. Cambria, D. Olsher, D. Rajagopa (2014) "Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis." In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI Press.

[16] Dadvar, M., de Jong, F.M.G., R.J.F., Ordelman, Trieschnigg, D.(2012) "Improved Cyberbullying Detection Using Gender Information." In: Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop, pp. 23-26., Belgium.

[17] Balakrishnan, V. (2015) "Cyberbullying among young adults in Malaysia: the roles of gender, age and internet frequency." Computers in Human Behavior, 46, 149e157.

[18] Q. Huang, V. K. Singh and P. K. Atrey (2014) "Cyber bullying detection using social and textual features." The 3rd ACM MM Workshop on Socially Aware Multimedia (IWSAM14), Orlando, FL, USA.

[19] K. Reynolds , A. Kontostathis , L. Edwards (2004) "Using Machine Learning to Detect Cyberbullying." Proceedings of the 10th International Conference on Machine Learning and Applications and Workshops, p.241-244, December 18-21.

[20] Waseem, Z., Hovy, D. (2016) "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter" In SRW@ HLT-NAACL (pp. 88-93).

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

[22] Nahar, V., Al-Maskari, S., Li, X., Pang, C (2014). "Semi-supervised Learning for Cyberbullying Detection in Social Networks." In ADC, pp. 160-171, 2014.

[23] Di Capua, M., Di Nardo, E., Petrosino, A. (2016) "Unsupervised cyber bullying detection in social networks." In Pattern Recognition (ICPR), 2016 23rd International Conference on (pp. 432-437). IEEE.

[24] Marzieh Mozafari, Reza Farahbakhsh, Noel Crespi.(2019) "A BERT-based transfer learning approach for hate speech detection in online social media. Complex Networks" 8th International Conference on Complex Networks and their Applications, pp.928-940.

[25] Razavi, A. H., Inkpen, D., Uritsky, S., Matwin, S (2010) "Offensive language detection using multi-level classification." In Canadian Conference on Artificial Intelligence, pp. 16-27, Springer Berlin Heidelberg.

[26] Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012) "Common sense reasoning for detection, prevention, and mitigation of cyberbullying." ACM Trans. Interact. Intell. Syst. 2, 3, Article 18(September 2012).

[27] Dadvar, M., Trieschnigg, D. and de Jong, F., May (2014) "Experts and machines against bullies: A hybrid approach to detect cyberbullies." In Canadian Conference on Artificial Intelligence, pp. 275-281, Springer International Publishing.

[28] Badjatiya, P., Gupta, S., Gupta, M., Varma, V. (2017) "Deep learning for hate speech detection in tweets". In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 759-760). International World Wide Web Conferences Steering Committee.

[29] Pavlopoulos, J., Malakasiotis, P., Androutsopoulos, I. (2017) "Deep Learning for User Comment Moderation." In Proceedings of the First Workshop on Abusive Language Online, Vancouver, Canada.

[30] Pitsilis, G. K., Ramampiaro, H., Langseth, H. (2018) "Detecting Offensive Language in Tweets Using Deep Learning" arXiv preprint arXiv:1801.04433.

[31] Zhang, Z., Robinson, D., Tepper, J. (2018) "Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network." In Proc. of ESWC (Vol. 18).

[32] Jason Brownlee PhD, https://machinelearningmastery.com/how-to-get-startedwith-deep-learning-for-computer-vision-7-day-minicourse/, Deep Residual Learning for Image Recognition.

[33] https://snappishproductions.com/blog/2018/01/03/classactivation-mapping-in-pytorch.html.

[34] https://github.com/Garima13a/YOLO-ObjectDetection/blob/master/YOLO.ipynb.

[35] https://colab.research.google.com/github/d2l-ai/d2len-colab/blob/master/chapterdeep learning computation/use gpu.ipynbscrollT o = mwrEJrSCo OR.

[36] https://github.com/opencv/opencv.