

Cyber Guardian: CNN-based Cyberbullying Prevention

Chaitanya Limaye¹, Bhavesh Bhatia², Chitra Atlani³, Siyona Singh⁴ and Rupali Soni⁵

¹⁻⁵V.E.S. Institute of Technology, Department of Computer Engineering, Mumbai, India

Email: 2020.chaitaniya.limaye@ves.ac.in, 2020.bhavesh.bhatia@ves.ac.in, 2020.chitra.atlani@ves.ac.in,
2020.siyona.singh@ves.ac.in, rupali.hande@ves.ac.in

Abstract— Social media apps are very common among young people due to the growing use of technologically advanced devices, which may be both beneficial and disadvantageous. This study demonstrates the viability of CNNs in cyberbullying detection through a systematic approach encompassing data preprocessing, analysis, and model training. The results suggest promising outcomes for utilizing deep learning techniques in combating online harassment. Our project's primary goal is to create a tool that uses social media intelligence and analysis methods to proactively detect and lessen cyberthreats across the many stages of product design and development. Cyberbullying is a relatively new phenomena that has affected people of all ages, particularly teenagers, on a socio-psychological level over the last ten years. As digital technology advances, young people are becoming more reliant on social media, which increases the likelihood of cyberbullying.

Index Terms— Cyberbullying, social-psychology, machine learning, product design, cybersecurity Cyberbullying Detection, Convolutional Neural Networks, Text Classification, preprocessing and Machine Learning.

I. INTRODUCTION

Information and communication technologies (ICTs) have become deeply integrated into our lives, blurring the lines between leisure, entertainment, learning, and even family interactions. While offering undeniable benefits like instant information access and global communication, ICTs also present challenges, particularly for adolescents and young adults who navigate both physical and virtual spaces. One such challenge is cyberbullying[1], a growing concern that involves intentionally inflicting harm on others through digital devices. With the advancement of technology, cybercrimes have become more sophisticated[2]. Recent studies, including the National Center for Education Statistics (NCES) 2019 report, reveal that cyberbullying is alarmingly prevalent, affecting approximately 20% of adolescents aged 12 to 18. A point to note is that 15% of these incidents occur online, highlighting the pervasiveness of this issue. Cyberbullying, as defined by Armitage et al. [3](2023), is characterised by repetitive and harmful behaviour using digital technologies, causing significant distress and long-term consequences for victims. These consequences include mental health issues, social difficulties, academic challenges, and even increased risk of criminal activity. The rapid evolution of technology has also significantly impacted the landscape of cybercrime. With the widespread use of smartphones, IoT devices, social media platforms, cloud storage, and cryptocurrency, cybercriminals have access to new avenues for malicious activities. As documented in various research reports, cybercrime costs the global economy trillions of dollars annually[4], and this figure is expected to rise.

The project aims to improve cybersecurity resilience in digital products and services by integrating social media

intelligence with product design, enabling proactive threat identification and mitigation.

II. MOTIVATION

The development of an application to detect cyberbullying is aimed at addressing the increasing number of cyber crimes and reducing emotional distress experienced by victims.

While technology offers immense benefits, its convenience attracts misuse, leading to the rise of "Cybercrime"[5]. This illicit activity thrives on the expanding digital landscape: internet access, connected devices, and a larger attack surface for individuals, businesses, and governments. From hacking and data breaches to identity theft, phishing, and cyberbullying, the spectrum of cybercrime is vast and destructive. These threats entail financial losses, privacy violations, and lasting emotional harm for victims, further compounded by cybercriminals' relentless innovation. Their sophisticated tools and techniques pose unique challenges to defence efforts. Beyond financial gains, cybercrime fuels social anxieties and psychological distress[6]. Cyberbullying and online harassment leave devastating emotional scars, highlighting the need for robust detection and investigation systems. The ubiquity of online threats demands proactive solutions. Our system is designed to combat cybercrime by leveraging cutting-edge technology and social media intelligence. By meticulously analysing online activity, we aim to detect malicious behaviour, mitigate its impact, and empower responsible online interactions by using NLP based ML algorithms like CNN.

III. RELATED WORK

Systems do exist in this domain, but they have some or the other setback, like they are either suitable for very specific conditions, or giving results that are less accurate. Following is the description of some of the various resources and what they lack:

In [7], a semi-supervised method for identifying cyberbullying by leveraging five distinct features (Sentimental Features, Sarcastic Features, Syntactic Features, Semantic Features and Social Features) that characterise cyberbullying content, employing the BERT model. Focusing solely on sentiment features, their BERT model achieved an accuracy of 91.90% after dual-cycle training, surpassing conventional machine learning models. The potential for the BERT model to deliver even higher accuracy hinges on access to extensive datasets. Integrating all proposed features outlined in their study could promise enhanced cyberbullying detection capabilities. An application harnessing these features could effectively identify and flag bullying content for further action. Additionally, combining complementary models with BERT can hold promise for developing a cutting-edge solution tailored to the nuances of cyberbullying detection in natural language processing tasks.

The goal is to develop an effective method for identifying and addressing online abusive content by combining NLP and ML to create a model that detects offensive language in English and Hinglish [8]. They found that CV slightly outperforms Term Frequency - Inverse Document Frequency (TFIDF) in accuracy and that Linear SVC and Stochastic Gradient Descent (SGD) offer better results in classifying bullying messages in Hinglish, with faster training times. However, deeper analysis in sentiment, semantics, and syntax could improve accuracy further. Integrating the model with various social media platforms can aid in reducing cyberbullying. The main challenge is obtaining large, accurately labelled datasets in Hinglish for ML training, as existing datasets are limited in size and reliability.

The work in [9] introduces a model designed to automatically detect cyberbullying content across multiple languages, addressing a crucial need in managing social media content and safeguarding users from the harmful effects of toxic remarks such as verbal attacks and offensive language. Their study evaluates the performance of various neural network models, with the CNN-BiLSTM network demonstrating the highest accuracy. Unlike the CNN model, which focuses solely on local word n-grams, the CNN-BiLSTM model can capture both local characteristics and global features, including long-term dependencies. The paper can explore the integration of image and video analysis to determine if cyberbullying detection can be automated across multimedia content. This is part of their future work.

Current approaches to cyberbullying detection, specifically focusing on session-based detection in social media was explored [10]. Social Media Session-based Cyberbullying Detection (SSCD) framework and review research progress within this framework, including model and dataset creation related to social media sessions was introduced. Their comparative experiments assess state-of-the-art models on SSCD datasets and suggest avenues for future research. They highlight the need to consider key cyberbullying characteristics like repetition and power imbalances in model design and dataset creation, emphasizing the potential of fine-grained detection methods. During the detection of cyberbullying in social media sessions, they encountered significant challenges like

enhancing transparency and quality in dataset management, advancing fine-grained detection capabilities, and ensuring model reliability and reproducibility that need addressing. A machine learning-based approach for cyberbullying detection and conducted evaluations using two classifiers, [11] SVM and Neural Network, along with TFIDF and sentiment analysis algorithms for feature extraction was introduced. Their model achieved 92.8% accuracy with Neural Network using 3-grams and 90.3% accuracy with SVM using 4-grams when combining TF IDF and sentiment analysis. Notably, the Neural Network demonstrated superior performance with an average f-score of 91.9%, compared to SVM's average f-score of 89.8%. Additionally, their model surpassed related work in accuracy and f-score metrics. Despite these advancements, detecting cyberbullying patterns is constrained by training data size. Larger datasets can be used which are essential for improved performance, making deep learning techniques more suitable due to their proven superiority over machine learning methods with larger datasets. The oversight of sentence semantics in existing academic methods are addressed by utilizing word2vec to train customized word embeddings [19]. They develop an LSTM-CNN architecture tailored for cyberbullying detection, surpassing traditional approaches on Twitter data. Their model, with a 97% ROC AUC score, includes a web application for toxicity-based tweet classification and a Telegram Bot for cyberbullying prevention. They also create Chrome Extensions for NSFW content moderation on WhatsApp Web. While their solution is effective, future work includes transitioning to Attention-based Transformers, expanding platform compatibility, and incorporating multimedia analysis and language support. Cyberbullying detection in Roman Urdu is focused [20], facing challenges due to its linguistic limitations and diverse structures. They employ advanced preprocessing techniques and a deep learning architecture tailored for Roman Urdu cyberbullying detection. Experimentation led to RNN-LSTM and RNN-BiLSTM models outperforming CNN after 20 epochs. Future directions include developing ensemble models for improved detection of harassment and hate speech, incorporating context-specific features, and addressing morphological variations for enhanced outcomes.

IV. PROPOSED SYSTEM

Our tool would have two sections: the backend and the frontend. The backend of the system uses a labelled dataset of text-based tweets to train various machine learning models, including CNN, Random Forest, and Boost, to ensure efficiency and accuracy. The chosen model is then displayed to users through the frontend. The frontend allows users to enter text from social media platforms, primarily tweets, and the tool predicts whether the text contains a motive of cyberbullying. The tool also displays other cyberbullying for non-categorized text and no cyberbullying for text without such motives. The website also offers the latest cyberbullying news from The Guardian, a complaint page for submitting complaints, and user authentication for security purposes.

V. METHODOLOGY

Implementing CNNs designed to process textual data through multiple layers. The model is trained and evaluated, considering variations in accuracy with alternative models, preprocessing done here is based on the NLP techniques used. Initially tweets, hashtags or any url is cleaned as the project is concerned with text only. Then punctuation marks also need to be removed. Then there is stopword removal, which cleans the majority of the unrequired content in the dataset. Then only keywords to be tokenized for cyberbullying detection exist in the dataset. Each & every word of the text, message or any other web content has to be converted into its base form i.e the indivisible unit known as morphemes A word can be in noun, verb, adjective or adverb form. Reducing it to base form will ease processing of that word for cyberbullying categories or not.

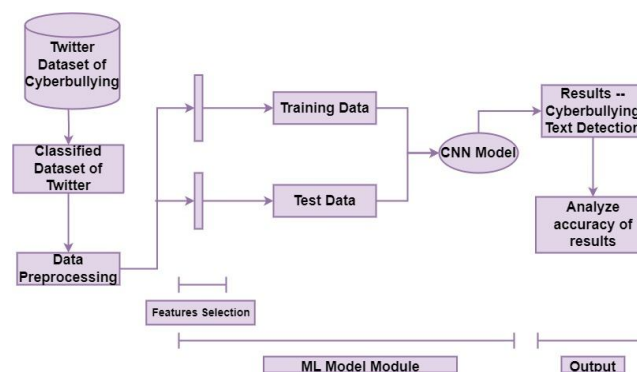


Fig 1: Architecture of Implementation of Model

A. Steps Implemented for Machine Learning Model

- 1) *Importing Libraries*: The implementation begins with importing necessary libraries, creating a foundation for subsequent processes.
- 2) *Importing Data*: Data acquisition is crucial; this step involves loading the dataset for training and evaluation.
- 3) *Exploratory Data Analysis (EDA)*: Understanding the dataset is vital. EDA provides insights into data distribution, trends, and patterns.
- 4) *Preprocessing Tweets with Tweet-preprocessor*: Leveraging the tweet-preprocessor library for cleaning, parsing, and tokenizing tweets. This step involves handling URLs, hashtags, mentions, reserved words, emojis, smileys, and supporting JSON and .txt files.
- 5) *Removing Stop Words*: Stop words, common in language, are removed to focus on meaningful tokens, reducing dataset size and training time.
- 6) *Stopword Removal*: Made function for stopwords for preprocessing and importing necessary libraries.
- 7) *Removing Punctuations and Converting Case*: Employing regular expressions and string functions to eliminate punctuations and convert text to lowercase.
- 8) *Tokenization*: Tokenizing words using NLTK, breaking down text into smaller units for further analysis.
- 9) *Word Cloud*: Implementing word cloud visualisation for a concise representation of textual data.
- 10) *Lemmatization*: Applying lemmatization for accurate grouping of inflected word forms into their base root mode.
- 11) *Counter*: Utilising Counter, a dictionary subclass, for counting hashable objects and obtaining insights into word frequency.
- 12) *Splitting Data into Test and Train*: Performing a train-test split to estimate model performance and ensure robust evaluation.
- 13) *Ordinal Encoder*: Encoding categorical features as integer arrays using ordinal encoding.
- 14) *Training the Model*: Implementing CNNs designed to process textual data through multiple layers. The model is trained and evaluated, considering variations in accuracy with alternative models. Totally we have evaluated 3 models as per the changes of Epochs required for better model experimentation. Result below is the 3rd model evaluation.

B. Dataset

A labelled dataset of nearly 47,000 tweets categorises cyberbullying types:

Religion, Age, Ethnicity , Gender, Another kind of online harassment, Not the kind of cyberbullying
There are about 8000 of each class in the data since it has been balanced.

C. Algorithmic Study

1. CNN (Convolution Neural Networks)

In this study, CNN was employed to model the relationship between tweet_text(independent variable) and cyberbullying_type(dependent variable) in the context of social media investigations against cybercrime. Compared to CNNs for text classification, language models have several differences. Here we discuss [12] general design principles of CNN language models; for a detailed description of specific architectures:

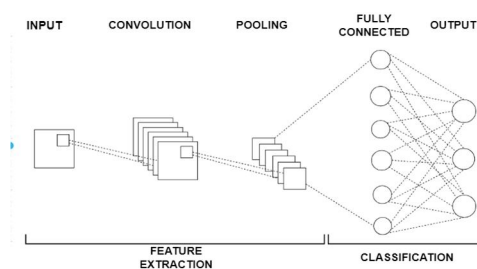


Fig 2 : CNN model evaluation architecture

Input Layer: Here, we give input to our model. The number of neurons in this layer is equal to the total number of features in our data. Trained with the help of Keras.

About Keras: Written in Python, it is a high-level neural network API that may be used with TensorFlow, Microsoft Cognitive Toolkit (CNTK), or Theano. Its development was focused on making quick experiments possible. Keras facilitates quick and simple convolutional neural network (CNN) prototyping.

In the context of CNNs (Convolutional Neural Networks), Keras provides a convenient and user-friendly interface for building, training, and deploying CNN models. It abstracts away many of the complexities involved in designing neural network architectures, allowing researchers and practitioners to focus more on the conceptual aspects of their models rather than low-level implementation details. Keras provides a modular and intuitive API that allows users to easily define the architecture of their CNN models, including layers such as convolutional layers, pooling layers, and fully connected layers. It also includes various activation functions, regularisation techniques, optimizers, and loss functions, making it easy to experiment with different configurations to improve model performance.

- *Convolutional Layers:* This layer extracts the feature from the input dataset given in the input layer. Our model is covered in Conv1D.
 - *Pooling Layers:* They minimise volume, which speeds up calculation, saves memory, and guards against overfitting.
 - *Fully Connected Layers:* They solve the final classification or regression task using the input from the preceding layer.
 - *Layer of Dropout:* During training, Dropout removes neurons from the network at random. As a result, the network is forced to acquire more resilient properties that are independent of any particular neuron.
 - *Activation Functions:* Activation layers give the network nonlinearity by adding this to the output of the layer that came before it.
 - *Output Layer:* Produces final prediction or classification.
- Our Model summarised as:*
- *Embedding:* A technique used to represent categorical data, often used for converting words or tokens into dense vectors.
 - *Convolutional Layer (Conv1D):* A layer that applies convolutional operation on 1D inputs, often used for feature extraction in sequential data.
 - *Max Pooling Layer (MaxPooling1D):* A layer that performs down-sampling by taking the maximum value over a fixed-size window, reducing the dimensionality of the input.
 - *Flatten Layer:* A layer that flattens the input into a one-dimensional tensor, often used to connect convolutional layers to fully connected layers.
 - *Dense Layer:* A fully connected layer where each neuron is connected to every neuron in the previous layer, responsible for learning non-linear transformations in data.

CNNs excel due to their ability to capture hierarchical features through convolution and pooling, reducing the need for manual feature engineering.

When designing a CNN language model, below are points noted[13]:

- 1) To Prevent information flow from future tokens
- 2) A left-to-right LM can only anticipate a token by using prior tokens; make sure your CNN only sees those tokens!
- 3) For language models, location information is crucial, as opposed to text categorization. Consequently, avoid using pooling (or use extreme caution when doing so).
- 4) When stacking many layers, remember to account for remnant connections.

It could be challenging to effectively train an extremely deep network if you stack numerous layers.

Convolutional models can have very long contexts, even though the typical context size for n-gram models is between four and five. Examine the example: a network with a context of seven tokens has just three convolutional layers and a tiny kernel size of three. Can be obtain a very long context length by stacking numerous layers.[14]

VI. RESULTS AND COMPARISON OF MODELS

Cnn Model Confusion Matrix

Test \ Predicted	age	ethnicity	gender	not bullying	religion	other_cyberbullying
age	2268	9	2	118	5	23
ethnicity	20	2243	3	14	5	121
gender	2	8	2013	247	8	122
not bullying	21	2	144	1230	74	923
religion	9	7	105	830	6	1342
other_cyberbullying	6	5	32	161	156	13

Fig 3: Confusion Matrix of Implemented Model

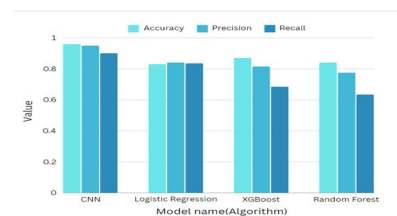


Fig 4: Graphical Representation of Algorithmic studies on our Dataset

A. Experimented Results of Different Machine Learning Model for Same Dataset

Model Comparison

- CNN: 96%, XGBoost: 87%, Random Forest: 84%, Logistic Regression: 83%

B. Comparisons of experimented Accuracies and summaries in evaluated model tests:

- Vocabulary size calculated from dataset of cleaned tweets: 56083
- Analysed top 50 words using tweets.

The convolutional model was analyzed with 35 epochs in 1055/1055-449s, with a loss of 0.0642 and an accuracy of 0.9665. The input layer was None, with 424 sequences. The Embedding layer was None, with 424 sequences. Convolutional layers were Conv1D_3, Conv1D_4, Conv1D_5, Max Pooling layers, and Flattening and Dense Layers.

- *Prediction*

```
new_text = "Black ppl aren't expected to do anything,"
"depended on for anything."
"yet free to participate, work,"
"enjoy the freedom of humans all across this globe."
"If you waste your energy on the past you will be"
"wasting it for building the future."
predicted_type = predict_cyberbullying_type(new_text, model_3, tokenizer)
print(f"Predicted cyberbullying Type: {predicted_type}")

1/1 [=====] - 0s 66ms/step
Predicted Cyberbullying Type: ethnicity
```

Fig. 5: Used as a Test case to predict the sentence

C. User Interface

Interactive Web Application for Cyberbullying News

- Updates on cyberbullying news and text prediction. It helps in filing complaints and tracking cyberbullying events. It includes 6 pages: Home, prediction, contact, about, and Login/Register.
- Techstacks: Uses HTML, CSS, BootStrap, JavascRipt technologies, Firebase for storage and authentication.
- Predicts messages only after successful login.

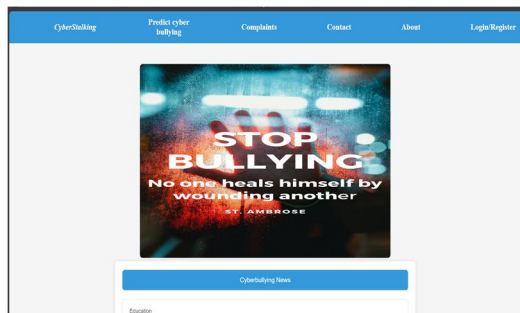


Fig. 6: Bully prevention animated pictures on home page

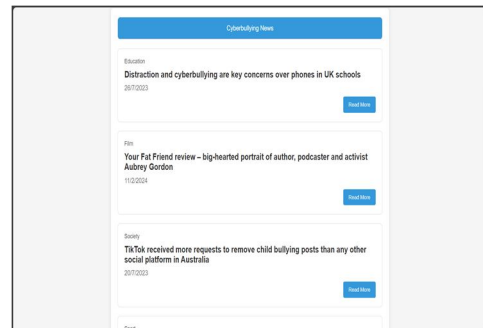


Fig. 7: Cyberbullying news updates on home page, integrated with Guardian News API

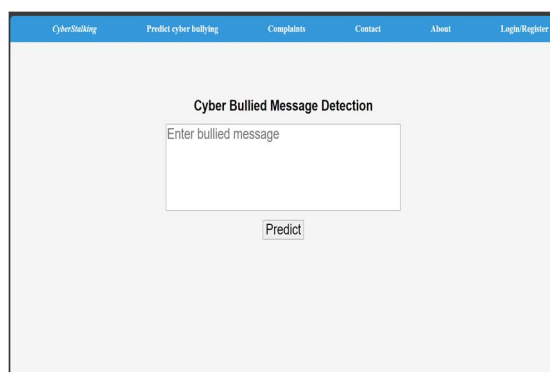


Fig. 8: Prediction page

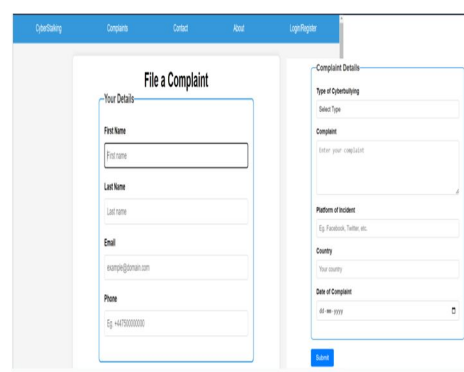


Fig. 9: Complaint Page

VII. CONCLUSION

The tool focuses on detecting cyberbullying by analyzing text input from users, particularly on social media platforms. It uses a CNN-based model to process and analyze the text, recognizing cyberbullying based on various categories such as gender, religion, ethnicity, age, other cyberbullying, or no cyberbullying. The tool is also available as a website, providing the latest news on cyberbullying from The Guardian. Users can submit complaints about cyberbullying content to the website, which can be directed to authorities. The CNN algorithm is more accurate in prediction, and the project can be extended to image, audio, and video datasets. The successful development of the tool will empower investigators, law enforcement agencies, and organizations to proactively detect and respond to cyberbullying, online harassment, and digital harassment issues on social media platforms. The project embodies Sustainable Goal Peace, Justice and Strong Institutions, aiming to create a safer, more inclusive environment through proactive intervention. The product can also be used as a service for social media handlers to predict bullying types and file complaints if found cyberbullied.

REFERENCES

- [1] Helsper, E.J.; Kalmus, V.; Hasebrink, U.; Ságvári, B.; de Haan, J. Country Classification: Opportunities, Risks, Harm and Parental Mediation; EU Kids Online, The London School of Economics and Political Science: London, UK, 2013, DOI: <http://eprints.lse.ac.uk/id/eprint/52023>
- [2] Ibrahim Arpacı & Omer Aslan (2023) Development of a Scale to Measure Cybercrime-Awareness on Social Media, *Journal of Computer Information Systems*, 63:3, 695-705, DOI: 10.1080/08874417.2022.2101160
- [3] R. Armitage, "Bullying in children: Impact on child health," *BMJ paediatrics open*, vol. 5, no. 1, 2021, DOI: 10.1136/bmjpo-2020-000939
- [4] Ross Anderson, Chris Barton, Rainer Böhme, Richard Clayton, Michel J. G. van Eeten, Michael Levi, Tyler Moore & Stefan Savage, *Measuring the Cost of Cybercrime*, 2013, DOI: https://doi.org/10.1007/978-3-642-39498-0_12
- [5] EMERGING TRENDS OF CYBER CRIME IN INDIA: A CONTEMPORARY REVIEW, Tanya Gupta, 2023, DOI: <http://dx.doi.org/10.37253/jlpt.v8i1.7839>
- [6] A Secure Open-Source Intelligence Framework For Cyberbullying Investigation, Sylvia Worlali Azumah, Victor Adewopo, Zag ElSayed, Nelly Elsayed, Murat Ozer. DOI: arXiv:2307.15225v2
- [7] Cyber Bullying Detection on Social Media using Machine Learning, Aditya Desai, Shashank Kalaskar, Omkar Kumbhar, and Rashmi Dhumal, 2021, DOI: <https://doi.org/10.1051/itmconf/20214003038>
- [8] Cyber Bullying Detection for Hindi-English Language Using Machine Learning, Ninad Mehendale, Karan Shah, Chaitanya Phadtare, and Keval Rajpara, 2022, DOI: <http://dx.doi.org/10.2139/ssrn.4116143>
- [9] An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques, Mitushi Raj, Samridhi Singh, Kanishka Solanki, and Ramani Selvanambi, 2022, DOI: <https://doi.org/10.1007/s42979-022-01308-5>
- [10] Session-based cyberbullying detection in social media: A survey, Peiling Yi, Arkaitz Zubiaga, 2023, DOI: <https://doi.org/10.1016/j.osnem.2023.100250>
- [11] Social Media Cyberbullying Detection using Machine Learning, John Hani Mounir, Mohamed Nashaat, Mostafaa Ahmed, and Zeyad Emad, 2019, DOI: Link
- [12] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. "Understanding Convolutional Neural Networks for Text Classification." Affiliations: 1 Computer Science Department, Bar Ilan University, Israel; 2 IBM Research, Haifa, Israel; 3 Intuit, Hod HaSharon, Israel; 4 Allen Institute for Artificial Intelligence, 2020, DOI: 1809.08037.pdf (arxiv.org)
- [13] "Convolutional Neural Network: Text Classification Model for Open Domain Question Answering" by S. M. Kamruzzaman and A. Mustafa 2020 DOI: Link
- [14] "Understanding Convolutional Neural Networks for Text Classification" by Denny Britz, 2018, DOI: <https://aclanthology.org/W18-5408/>
- [15] Convolutional Neural Networks for Sentence Classification by Yoon Kim, 2014, DOI: <https://aclanthology.org/D14>
- [16] R. R. Dalvi, S. Baliram Chavan and A. Halbe, Detecting A Twitter Cyberbullying Using Machine Learning, *ICICCS*, pp. 297–301, doi:10.1109/ICICCS48265.2020.9120893. (2020)
- [17] S. M. Kargutkar and V. Chitre, A Study of Cyberbullying Detection Using Machine Learning Techniques, *ICCMC*, pp. 734–739, doi:10.1109/ICCMC48092.2020.ICCMC-000137. (2020)
- [18] Trana R.E., Gomez C.E., Adler R.F. (2021) Fighting Cyberbullying: An Analysis of Algorithms Used to Detect Harassing Text Found on YouTube. In: Ahram T. (eds) *Advances in Artificial Intelligence, Software and Systems Engineering. AHFE 2020. Advances in Intelligent Systems and Computing*, vol 1213. Springer, Cham. https://doi.org/10.1007/978-3-030-51328-3_2. (2020)
- [19] Mihir Gada, Kaustubh Damania, Smita Sankhe, Cyberbullying Detection using LSTM-CNN architecture and its applications, doi: <https://doi.org/10.1109/ICCCI50826.2021.9402412> (2021)
- [20] Amirita Dewani, Mohsin Ali Memon, Sania Bhatti, Cyberbullying detection: advanced preprocessing techniques & deep learning architecture for Roman Urdu data, doi: <https://doi.org/10.1186/s40537-021-00550-7>(2021)