Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**

A framework integrating Machine Learning and Fuzzy Logic for the identification of cyber-hate, using(Voting Classifier and Stacking Classifier)

K. JAYA KRISHNA, I. Venkata Abhilash

#1 Associate Professor Department of Master of Computer Applications
#2 Pursuing M.C.A
QIS COLLEGE OF ENGINEERING & TECHNOLOGY
Vengamukkapalem(V), Ongole, Prakasam dist., Andhra Pradesh- 523272

ABSTRACT

Globally, social media have revolutionized how people connect and share information. However, the rise of these platforms has led to the proliferation of cyber hatred, which is a significant concern that has garnered the attention of researchers. To address this issue, We proposes a various solutions, utilizing Machine learning and Deep learning techniques such as Naive Bayes, Logistic Regression, Convolutional Neural Networks, and Recurrent Neural Networks. These methods rely on a mathematical approach to distinguish one class from another. However, when dealing with sentiment-oriented data, a more “critical thinking” perspective is needed for accurate classification, as it provides a more realistic representation of how people interpret online messages. This study applied two machine learning classifiers, Multinomial Naive Bayes and Logistic Regression, to four online hate datasets. The results of the classifiers were optimized using bio-inspired optimization techniques such as Particle Swarm Optimization and Genetic Algorithms, in conjunction with Fuzzy Logic, to gain a deeper understanding of the text in the datasets.

INTRODUCTION

It was the advancement of technology and the impulse of human communication that led to the evolution of social media, which altered how individuals interact online. Prior to the introduction of Information Communication Technology (ICT), human interactions were largely confined to geographical locations; however, Online Social Networks (OSNs) have eliminated geographical barriers.This has prompted researchers to investigate the potential of utilizing Machine Learning and Deep Learning techniques to design automated systems capable of detecting and preventing cyber-hate. Considering the vast amount of content that can be found on OSNs related to aggressive and anti-social behaviour, an Optimized Machine Learning-Based framework is to help identify online hate using fuzzy logic techniques. Several different machine learning models have been implemented, such as, Multinomial Naive Bayes and Logistic Regression, in conjunction with the Bio-Inspired Optimization methods, Genetic Algorithm and Particle Swarm Optimization. The implementation of Particle swarm Optimization selects the best feature selection subset that better represents the feature selection space.

We proposes a various solution, utilizing Machine learning and Deep learning techniques such as Naive Bayes, Logistic Regression, Convolutional Neural Networks, and Recurrent Neural Networks. These methods rely on a mathematical approach to distinguish one class from another. However, when dealing with sentiment-oriented data, a more “critical thinking” perspective is needed for accurate classification, as it provides a more realistic representation of how people interpret online messages. This study applied two machine learning classifiers, Multinomial Naive Bayes and Logistic Regression, to four online hate datasets. The results of the classifiers were optimized using bio-inspired optimization techniques such as Particle Swarm Optimization and Genetic Algorithms, in conjunction with Fuzzy Logic, to gain a deeper understanding of the text in the datasets. The proliferation of cyber hatred on global social media platforms poses a pressing issue, necessitating effective solutions. Existing approaches, relying on machine learning and deep learning techniques, lack a nuanced perspective when dealing with sentiment-oriented data. This study aims to enhance the accuracy of online hate classification by introducing a "critical thinking" approach and optimizing results using bio-inspired techniques like Particle Swarm Optimization and Genetic Algorithms, coupled with Fuzzy Logic, to gain deeper insights into the textual content of hate datasets.

LITERATURE SURVEY

3.1 Social media cyberbullying detection using machine learning:

https://www.researchgate.net/publication/333506989_Social_Media_Cyberbullying_Detection_using_Machine_Learning

ABSTRACT: With the exponential increase of social media users, cyberbullying has been emerged as a form of bullying through electronic messages. Social networks provides a rich environment for bullies to uses these networks as vulnerable to attacks against victims. Given the consequences of cyberbullying on victims, it is necessary to find suitable actions to detect and prevent it. Machine learning can be helpful to detect language patterns of the bullies and hence can generate a model to automatically detect cyberbullying actions. This paper proposes a supervised machine learning approach for detecting and preventing cyberbullying. Several classifiers are used to train and recognize bullying actions. The evaluation of the proposed approach on cyberbullying dataset shows that Neural Network performs better and achieves accuracy of 92.8% and SVM achieves 90.3. Also, NN outperforms other classifiers of similar work on the same dataset.

3.2 Modeling the detection of textual cyberbullying:

https://ojs.aaai.org/index.php/ICWSM/article/view/14209

ABSTRACT: The scourge of cyberbullying has assumed alarming proportions with an ever-increasing number of adolescents admitting to having dealt with it either as a victim or as a bystander. Anonymity and the lack of meaningful supervision in the electronic medium are two factors that have exacerbated this social menace. Comments or posts involving sensitive topics that are personal to an individual are more likely to be internalized by a victim, often resulting in tragic outcomes. We decompose the overall detection problem into detection of sensitive topics, lending itself into text classification sub-problems. We experiment with a corpus of 4500 YouTube comments, applying a range of binary and multiclass classifiers. We find that binary classifiers for individual labels outperform multiclass classifiers. Our findings show that the detection of textual cyberbullying can be tackled by building individual topic-sensitive classifiers.

3.3 Detecting cyberbullying: Query terms and techniques:

https://www.researchgate.net/publication/262238159_Detecting_cyberbullying_Query_terms_and_techniques

ABSTRACT: In this paper we describe a close analysis of the language used in cyberbullying. We take as our corpus a collection of posts from Formspring.me. Formspring.me is a social networking site where users can ask questions of other users. It appeals primarily to teens and young adults and the cyberbullying content on the site is dense; between 7% and 14% of the posts we have analyzed contain cyberbullying content. The results presented in this article are two-fold. Our first experiments were designed to develop an understanding of both the specific words that are used by cyberbullies, and the context surrounding these words. We have identified the most commonly used cyberbullying terms, and have developed queries that can be used to detect cyberbullying content. Five of our queries achieve an average precision of 91.25% at rank 100. In our second set of experiments we extended this work by using a supervised machine learning approach for detecting cyberbullying. The machine learning experiments identify additional terms that are consistent with cyberbullying content, and identified an additional querying technique that was able to accurately assign scores to posts from Formspring.me. The posts with the highest scores are shown to have a high density of cyberbullying content.

3.4 Improved cyberbullying detection using gender information:

https://www.researchgate.net/publication/230701861_Improved_Cyberbullying_Detection_Using_Gender_Information

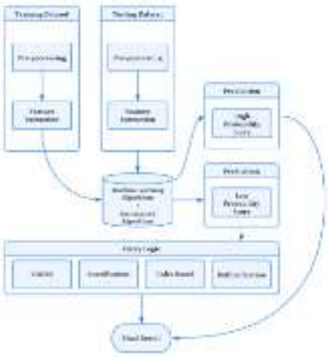
ABSTRACT: As a result of the invention of social networks, friendships, relationships and social communication are all undergoing changes and new definitions seem to be applicable. One may have hundreds of "friends" without even seeing their faces. Meanwhile, alongside this transition there is increasing evidence that online social applications are used by children and adolescents for bullying. State-of-the-art studies in cyberbullying detection have mainly focused on the content of the conversations while largely ignoring the characteristics of the actors involved in cyberbullying. Social studies on cyberbullying reveal that the written language used by a harasser varies with the author"s features including gender. In this study we used a support vector machine model to train a gender-specific text classifier. We demonstrated that taking gender-specific language features into account improves the discrimination capacity of a classifier to detect cyberbullying.

3.5 Towards user modelling in the combat against cyberbullying:

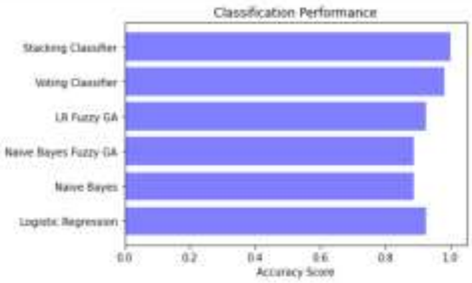
https://www.researchgate.net/publication/230701870_Towards_User_Modelling_in_the_Combat_Against_Cyberbullying

ABSTRACT: Friendships, relationships and social communications have all gone to a new level with new definitions as a result of the invention of online social networks. Meanwhile, alongside this transition there is increasing evidence that online social applications have been used by children and adoles-cents for bullying. State-of-the-art studies in cyberbullying detection have mainly focused on the content of the conversations while largely ignoring the users involved in cyberbullying. We hypothesis that incorporation of the users' profile, their characteristics, and post-harassing behaviour, for instance, posting a new status in another social network as a reaction to their bullying experience, will improve the accuracy of cyberbullying detection. Cross-system analyses of the users' behaviour -monitoring users' reactions in different online environments -can facilitate this process and could lead to more accurate detection of cyberbullying. This paper outlines the framework for this faceted approach.

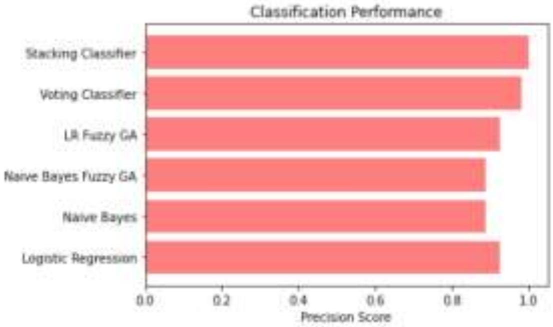
SYSTEM ARCHITECTURE:



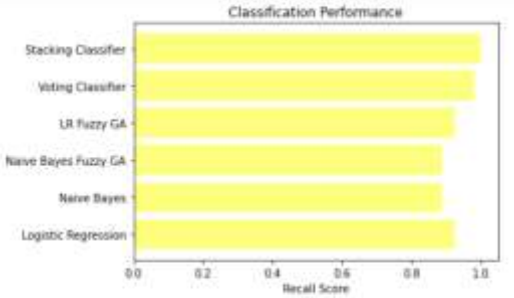
SCREENS:



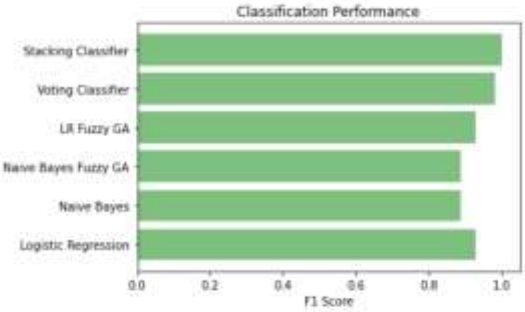
ACCURACY COMPARISION GRAPH



PRECISION COMPARISION GRAPH

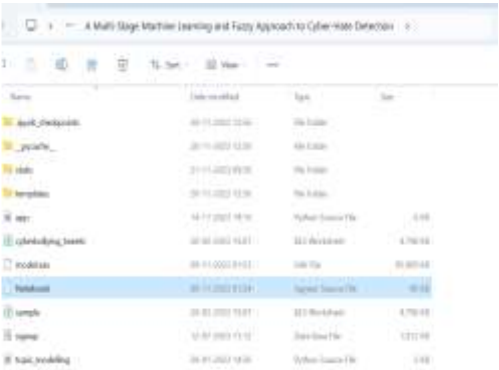


RECALL COMPARISION GRAPH



F1 SCORE COMPARISION GRAPH

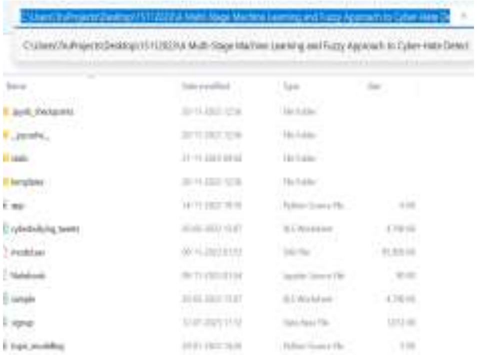
STEP 1



A Multi-Stage Machine Learning and Fuzzy Approach to Cyber-Hate Detection

Name	Date modified	Type	Size
.ipynb_checkpoints	[illegible]	File folder	
__pycache__	[illegible]	File folder	
static	[illegible]	File folder	
templates	[illegible]	File folder	
app	[illegible]	Python Source File	[illegible]
cyberbullying_tweets	[illegible]	XLS Worksheet	[illegible]
model.sav	[illegible]	SAV File	[illegible]
Notebook	[illegible]	Jupyter Source File	[illegible]
sample	[illegible]	XLS Worksheet	[illegible]
signup	[illegible]	Data Base File	[illegible]
topic_modelling	[illegible]	Python Source File	[illegible]

STEP 2

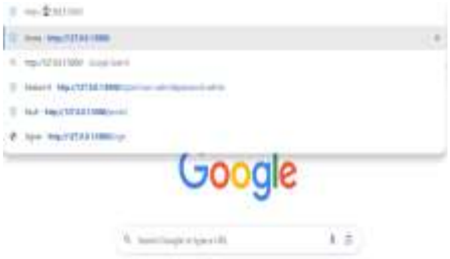


C:\Users\TruProjects\Desktop\15112023\A Multi-Stage Machine Learning and Fuzzy Approach to Cyber-Hate De

C:\Users\TruProjects\Desktop\15112023\A Multi-Stage Machine Learning and Fuzzy Approach to Cyber-Hate Detect

Name	Date modified	Type	Size
.ipynb_checkpoints	[illegible]	File folder	
__pycache__	[illegible]	File folder	
static	[illegible]	File folder	
templates	[illegible]	File folder	
app	[illegible]	Python Source File	[illegible]
cyberbullying_tweets	[illegible]	XLS Worksheet	[illegible]
model.sav	[illegible]	SAV File	[illegible]
Notebook	[illegible]	Jupyter Source File	[illegible]
sample	[illegible]	XLS Worksheet	[illegible]
signup	[illegible]	Data Base File	[illegible]
topic_modelling	[illegible]	Python Source File	[illegible]

STEP 3

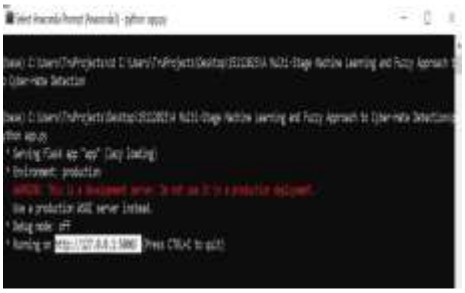


Home - http://127.0.0.1:5000

http://127.0.0.1:5000 - Google Search

Search Google or type a URL

STEP 4

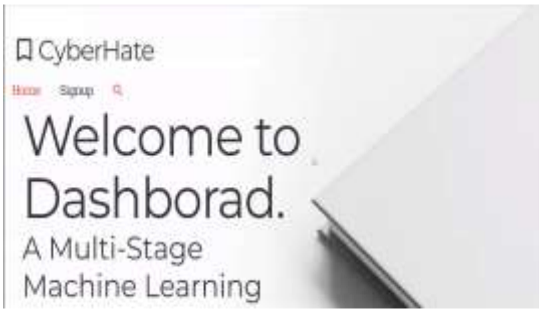


Select Anaconda Prompt (Anaconda3) - python app.py

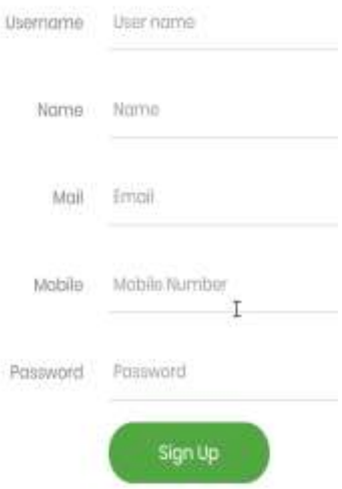
```
(base) C:\Users\TruProjects>cd C:\Users\TruProjects\Desktop\15112023\A Multi-Stage Machine Learning and Fuzzy Approach t
o Cyber-Hate Detection

(base) C:\Users\TruProjects\Desktop\15112023\A Multi-Stage Machine Learning and Fuzzy Approach to Cyber-Hate Detection>p
ython app.py
 * Serving Flask app "app" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: off
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

STEP 5

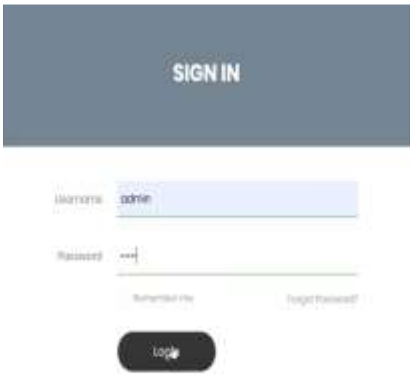


STEP 6



Username	User name
Name	Name
Mail	Email
Mobile	Mobile Number
Password	Password

STEP 7

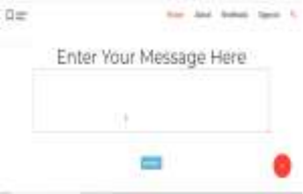


SIGN IN

Username admin

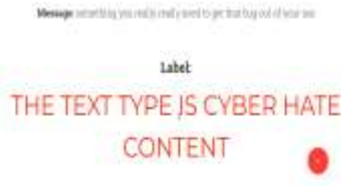
Password ••••

STEP 8



Enter Your Message Here

STEP 9

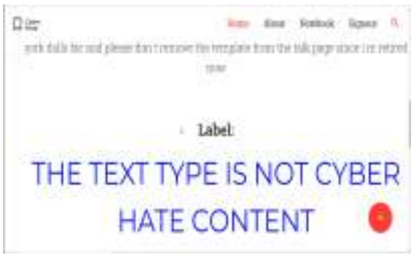


Message: [illegible]

Label:

THE TEXT TYPE IS CYBER HATE CONTENT

STEP 10



[illegible] please don't remove the template from the talk page since i'm retired [illegible]

Label:

THE TEXT TYPE IS NOT CYBER HATE CONTENT

CONCLUSION

In this work, We proposes an optimized machine learning - fuzzy logic approach for identifying hate speech in social media posts. The novelty of the approach lies in the incorporation of bio-inspired optimization techniques along with fuzzy logic to facilitate a deeper understanding of the linguistic aspects of the text. The proposed approach offers several advantages, such as the reduction of data dimensionality resulting from the implementation of optimization, which accelerates the classification process. Additionally, applying fuzzy logic resolves linguistic issues and provides a better understanding of text sentiment. Both GA and PSO are evolutionary search methods that refine values over time using probabilistic and deterministic rules to improve them over time. We combines these two optimization models with fuzzy logic independently on four publicly available datasets: Maryland, Davidson, Formspring, and OLID. Compared to two state-of-the-art supervised machine learning classifiers, such as Logistic Regression and Multinomial Naive Bayes, the optimized fuzzy rule-based method consistently outperforms them with regard to accuracy and F1 scores. Future work will examine General Adversarial Networks (GANs), a deep generative reinforcement learning model that addresses the challenge of imbalance by augmenting the dataset with hateful tweets. This will be done by employing a two-component framework: a generator network and a discriminator network.

REFERENCES

[1] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, "Social media cyberbullying detection using machine learning," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 5, pp. 703–707, 2019.

[2] B. Vidgen, E. Burden, and H. Margetts, "Social media cyberbullying detection using machine learning," Alan Turing Inst., London, U.K. Tech. Rep, Feb. 2022. [Online]. Available: https://www.ofcom.org.uk/__ data/assets/pdf_file/0022/216490/alan-turing-institute-reportunderstanding-online-hate.pdf

[3] 4.4.1 A Sampling of Cyberbullying Laws Around the World. Accessed: Nov. 1, 2023. [Online]. Available: https://socialna-akademija.si/joining forces/4-4-1-a-sampling-of-cyber-bullying-laws-around-the-world/

[4] The EU code of Conduct on Countering Illegal Hate Speech Online. Accessed: Nov. 1, 2022. [Online]. Available: https://commission. europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/ combatting-discrimination/racism-and-xenophobia/eu-code-conductcountering-illegal-hate-speech-online_en

[5] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in Proc. Int. AAAI Conf. Web Social Media, vol. 5, no. 3, Barcelona, Spain, 2011, pp. 11–17.

[6] A. Kontostathis, K. Reynolds, A. Garron, and L. Edwards, "Detecting cyberbullying: Query terms and

techniques," in Proc. 5th Annu. ACM Web Sci. Conf., May 2013, pp. 195–204.

[7] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," in Proc. Content Anal. Web, Madrid, Spain, 2009, pp. 1–7.

[8] M. Dadvar, F. D. Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in Proc. 25th Dutch-Belgian Inf. Retr. Workshop, Ghent, Belgium, 2012, pp. 1–3.

[9] M. Dadvar, R. Ordelman, F. De Jong, and D. Trieschnigg, "Towards user modelling in the combat against cyberbullying," in Proc. 17th Int. Conf. Appl. Natural Lang. Process. Inf. Syst., 2012, pp. 277–283.

[10] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops, Honolulu, HI, USA, Dec. 2011, pp. 241–244.

[11] H. Hosseinmardi, S. A. Mattson, R. Rafiq, R. Han, Q. Lv, and S. Mishra, "Poster: Detection of cyberbullying in a mobile social network: Systems issues," in Proc. 13th Annu. Int. Conf. Mobile Syst., Appl., Services, May 2015, p. 481.

[12] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, "Mean birds: Detecting aggression and bullying on Twitter," in Proc. ACM Web Sci. Conf., New York, NY, USA, Jun. 2017, pp. 13–22.

[13] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," Comput. Hum. Behav., vol. 63, pp. 433–443, Oct. 2016.

[14] V. S. Babar and R. Ade, "A review on imbalanced learning methods," Int. J. Comput. Appl., vol. 975, no. 2, pp. 23–27, 2015.

[15] N. Aggrawal, "Detection of offensive tweets: A comparative study," Comput. Rev. J., vol. 1, no. 1, pp. 75–89, 2018.

[16] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi,"The social world of content abusers in community question answering,"in Proc. 24th Int. Conf. World Wide Web, Florence, Italy, May 2015, pp. 570–580.

[17] P. Fortuna, ''Automatic detection of hate speech in text: An overview of the topic and dataset annotation with hierarchical classes,'' M.S. thesis, Dept. Engenharia, Univ. Porto, Porto, Portugal, 2017.

[18] S. O. Sood, J. Antin, and E. Churchill, ''Using crowdsourcing to improve profanity detection,'' in Proc. AAAI Spring Symp., Stanford, CA, USA, 2012, pp. 69–74.

[19] R. Zhao, A. Zhou, and K. Mao, ''Automatic detection of cyberbullying on social networks based on bullying features,'' in Proc. 17th Int. Conf. Distrib. Comput. Netw., Jan. 2016, Art. no. 43.

[20] V. Nahar, S. Unankard, X. Li, and C. Pang, ''Sentiment analysis for effective detection of cyber bullying,'' in Proc. Asia–Pacific Web Conf., 2012, pp. 767–774.

[21] V. Nahar, X. Li, and C. Pang, ''An effective approach for cyberbullying detection,'' Commun. Inf. Sci. Manage. Eng., vol. 3, no. 5, p. 238, 2013.

[22] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, ''Pre-diction of cyberbullying incidents in a media-based social network,'' in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), Aug. 2016, pp. 186–192.

[23] R. Duwairi, A. Hayajneh, and M. Quwaider, ''A deep learning framework for automatic detection of hate speech embedded in Arabic tweets,'' Ara-bian J. Sci. Eng., vol. 46, no. 4, pp. 4001–4014, Apr. 2021.

[24] A. Al-Hassan and H. Al-Dossari, ''Detection of hate speech in Arabic tweets using deep learning,'' Multimedia Syst., vol. 28, no. 6, pp. 1963–1974, Dec. 2022.

[25] G. Rizos, K. Hemker, and B. Schuller, ''Augment to prevent: Short text data augmentation in deep learning for hate-speech classification,'' in Proc. 28th ACM Int. Conf. Inf. Knowl. Manage., Beijing, China, Nov. 2019, pp. 991–1000.

Mr. K. Jaya Krishna is an Associate Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He earned his Master of Computer Applications (MCA) from Anna University, Chennai, and his M.Tech in Computer Science and Engineering (CSE) from Jawaharlal Nehru Technological University, Kakinada (JNTUK). With a strong research background, he has authored and co-authored over 90 research papers published in reputed peer-reviewed Scopus-indexed journals. He has also actively presented his work at various national and international conferences, with several of his publications appearing in IEEE-indexed proceedings. His research interests include Machine Learning, Artificial Intelligence, Cloud Computing, and Programming Languages. He is committed to advancing research and fostering innovation while mentoring students to excel in both academic and professional pursuits

.

Mr. I Venkata Abhilash has received his MCA (Masters of Computer Applications) from QIS college of Engineering and Technology Vengamukkapalem(V), Ongole, Prakasam dist., Andhra Pradesh-523272 affiliated to JNTUK in 2023-2025