TACKLING CYBERBULLYING ON SOCIAL MEDIA A MACHINE LEARNING FOR EFFECTIVE DETECTION

Priyatharsini .C [1], Satyendra Kumar [2], Navaneethakrishan K [3], Nishanth K [4], Krishna Raj K [5]

[1] At Mahendra Engineering College, I serve as an Assistant Professor within the Department of Computer Science and Engineering.
[2, 3, 4, 5] Undergraduate students, Department of Computer Science and Engineering

Corresponding Author
Satyendra Kumar,
satyendrakumar4513@gmail.com

DOI
10.29121/shodhkosh.v5.i7.2024.4954

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

Over the past few years, social media and online social networks (OSN) have seen a sharp rise in popularity. However, the main issues with social media sites and online networks are security and privacy. However, attention must be paid to the grave issue of cyberbullying (CB) on social media platforms. An intentional, tenacious, and forceful reaction The phrase "cyberbullying" (CB) refers to behaviors on information and communication technology (ICT) platforms, such as social media, the internet, and mobile devices. Deep learning (DL) methods are needed for the identification and classification of CB in social networks in order to counter this trend. An novel method that combines deep learning and feature subset selection for social networks is called feature subset selection with deep learning based CB detection and categorization (FSSDL-CBDC). Three steps make up the suggested FSSDL-CBDC method: preprocessing, classification, and feature selection. combining a system for automatically filtering social media accounts with a A deep learning method system for automatically identifying and classifying cyberbullying on social media platforms is given in this study. Proactive steps to protect user safety and wellness are becoming more and more crucial due to the prevalence of online abuse and cyberbullying. With cutting-edge techniques for text and audio-visual data analysis on social media, the suggested system attains a 99.983% accuracy rate. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are two categories of deep learning models. All things considered, the experimental findings showed that the FSSDL-CBDC technique outperformed the other options in several areas..

Keywords: IoT, AI, Deep Learning, Sensor

1. INTRODUCTION

The generation of young people today, known as "digital natives," grew up during a time when new technology predominated and communications were nearly instantaneous. Building bonds with people and communities has never been simpler as a consequence. Social networking sites are being used by teenagers more and more, which has made them vulnerable to bullying. Adversarial comments demoralize teens and have a detrimental psychological effect on them.

In this work, we have developed supervised learning methods for cyberbullying detection. Cyberbullying refers to the act of harassing someone via the internet. Although it has always been an issue, knowledge of its consequences for young people has recently increased. By using machine learning, we can develop rules that automatically recognize text that constitutes cyberbullying and pinpoint the language patterns that bullies and their victims use. They also provide as a place for social interaction and the upkeep of existing friendships. On the negative side, social media facilitates the creation of new relationships as well as the maintenance of existing ones.

How to cite this article (APA): C, P., Kumar, S., K, N., K, N., and K, K.R. (2024). Tackling Cyberbullying on Social Media a Machine Learning for Effective Detection. *ShodhKosh: Journal of Visual and Performing Arts, 5*(7), 1082–1089. doi: 10.29121/shodhkosh.v5.i7.2024.4954

Conversely, the negative aspect of social media is that it increases the likelihood that children will encounter risky situations including cyberbullying, depressive symptoms and suicidal thoughts, as well as grooming and inappropriate sexual activity. Using social media, bullies can readily target their victims outside of school premises because users can often remain anonymous and are reachable 24/7. It is common practice to define the definition of cyberbullying and online harassment as a classification issue. Techniques often used for subject identification, emotion analysis, and document classification can also be used to detect cyberbullying by leveraging characteristics of messages, senders, and recipients.It should be emphasized, nevertheless, that recognizing abusive content by itself is not as difficult as recognizing cyberbullying. To classify a message as cyberbullying, additional It might be required to offer proof that it's a part of a bigger online harassment campaign directed at the victim. Cyberbullying and the pervasive use of social media are expanding simultaneously. Cyberbullying victims run a major risk to their physical and emotional health. While there is a scheme in place to identify bullying, there isn't as much activity focused on keeping an eye on social media for indications of cyberbullying. Thus, the primary objective of the proposed system is to detect instances of cyberbullying on social networks through natural language analysis.

2. LITERATURESURVEY

M.Di.Capua,et,al.[1].The four areas that highlight the challenge of how to construct a model of unsupervised internet bullying that includes both conventional textual elements and additional "social features" are sentiment community, semantic, and syntactic features. The author has used the Growing Hierarchical Self Map editing network (GHSOM), which contains an insertion layer with 20 items, 50 neurons, and a 50 x grid. Using the integration algorithm k-means, M. Di Capua and colleagues were able to isolate the GHSOM in the Formspring database from the input database. When this hybrid strategy is utilized unsupervised, the results are better than with the previous method. The author looked through the YouTube database at three p.m.

There are three different machine learning models available: Decision Tree Classifier (C4.5), Naive Bayes Classifier, and Support Vector Machine (SVM) with Linear Kernel. It was demonstrated that the combined effects of hate speech resulted in lower accuracy in the YouTube database as compared to FormSpring trials. This is due to the fact that syntactical features and text analysis function differently on each side. The Twitter Database's hybrid approach produced an F1 Score and a low recall score. The authors' concept can be refined and applied to develop initiatives that mitigate the negative effects of cyberbullying.

J. Yadav et al. [2] , A novel method of detecting cyberbullying on social media platforms is proposed by fusing a single line neural network with a BERT model. The model was tested using the Form spring forum and the Wikipedia database. The suggested approach yielded 98% accuracy in spring form datasets and 96% accuracy in a respectably sized Wikipedia database when performance was compared to earlier models. Because of its size g, the suggested model produced better results from the Wikipedia database with less sampling; however, I Multiple samples are required for the spring data form.

R. R. Dalvi, et al. [3] Show Twitter the benefits of using Classified Supervised Machine Learning Algorithms to detect and stop online exploitation. The live Twitter API is used to collect tweets and data for this investigation. Support Vector Machine and Naive Bayes tests are applied to the gathered data sets by the suggested model. Utilize its TFIDF vectorizer to eliminate a feature. The outcomes demonstrate the accuracy of the Vector Support-based online bullying model. The machine outperforms Naive Bayes about 71.25% of the time, compared to 52.75%.

Trana R.E., et al. [4] Creating a machine learning model that reduces unique events—such as text taken from meme images—was the aim. One of the writers' websites has about 19,000 text views that have been posted to YouTube. This study focuses on the performance of three learning machines: Ignorant Bayes and Help Vector Computers We test the gadget and the convolutional neural network on the YouTube page, and we contrast the outcomes with the capabilities of the most recent version of Form.The sections of the YouTube website that address algorithms related to cyberbullying are rarely highlighted. In the four categories of race, nationality, politics, and general, Naive Bayes outperformed SVM and CNN. While CNN belongs to the same gender group, SVM outperformed the inexperienced Naïve Bayes, and all three algorithms showed comparable accuracy in the middle body group. False data was utilized in the study's conclusions to differentiate between violent and non-violent episodes. If more research is required to determine whether the YouTube website offers a better context for collections connected to violence, it can concentrate on developing a two-part separation system that will be used to assess text extracted from photographs.

N. Tsapatsoulis, et al. [5] an extensive analysis of the latest data on cyberbullying from Twitter. It is highlighted how important it is to recognize the many types of Twitter abusers. According to Kwe paper, a number of actions must be taken in order to develop Internet traffic monitoring software that is successful and profitable. The features, machine learning models, and styles required for data classification and recording platforms for model research are provided by these technologies. The process project on gaining cyberbullying technology through machine reading will begin with this document.

3. METHODOLOGY

A. PROPOSED METHOD

Web technologies and Python will be used in the development of this project. Initially, we look for, locate, and load the dataset in order to train the model. I import the data, preprocess it, and then send it to the Tf-Idf. After that, it trains the data set and builds the model independently using CNN and Naive Bayes techniques. Next, we'll use the FLASK framework to create a web application. It first imports real-time tweets from Twitter, then analyzes the imported tweets using the developed model to determine if any of the text or photos are examples of cyberbullying. We utilize Mysql as the database, Python as the backend, and HTML, CSS, JavaScript, and other markup languages as the frontend.The dangerous circumstances, such as cyberbullying, signs of melancholy and suicide thoughts, grooming or sexually inappropriate behavior. Because users may frequently stay anonymous and are reachable around-the-clock, Bullies can easily target their victims outside of school premises by using social media. It is customary to frame the identification of online harassment and Cyberbullying is categorized as a problem. Techniques often used for subject identification, emotion analysis, and document classification can also be used to detect cyberbullying by leveraging characteristics of messages, senders, and recipients. It should be emphasized, nevertheless, that recognizing abusive content by itself is not as difficult as recognizing cyberbullying. To demonstrate that a single hurtful message is a part of a wider cyberbullying effort aimed at users, more information may be required. The usage of social media and cyberbullying are both on the rise.

B. ARCHITECTURE

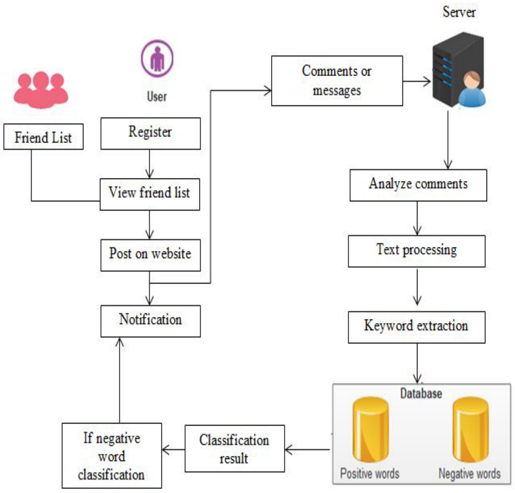


Fig 1: Block diagram

Currently, one of the most popular interactive platforms for sharing, interacting, and exchanging a substantial amount of knowledge about human existence is an online social network, or OSN. Online social networks (OSNs) must allow users to manage the notes put up in their personal space to stop dangerous information from spreading. OSNs now provide some limited support for this requirement. To bridge the gap, in this work we propose a technique that directly grants OSN users control over the statements that are written on their walls. This is made possible by a machine learning-based soft classifier that automatically analyzes messages to enable content-based filtering, and a framework based on rules that users may define to describe the filtering criteria to be applied to their walls. Extracting and selecting a set of identifying and characterization features for an effective Short Text Classifier (STC) is the main objective of the work. The list of phrases that have been categorized and utilized to determine whether or not the list contains any offensive terms is provided below. To remove any inappropriate information, any obscene terms found in the communication will be placed to the Blacklists. Ultimately, a message stating the outcome will be displayed on the user's wall; it will not contain any inappropriate language. Through the use of blacklists, a system automatically filters unsolicited messages based on the relationships and traits of the message producers as well as the message content. Two of the most significant modifications include allowing users to construct Filtering Rules (FRs) and improving the set of features utilized in the categorization process. The semantics of the filtering rules are also changed to better fit the domain under evaluation.

4. SYSTEM IMPLEMENTATION

Reaching your audience through social media marketing is quite effective, particularly when offering specialized goods or services. Research has demonstrated the efficacy of social media in augmenting revenue, elevating brand recognition, fostering brand image development, and increasing the likelihood of a brand going viral. Businesses attempting to connect with consumers on social media have found that machine learning has completely changed the game. As a social media agency, we are aware of the requirements for optimizing social media marketing with artificial intelligence.

Machine learning is being used by social media to both identify and group data. Machines are used by social media technologies to decide which features should be shown to certain users. They collect user data and utilize machine learning to analyze it for potential future uses. Images, videos, audio, test mixes, messages, and likes are all processed by machine learning. However, machine learning uses algorithms in order to collect various types of data and provide it to users so they can complete their duties. Data analysis and machine learning data clustering are essential in social media. As a result, they illustrate the features of uploading photos using their extremely predictable algorithm and graphical data display.

SHORT TEXT CLASSIFIER (STC)

A technique for categorizing brief textual items, such Tweets, Facebook postings, online reviews, and more, is called short text classification. The merger of deep learning, machine learning, and natural language processing (NLP) approaches helps produce meaningful and relevant categories from small quantities of text input. Compared to conventional text classification techniques like word embedding, which groups related words in a data set using a conventional bag-of-words (BOW) model, short text classification is more accurate. Word embeds, for instance, may presume that since apples and bananas are both considered fruits, they belong in the same category.

This approach does not show more complex links between words and their context—it only labels data. In order to understand the semantic connections between different texts, short text classifiers go beyond the conventional method by utilizing sophisticated classification models and tactics including support vector machines, decision trees, and the naive Bayes classifier. Their capacity to extract more context from the text increases the accuracy of the classification.

The goal of short text classification is to get rid of all the obstacles that stand in the way of useful insights from brief texts. With the use of machine learning and data analytics, it creates a classification model that is sensitive to the subtleties of spoken language.

5. RESULT AND DISCUSSION

An sample of bullying-related content that the detection aircraft has determined to be bullying is seen in the screenshot. The results indicate that "this is a type of bullying content." The phrase "This is not a type of bullying content" will appear in the results if the text contains no instances of cyberbullying. When it's done, the Django framework displays "NAME, USERNAME, PASSWORD" for the register option. By choosing the "LOGIN" option, the user can sign in using their current account. Research on cyberbullying has frequently concentrated on identifying specific cyberbullying "attacks," ignoring more subdued forms of the activity as well as posts made by victims and onlookers.

However, these posts might as well prove that cyberbullying occurs. The main contribution of this paper is the method presented for automatically identifying cyberbullying signals on social media, which includes posts by bullies, victims, and bystanders, among other types of cyberbullying. An manually annotated corpus of cyberbullying in English and Dutch was used to evaluate our approach. This allowed us to show how easily, given the availability of annotated data, our method might be translated to other languages.

It can be split into the two steps of learning that are mentioned below. With this strategy, we suggested building an IDS with a deep learning technique based on HCRNNs. If the datasets have all the criteria, our suggested HCRNNIDS is computationally efficient and provides greater accuracy with a low likelihood of a FAR.

A. APPROACH

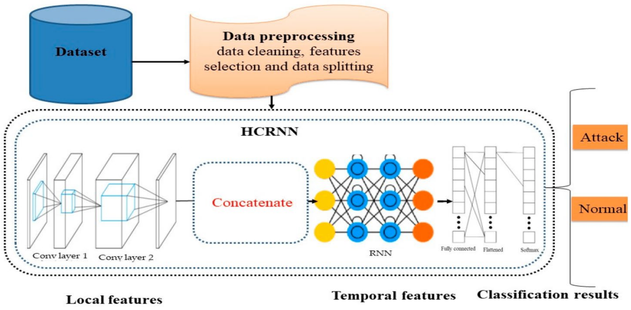


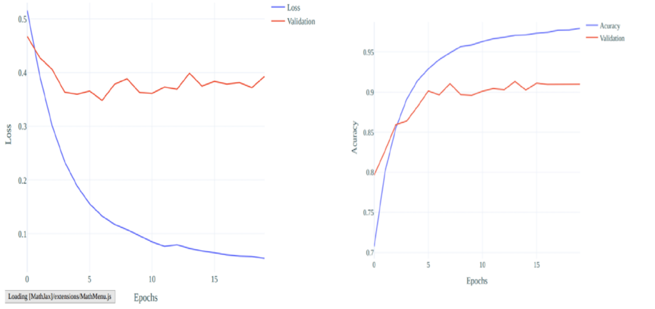
Fig 2: Overview of the HCRNNIDS

A CNN is made up of two basic parts, as the HCRNNIDS overview demonstrates: (i(i) a classifier, and (ii) a feature extractor. The two layers that make up the feature extractor are the pooling and convolution layers. The feature map, or extracted output, serves as the input for the second classification component. In this way, CNN picks up on the local characteristics quite well. The flaw, though, is that it fails to recognize the time dependence of key characteristics. Thus, we added recurrent layers following the CNN layers in order to more robustly capture both spatial and temporal variables. By doing so, we were able to properly handle the vanishing and inflating gradient difficulties, which enhances our capacity to recognize temporal and spatial connections and extract useful information from variable extent sequences. The CNN processes the input in the HCRNN network initially, and the CNN's output is then transmitted via the recurrent layers to produce sequences at every time interval that aid in the modeling of both temporal and spatial aspects. Subsequently, the sequence vector is supplied into a After going through a fully linked layer, the probability distribution across the classes is passed through a soft max layer. The data pre-processing section initially arranged and prepared the network flow. The conversions required for the data formats supported by IDS and HCRNNIDS were part of the pre-processing.

B. EXPERIMENTAL SETTING

This section presents the experiment's setup and explains their significance. Currently, Python 3.13 is used on Google Colab GPUs. Tensorflow, a toolkit for computer vision applications, and natural language processing (NLP) are used to develop the proposed cyberbullying model as well as the other baseline models. Finding the best hyperparameters for the dense layer and removing unnecessary elements, such as the number of hidden nodes, were two ways to make the model simpler. A 35,873 tweet input matrix was created, and the Tensorflow framework's tokenization was used to evaluate the word sequence in order to help the cyberbullying model find relevant information inside the text and comprehend the context.

Before tokenization, a pretreatment step was carried out to get rid of missing and duplicate documents, improper text formats, and text content that was lost. Words that had no significance were eliminated from the text; by doing so, noise levels and feature set dimensions were decreased and they were kept from interfering with text processing for its intended use.



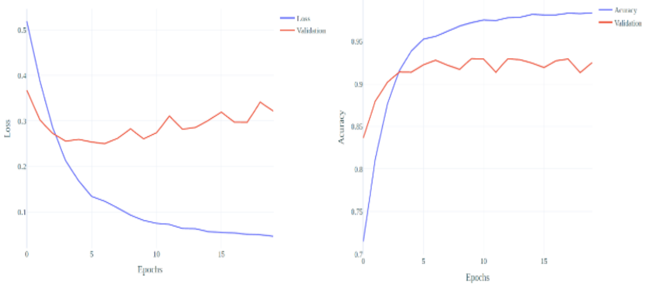
(a) Proposed stacked ensemble learning model loss. (b) Proposed stacked ensemble learning model testing accuracy.

Fig 3: Utilizing the Twitter dataset, test the suggested stacked ensemble model's loss and accuracy

The amount of tweets with more words than others is counted using the CBOW and Word2vec approach skip-gram model architectures. This method makes use of both training models, the basic idea of which is to predict a word's surrounding context depending on whether it is present or absent in the context in which it is being used. Because of the similarities between the CBOW and Skip-gram models, we shall try to present their derivation once more.

Due to the high computational cost of these models, negative sampling and hierarchical SoftMax training techniques were used. In Hierarchical SoftMax, every word is represented by An almost binary-looking frequency-based Huffman tree represents the vocabulary units at the output. A Huffman output layer is used in place of the output layer of the hierarchical SoftMax CBOW model. In order to verify and assess the performance of our suggested stacked ensemble model on additional social media platforms, we also carried out a follow-up experiment based on Facebook and Twitter groups using datasets gathered from the literature. This dataset is based on questionable behavior, which usually manifests as threats, derogatory language, and racist cyberbullying.

The suggested model was trained over the course of 20 epochs using the experimental setup previously mentioned. The model loss in relation to the model validation is shown in Figure 4. Figure 3 displays the stacked model accuracy against model validation. The following subsections provide a description of the baseline models', ensemble stacked model's, baseline BERT's, and modified BERT's experimental results:



(a) Proposed stacked ensemble learning model loss. (b) Proposed stacked ensemble learning model testing accuracy.

Fig 4: Analyze the proposed stacked ensemble model's accuracy and loss using Twitter and Facebook social media datasets.

This study evaluated the effectiveness of a proposed model in distinguishing between cyberbullying and non-cyberbullying using a range of assessment markers. In the process of identifying cyberbullying, three deep learning-based models—the stacking model, BERT, and a modified BERT model—were created. In the scientific community, knowing the evaluation criteria is necessary to comprehend the performance of competing models. Cyberbullying classifiers for social media platforms such as Facebook and Twitter are frequently evaluated using the following criteria: The accuracy of cyberbullying prediction models can be calculated by taking the percentage of correctly identified tweets and dividing it by the total number of tweets. As a result, you can use the following computation.

This method is based on the inverse spectrum pyramid (ISP) decomposition method for image representation, a ground-breaking method of digital picture encoding. The model uses a multi-layer neural network architecture inspired by the human visual system to analyze and represent images. The neural model has the ability to enable reliable and self-adaptive image representation beyond pixel level by simulating the non-classical receptive field and local feedback control circuit of a ganglion cell.

6. CONCLUSION

A method for filtering undesired signals from OSN barriers was proposed in this work. The system uses a machine learning soft classifier to impose customizable content-dependent FRS. The major goal of the work is to extract and select a set of defining and distinguishing criteria for a reliable short text classifier. Additionally, the system's flexibility with regard to filtering possibilities is increased by BL management. This endeavor is a precursor to a larger undertaking. Our initial success with the classification procedure motivates us to carry out additional research to raise the standard of classification. The ML soft classifier is used by this system to weed out unnecessary messages. The use of BL increases the filtration system's adaptability. When deciding whether to add a user to the BL, we will take a more sophisticated technique that takes into account the system design. The system includes tools for classification and a robust rule layer using an An extensible language is used to define Filtering Rules (FRs), which provide users control over what information appears on their walls and what doesn't.

FRs provide a wide range of filtering criteria that can be combined and customized to meet the needs of the user. To be more precise, FRs employ the machine learning classification result, user relationships, and profiles to decide which filtering criterion to apply.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Humera Aqeel, “A Hybrid Classifier of Cyber Bullying Detection in Social Media Platforms”, IEEE Conference on Emerging Research in Electronics, Computer Science and Technology2022

Chetna Sharma, “Cyber-Bullying Detection Via Text Mining and Machine Learning”, IEEE Conference on Computing Communication and Networking Technologies, 2021

Peidong Zhang, “Detect Chinese Cyber Bullying by Analyzing User Behaviors and Language Patterns”, IEEE International Symposium on Autonomous Systems, 2019

Djedjiga Mouheb, “Detection of Offensive Messages in Arabic Social Media Communications”, IEEE Conference on Innovations in Information Technology, 2018

Daphney-Stavroula Zois, “Optimal Online Cyberbullying Detection”, IEEE Conference on Acoustics, Speech and Signal Processing, 2018

B. Wang, G. Chen, L. Fu, L. Song, and X. Wang, “Drimux: Dynamic rumor influence minimization with user experience in social networks,” in Proc. 30th AAAI Int. Conf. Artif. Intell., Feb. 2016.

L. Fu, W. Huang, X. Gan, F. Yang, and X. Wang, “Capacity of wireless networks with social characteristics,” IEEE Trans. Wireless Commun., vol. 15, pp. 1505–1516, Feb. 2016.

D. N. Yang, H. J. Hung, W. C. Lee, and W. Chen, “Maximizing acceptance probability for active friending in online social networks,” in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2013, pp. 713–721.

C. Budak, D. Agrawal, and A. E. Abbadi, “Limiting the spread of misinformation in social networks,” in Proc. 20th Int. Conf. World Wide Web, 2011, pp. 665–674.

X. Rong and Q. Mei, “Diffusion of innovations revisited: From social network to innovation network,” in Proc. 22Nd ACM Int. Conf. Inf. Knowl. Manag., 2013, pp. 499–508.

U. Azam, H. Rizwan and A. Karim, "Exploring data augmentation strategies for hate speech detection in roman urdu", Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 4523-4531, 2022.

U. Azam, H. Rizwan and A. Karim, "Exploring data augmentation strategies for hate speech detection in roman urdu", Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 4523-4531, 2022.

M. Bilal, A. Khan, S. Jan, S. Musa and S. Ali, "Roman urdu hate speech detection using transformer-based model for cyber security applications", Sensors, vol. 23, no. 8, pp. 3909, 2023.

M. M. Khan, K. Shahzad and M. K. Malik, "Hate speech detection in roman urdu", ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 20, no. 1, pp. 1-19, 2021.

M. M. Khan, K. Shahzad and M. K. Malik, "Hate speech detection in roman urdu", ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 20, no. 1, pp. 1-19, 2021.

M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed and M. T. Sadiq, "Automatic detection of offensive language for urdu and roman urdu", IEEE Access, vol. 8, pp. 91213-91226, 2020.

T. Sajid, M. Hassan, M. Ali and R. Gillani, "Roman urdu multi-class offensive text detection using hybrid features and svm", 2020 IEEE 23rd International Multitopic Conference (INMIC), pp. 1-5, 2020.

T. Sajid, M. Hassan, M. Ali and R. Gillani, "Roman urdu multi-class offensive text detection using hybrid features and svm", 2020 IEEE 23rd International Multitopic Conference (INMIC), pp. 1-5, 2020.

BIOGRAPHICAL NOTE

Ms. Priyatharsini C. completed her Master's in Computer Science in 2014. She's interested in Artificial Intelligence and Machine Learning and has published 10 papers in international journals. With 10 years of teaching experience, she's now an Assistant Professor at Mahendra Engineering College in Tamil Nadu, India.

Satyendra Kumar is studying Computer Science and Engineering at Mahendra Engineering College in Tamil Nadu, India. His focus is on Data Analysis and Machine Learning.

Navaneethakrishan K is enrolled in the Bachelor's program for computer science and engineering at Mahendra Engineering College, located in Namakkal District, Tamil Nadu, India. His academic pursuits align with his passion for software development.

Nishanth K is a student at Mahendra Engineering College, Namakkal Dt, Tamil Nadu, India. He is currently studying Bachelor's degree in computer science and engineering. He is interested in software development.

Krishna Raj K is a student at Mahendra Engineering College, Namakkal Dt, Tamil Nadu, India. He is currently studying Bachelor's degree in computer science and engineering. He is interested in software development.