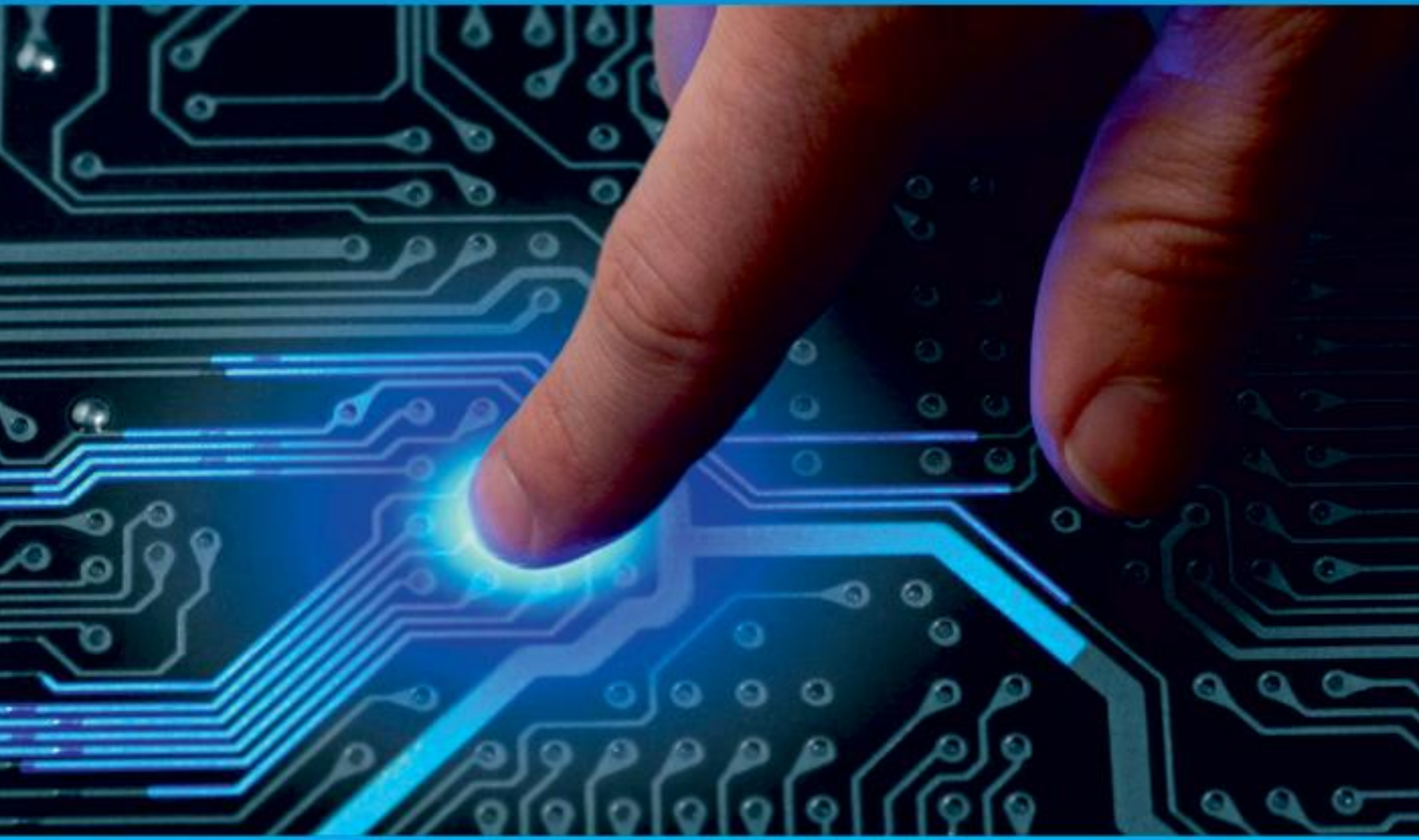
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205333 |

A Novel Approach to Detecting and Assessing Cyber Bullying Severity using Deep Learning Algorithms

Ashokkumar R[1], Akashpradeep M[2], Pradeep V[3], Rahul S[4], Balamurali S[5]

Professor, Department of Computer Science and Engineering, Mahendra Institute of Engineering and Technology, Namakkal, Tamil Nadu, India[1]

Student, Department of Computer Science and Engineering, Mahendra Institute of Engineering and Technology, Namakkal, Tamil Nadu, India [2 3 4 5]

ABSTRACT: Everyone has the right to express themselves freely. But this right is being misused to discriminate against and injure other individuals under the pretence of free speech. Hate speech is the term used to describe this bias. Language that conveys hatred towards an individual or a group of individuals due to characteristics such as race, religion, ethnicity, gender, nationality, disability, or sexual orientation is unmistakably defined as hate speech. It can be expressed verbally, in writing, through gestures, or visually when someone is attacked due of the group to which they belong. Hate speech has been more prevalent both offline and online in recent years.

KEY WORDS: Social media, Hate Speech, Machine learning, Deep learning, Text mining

I. INTRODUCTION

People can communicate easily and broadly using social media, which is more crucial than anything else. Engage in online conversation and freely express your ideas and opinions with others. It is now a necessary component of day-to-day existence. People are particularly vulnerable to abuse or harassment during this phase from those who display hate in a number of contexts, such as politics, sexism, racism, and other forms. Cybertronic, online nuisance, and blackmail are becoming more and more common uses on these social media sites. Thanks to social networking sites (SNS), we may now readily communicate with a range of societies or organizations that interest us.Many technologies have advanced to the point that a significant section of the public can access these websites, such as portable devices and high-speed internet. Handlers in these networks are mostly less than thirty years old. Using the vast amounts of data available on various social networking platforms, researchers have undertaken extensive research in a range of disciplines. Sentiment analysis is a popular academic field that heavily leverages social media data.

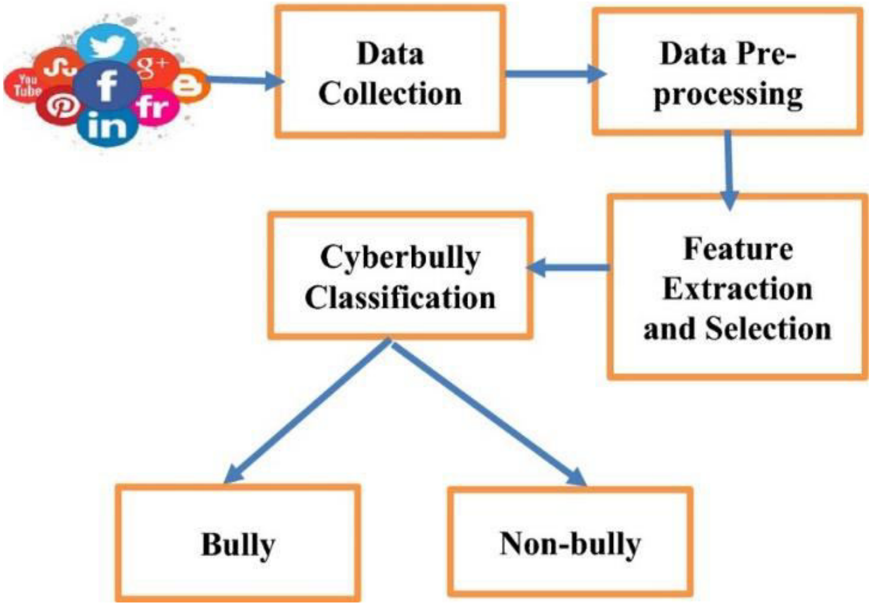


Fig 1: Social media types

II. RELATED WORK

A. A Framework for Hate Speech Detection Using Deep Convolutional Neural Network (IEEE-2020)
In this study, a deep convolutional neural network is used to detect hate speech on Twitter. To find the HS-related tweets on Twitter, machine learning-based classifiers including LR, RF, NB, SVM, DT, GB, and KNN were first employed. The features were retrieved using the tf-idf approach. Nonetheless, using a 3:1 train-test dataset, the best machine learning model, or SVM, could only accurately predict 53% of high school tweets. The unbalanced dataset might be the cause of the poor forecast of HS tweets; as a result, the model was biassed towards the NHS tweets prediction because it included the bulk of the occurrences.

B. Deep Learning Models for Multilingual Hate Speech Detection (arxiv-2020)
We provide the first extensive examination of hate speech in many languages in this research. We employ deep learning models to create classifiers for multilingual hate speech categorization using 16 datasets in 9 languages. We conduct several studies for a range of languages in both monolingual and multilingual situations, with minimal and high resource requirements. Overall, we find that BERT models are more successful for high resource requirements, whereas LASER + LR is more effective for low resource requirements. Finally, we provide a catalogue that

C. Advances in Machine Learning Algorithms for Hate Speech Detection in social media: A Review (IEEE-2021)
The current state of automated hate speech detection on social media was examined in this article. Though study on hate speech as a social issue has long been in the humanities and arts, it is still relatively new in the computing field. In order to keep researchers informed, it is necessary to provide them with updates on any advancements or improvements achieved on a regular basis. We examined methods from deep learning, ensemble, and traditional machine learning techniques. in social online hate speech detection. This study discovered that traditional machine learning (ML) methods for detecting hate speech have received more research attention than ensemble and deep learning methods. This implies that by utilising ensemble and deep learning techniques, researchers may investigate hate speech detection further.

.D. Hate Speech Detection via Multi-Faceted Text Representations (ACM-2021)
In this study, we introduced a unique deep learning framework, called deephate, for automated detection of hate speech using multi-faceted text representations. We conducted extensive testing and found that Deephate beat the state-of-the-art baselines on three publicly available real-world datasets. Additionally, we have conducted an empirical analysis of the Deephate model and have identified the key components that are most helpful in identifying hate speech on online social media platforms. Our examination of prominent features helped to explain Deephate's categorization of hate in some way. Online hate speech is a significant problem that undermines the unity of social media networks and even poses threats to public safety in our towns. Driven by this growing problem, scholars have created several conventional machine learning and deep learning techniques to automatically identify hate speech on online social media sites.

E. Bilstm with deep CNN and hierarchical attention for hate speech detection (ELSEVIER-2022)
In order to detect hate speech, a novel deep learning model called bichat was suggested in this study. It combines a deep convolutional network called bilstm with a hierarchical attention mechanism and BERT-based contextual embedding. In contrast to other models, bichat leverages the capabilities of deep CNN attention mechanisms, context-incorporating embeddings, and bilstm to acquire knowledge about long-range contextual connections and geographical data. The suggested model has been tested on the HD1, HD2, and HD3 benchmark Twitter datasets. The bichat model performed better than the SOTA and baseline approaches, according to the experimental evaluation. In order to determine the effectiveness of the different neural network components included in the suggested model, we also conducted an ablation research.

III. PROPOSED ALGORITHMS

One of the most widely used interactive platforms for sharing, interacting, and exchanging a lot of personal data is the Internet Social Network (OSN).A flexible rule-based system that allows users to customise the filtering criteria applied on their walls and a machine learning-based soft classifier that automatically recognises messages in support of material filtering are used to achieve this. Making use of deep learning (DL) text categorization algorithms, each brief text message is appropriately classified into one or more categories according to its content.

A set of distinguishing and characterizing features must be selected, and this takes up most of the work in constructing a rigorous VADER algorithm.

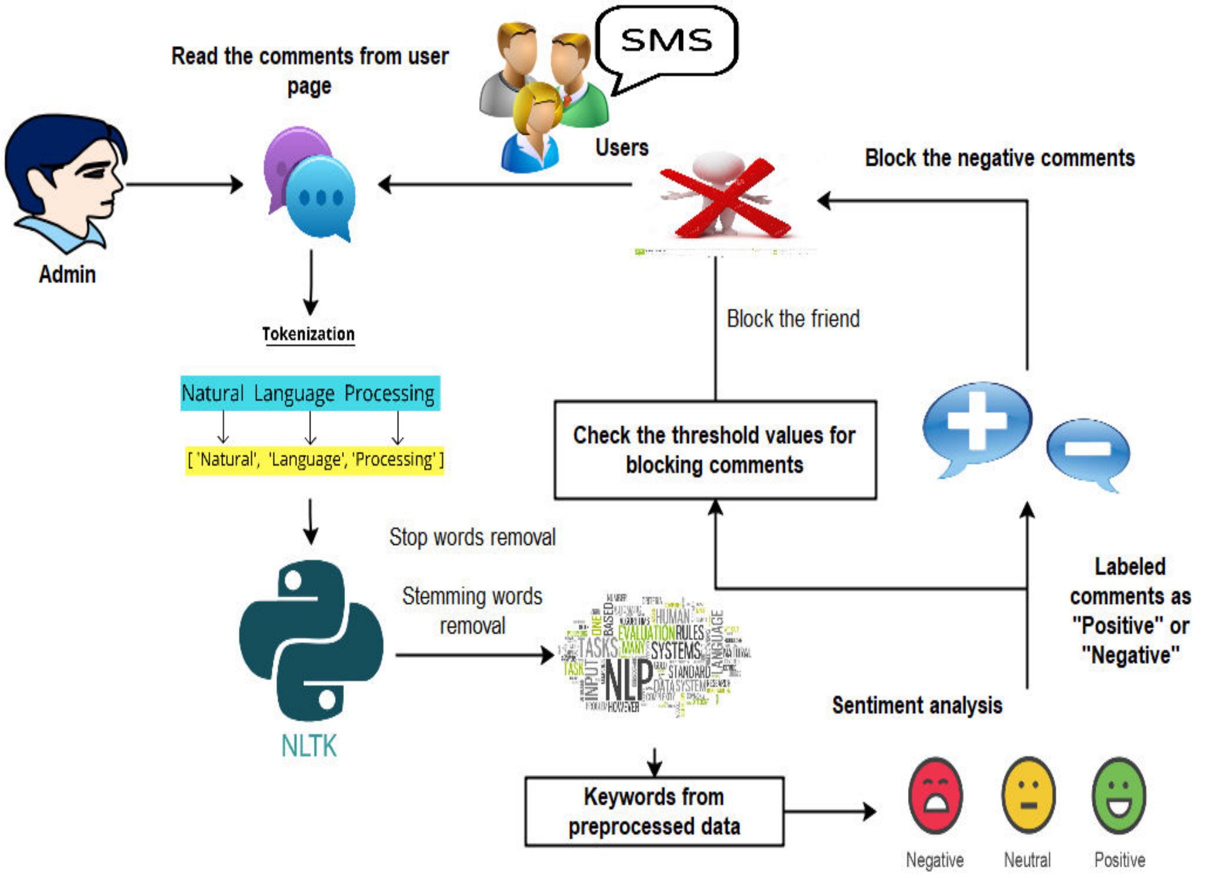

Read the comments from user page
SMS
Users
Block the negative comments
Admin
Tokenization
Natural Language Processing
['Natural', 'Language', 'Processing']
Block the friend
Check the threshold values for blocking comments
Stop words removal
Stemming words removal
NLTK
NLP
Labeled comments as "Positive" or "Negative"
Sentiment analysis
Keywords from preprocessed data
Negative
Neutral
Positive

Fig 2: Proposed Framework

A. FRAMEWORK CONSTRUCTION

A social networking service, sometimes referred to as a social networking site, SNS, or social media, is an online platform that enables users to establish social media accounts or connect with others who have comparable hobbies, pursuits, backgrounds, or connections in real life. A definitional difficulty is presented by the diversity and ever-changing array of hold and built-in social networking platforms in the internet world.Human contact that takes place in virtual communities and networks—where individuals develop, exchange, and/or transmit ideas and information—is referred to as a "social network". Provide a graphical user interface (GUI) that enables visual cues and integrated visuals to be used in user-to-user communication.

B.WORDS EXTRACTION

Social networking is playing a more and bigger role in daily online life as social apps and websites multiply. User comment areas are among the several elements seen in the majority of traditional internet media. Social media platforms are used by businesses for marketing, brand promotion, customer relationship building, and business development. This module allows us to leave comments on an online social network.

C.TEXT MINING ALGORITHM

The majority of efforts aimed at developing a potent deep learning classifier concentrate on the identification and selection of a collection of distinguishing and defining characteristics. The following are the steps of the text mining algorithm:

- Text-based review tokenization as a single phrase
- Examine n-, big-, and unigrams

- Eliminate stop words, examine word stems, and eliminate special characters
- Lastly, take out important words and phrases
- Examine long terms that have appropriate alternatives

D.DEEP LEARNING ALGORITHM

A database of categorised terms is formed here, and the words are then checked for any offensive words. If the user says any vulgar terms, it will be sent to the Watch lists that will eliminate such words.

1. Collect comments or text data with samples of both hateful and non-hateful language.
2. Remove from the text any extraneous characters, symbols, or information.
3. To create the hate speech detection model, use an effective deep learning strategy like VADER Algorithm
4. By providing the model the preprocessed data and modifying its settings depending on the offered examples, you may train the model to recognise hate speech.
5. Integrate the trained model into a programme or system to quickly and accurately identify hate speech.
6. Integrate the trained model into a programme or system to quickly and accurately identify hate speech.

RULES IMPLEMENTATION

Users should be able to specify constraints on text creators through the filtering rules. Thus, creators to whom a filtration rule applies should be chosen based on a variety of criteria, one of which is impinging conditions on the properties of the profile page. This allows us to set guidelines that are exclusive to younger artists, artists who adhere to a certain religion or political philosophy, or artists who, in our opinion, lack expertise in a given field

ALERT SYSTEM

BLs are managed directly by system, which ought to be capable of determining who is embedded in the BL as well as when consumer customer loyalty in the BL is complete. This knowledge is in the framework by a set of guidelines designed to improve flexibility; the rules on BL. The server generates rules for setting thresholds. We can block friends who leave negative comments based on threshold values. Finally, give users mobile notifications.

IV. RESULTS

In this chapter, we can construct the social network using Python as front end and MYSQL as Back end. It is possible to evaluate the system's performance in terms F-measure parameter.

The performance of the system is evaluated using Precision, Recall and F-measure.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{F measure} = 2* \frac{\text{Precision}*\text{Recall}}{\text{Precision+Recall}}$$

The performance evaluation result is shown in fig 3.

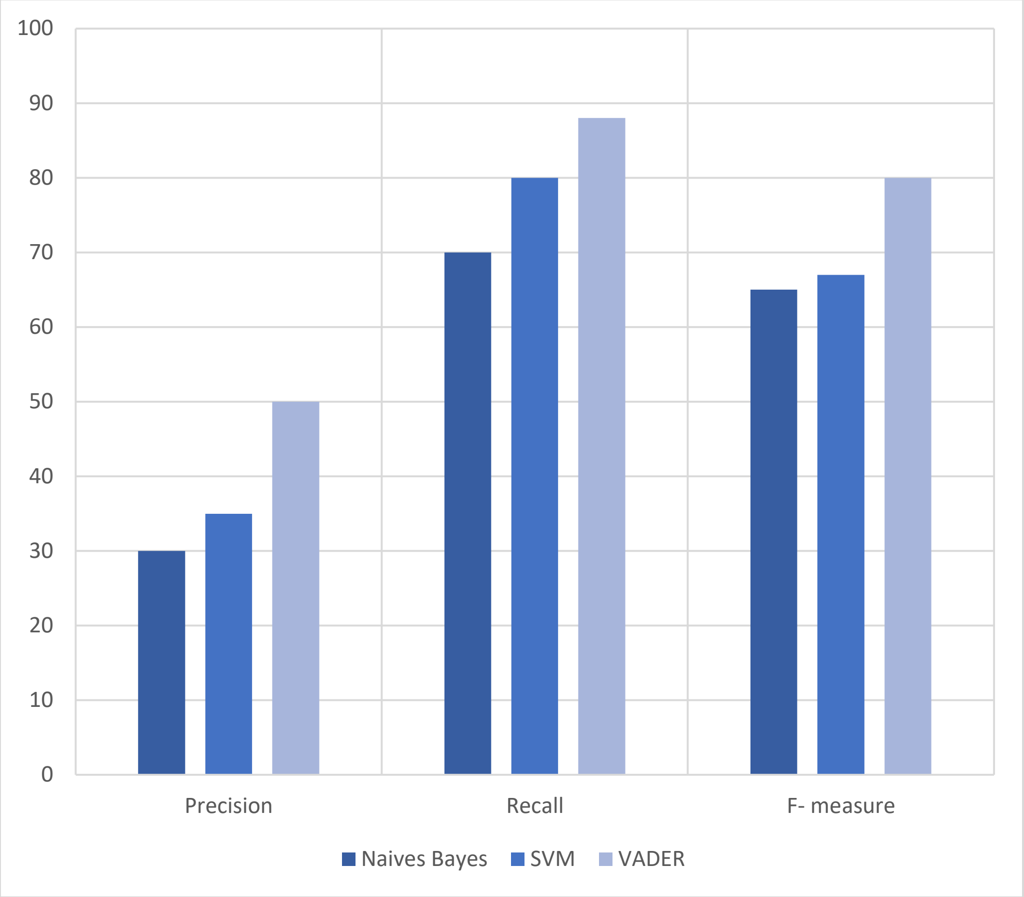


Fig 3: Performance chart

V.CONCLUSION AND FUTURE WORK

In this project, we showed how to filter unsolicited communications from OSN walls. The system enforces a content-dependent filtering rules system—which may be customized—using a DL soft classifier. The most time-consuming part of creating a reliable short text classifier is extracting and choosing a collection of defining and discriminant characteristics. Furthermore, the handling of BLs improves the system's versatility in terms of filtering choices.We plan to use similar strategies to infer BL rules and FRs in the future. We can enhance the framework in the future to implement this approach in a variety of languages with higher accuracy. Also included is the semi-supervised technique to unlabeled data analysis.

REFERENCES

1. Roy, Pradeep Kumar, et al. "A framework for hate speech detection using deep convolutional neural network." IEEE Access 8 (2020): 204951-204962.
2. Aluru, Sai Saketh, et al. "Deep learning models for multilingual hate speech detection." arXiv preprint arXiv:2004.06465 (2020).
3. Mullah, NanlirSallau, and Wan Mohd Nazmee Wan Zainon. "Advances in machine learning algorithms for hate speech detection in social media: a review." IEEE Access 9 (2021): 88364-88376.
4. Cao, Rui, Roy Ka-Wei Lee, and Tuan-Anh Hoang. "DeepHate: Hate speech detection via multi-faceted text representations." Proceedings of the 12th ACM Conference on Web Science. 2020
5. Khan, Shakir, et al. "BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection." Journal of King Saud University-Computer and Information Sciences 34.7 (2022): 4335-4344.
6. Mozafari, Marzieh, Reza Farahbakhsh, and Noel Crespi. "A BERT-based transfer learning approach for hate speech detection in online social media." Complex Networks and Their Applications VIII: Volume 1 Proceedings

of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8. Springer International Publishing, 2020.

7. Rabiul Awal, Md, et al. "AngryBERT: Joint Learning Target and Emotion for Hate Speech Detection." arXiv e-prints (2021): arXiv-2103.
8. Alkomah, Fatimah, and Xiaogang Ma. "A literature review of textual hate speech detection methods and datasets." Information 13.6 (2022): 273.
9. Malik, Jitendra Singh, Guansong Pang, and Anton van den Hengel. "Deep learning for hate speech detection: a comparative study." arXiv preprint arXiv:2202.09517 (2022).
10. Khan, Shakir, et al. "HCovBi-caps: hate speech detection using convolutional and Bi-directional gated recurrent unit with Capsule network." IEEE Access 10 (2022): 7881-7894.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING