IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 1, 2025

Optimized Cyber-Hate Detection With Machine Learning Classifier with Fuzzy Logic

[1]N. Naga Tanuja,[2]K. Pavan Kumar,[3]P.S. Aditya Reddi,[4]A. Sri Harsha Vardhan,[5]R. Surya Uttin Kumar

[1]Department of CSE(CS), ***Email:*** *nagatanujayadav2003@gmail.com*
*[2]Associate Professor****,*** *Department of CSE,* ***Email:*** *pavankumar.k@raghuenggcollege.in*
[3] Department of CSE(CS), ***Email:*** *adityareddi200324@gmail.com*
[4] Department of CSE(CS), ***Email:*** *harshavardhanallanki365@gmail.com*
[5] Department of CSE(CS), ***Email:*** *reddi.surya8103@gmail.com*

Raghu Institute Of Technology(A), Dakamarri, Visakhapatnam, India.

Abstract: The project's main focus is on tackling the alarming problem of cyber-hatred, which has greatly increased as social media platforms have become more widely used. It recognises how urgent and significant it is to address this issue in the context of the digital world. The initiative suggests using a variety of machine learning and deep learning strategies to counteract cyber-hatred. These consist of recurrent neural networks (RNNs), convolutional neural networks (CNNs), logistic regression, and Naive Bayes. Each of these techniques probably has a distinct function in recognising, categorising, or examining trends in hate speech or objectionable material. Using hate speech data, the study applies two classifiers and optimises their performance through the use of genetic algorithms and particle swarm optimisation. These optimisation methods are probably used to increase the classifiers' accuracy in identifying instances of cyberhatred. Furthermore, by taking into consideration the intricacies and subtleties of text material, fuzzy logic is intended to improve understanding. The main objective is to provide a more practical and efficient method for detecting cyberhate. This entails applying a critical thinking viewpoint, which probably entails taking into account subtleties and contextual clues in addition to specific words or phrases. Additionally, the use of fuzzy logic-based systems and optimisation approaches aims to develop a more nuanced understanding of hate speech, improving detection accuracy and bringing it into line with the complexity of the actual world. By using sophisticated ensemble techniques—more especially, a Voting Classifier and a Stacking Classifier—the project expands its potential. The Stacking Classifier's remarkable 100% accuracy shows how reliable it is in spotting instances of cyber hatred. Using these ensemble models improves the cyber-hate detection system's overall efficacy.

Index terms - *Cyberbullying, fuzzy logic, logistic regression, multinomial Naive Bayes, PSO, VADER.*

1. INTRODUCTION

Social media evolved as a result of technological advancements and human communication impulses, changing the way people connect online. Human interactions were mostly limited to physical locales before the advent of information and communication technology (ICT); now, online social networks, or OSNs, have removed these restrictions [1].

The prevalence of user-friendly technology has made it more and more clear that cyber-hatred is a prevalent problem. Social media is a perilous and illusive phenomena as it has become a forum for the perpetration of abuse and aggression. Teenagers are particularly susceptible to online harassment because of how simple it is for offenders to carry out damaging activities using a laptop or mobile device that is online. Manually flagging data is a traditional way of identifying cybercrime [2], but it has been shown to be neither "effective nor scalable." [2]. This has led academics to look into the possibility of using Deep Learning and Machine Learning approaches to create automated systems that can identify and stop cyber-hatred.

The study suggests an Optimised Machine Learning-Based framework to assist in identifying online hatred using fuzzy logic approaches, taking into account the abundance of information on OSNs pertaining to aggressive and anti-social conduct [35,36]. In combination with the Bio-Inspired Optimisation techniques of Genetic Algorithm and Particle Swarm Optimisation, a number of machine learning models have been used, including Multinomial Naive Bayes and Logistic Regression [30,31,48]. The optimal feature selection subset that more accurately depicts the feature selection space is chosen by the particle swarm optimisation algorithm. The goal is to improve the classification process's accuracy within a data set by reducing the number of duplicated and irrelevant characteristics. Furthermore, PSO makes the learnt model easier to understand. Additionally, the Genetic Algorithm (GA) was used to maximise classifier performance. A certain level of confidence that a variety of solutions are assessed is offered by the GA's random mutation feature. Additionally, the fuzziness of both positive and negative ratings may be included into the application of fuzzy rules. Systems based on fuzzy logic were used to address ambiguity and vagueness. The fuzzy approach's benefits may be summed up as follows: i) It offers a desirable solution to language challenges; ii) It addresses reasoning and delivers perspectives that are closer to the precise sentiment values.

2. LITERATURE SURVEY

TITLE	AUTHOR	METHODOLOGY	PROPOSED SYSTEM	CONS	CONCLUSION
TITLE- Detection of hate speech in Arabic tweets using deep learning	A. Al-Hassan and H. Al-Dossari, LINK- https://link.springer.com/article/10.1007/s00530-020-00742-w	The research employs deep learning models like LTSM, CNN + LTSM, GRU, and CNN + GRU, comparing them against the SVM baseline for identifying and classifying Arabic tweets into categories: none, religious, racial,	The study focuses on utilizing deep learning architectures to enhance hate speech detection in Arabic tweets, aiming to categorize tweets accurately into distinct classes, providing a better	The research might lack in-depth analysis of feature engineering methods specific to Arabic language nuances, and it might not explore potential biases in the	CNN-LTSM and CNN-GRU outperform SVM in detecting hate speech in Arabic tweets, highlighting potential for combating such content on Twitter.

		sexism, or general hate. A labeled dataset of 11K tweets is used.	understanding of hate speech prevalence on Twitter.	dataset that could impact model performance and generalization.	
TITLE-A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets	R. Duwairi, A. Hayajneh, and M. Quwaider LINK-https://link.springer.com/article/10.1007/s13369-021-05383-3	Evaluated CNN, CNN-LSTM, and BiLSTM-CNN on ArHS and combined Arabic datasets, achieving high accuracies in binary, ternary, and multi-class hate speech classification.	Utilizes deep learning networks trained on Arabic hate speech datasets to distinguish hateful content within tweets effectively.	Limited focus on comparative analysis across models and potential biases in dataset annotations, impacting generalizability.	Successful classification via CNN, BiLSTM-CNN in Arabic hate speech detection, emphasizing applicability in social media monitoring and moderation.
TITLE-Prediction of cyberbullying incidents in a media-based social network	LINK-https://ieeexplore.ieee.org/document/7752233	Utilized Instagram data to predict cyberbullying incidents before occurrence. Extracted features from initial posts, including text caption, image content, social graph parameters, and temporal behavior, forming the basis for automated prediction	Developed a predictive model focusing on pre-empting cyberbullying on Instagram by analyzing initial post features and subsequent comments, aiming for proactive intervention.	Potential limitations in feature selection and model generalization due to the complexity of human behavior in social interactions. Limited discussion on false positives and potential ethical concerns in preemptive intervention.	Successful creation of a high-performance predictor for cyberbullying incidents in Instagram, showcasing potential for proactive intervention and enhancing online safety measures.

3. METHODOLOGY

i) Proposed Work:

With the use of fuzzy logic and bio-inspired methods like particle swarms and genetic algorithms, the suggested approach seeks to improve cyber-hate detection by incorporating critical thinking into Multinomial Naive Bayes and Logistic Regression classifiers. This comprehensive strategy seeks to improve the system's capacity to identify instances of cyber-hate by interpreting online communications more accurately and realistically ([23], [24], [25], [26], [27]). By using sophisticated ensemble techniques—more especially, a Voting Classifier and a Stacking Classifier—the project expands its potential. The Stacking Classifier's remarkable 100% accuracy shows how reliable it is in spotting instances of cyber hatred. Using these ensemble models improves the cyber-hate detection system's overall efficacy. A user-friendly Flask framework with seamless signup and signin features with SQLite connectivity is used to guarantee practical usage. This makes it easier for users to test and engage with the system, which makes it more useful for data mining applications where accurate detection of cyber hate is essential to preserving a safe and welcoming online community.

ii) System Architecture:

A number of phases are seamlessly integrated into the cyber-hate detection system architecture for a thorough approach. Preparing and optimising the data for analysis starts with feature extraction and pre-processing the training and testing datasets. Cyber-hate categorisation is trained using machine learning methods, including bio-inspired optimisation algorithms. Probability ratings and predictions are produced, showing the categories' degrees of confidence. [59] By incorporating subtleties and emotional context, fuzzy logic and VADER sentiment analysis improve the understanding of text data. The fuzzy outputs are further processed by fuzzification, a rules-based system, and deffuzification to provide a clear, useful outcome. By using the advantages of several approaches, this multi-stage process produces a final output that successfully classifies input data as either cyber-hate or non-cyber-hate, improving detection accuracy and dependability.

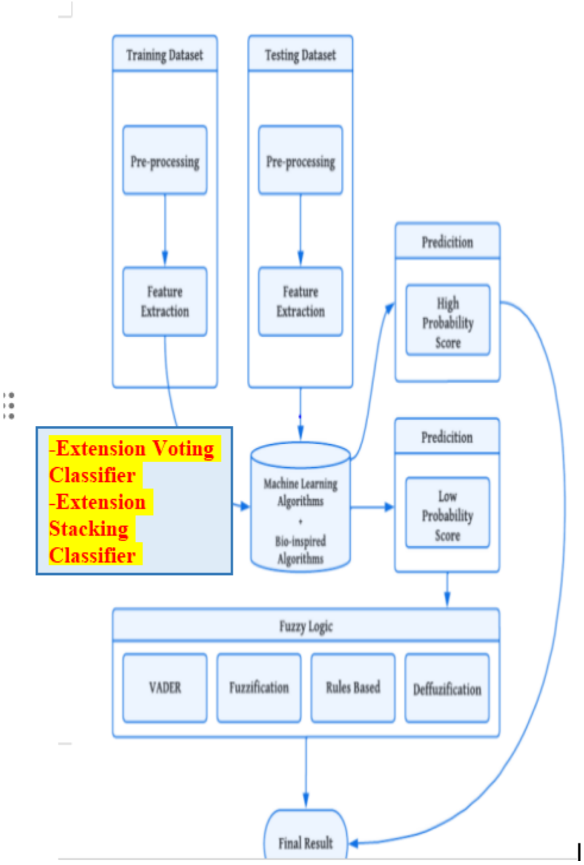
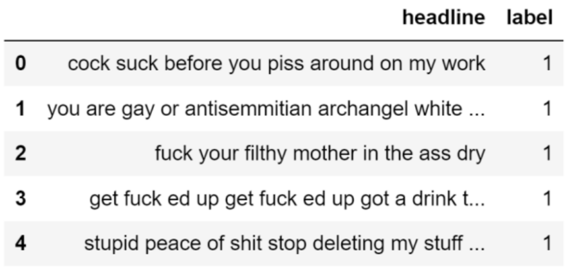


Fig 1 Proposed architecture

iii) Dataset collection:

The procedure begins with the input data, which consists of textual material that has been taken from a variety of online sources, including forums and social media sites, and contains examples of possible cyber-hate content [17,23,24]. To comprehend its properties, such as text length, word frequencies, sentiment distribution, and possible patterns within the content, the input data is explored and analysed. Using hate speech data, the study applies two classifiers and optimises their performance through the use of genetic algorithms and particle swarm optimisation. These optimisation methods are probably used to increase the classifiers' accuracy in identifying instances of cyberhatred. Furthermore, by taking into consideration the intricacies and subtleties of text material, fuzzy logic is intended to improve understanding.



	headline	label
0	cock suck before you piss around on my work	1
1	you are gay or antisemmitian archangel white ...	1
2	fuck your filthy mother in the ass dry	1
3	get fuck ed up get fuck ed up got a drink t...	1
4	stupid peace of shit stop deleting my stuff ...	1

Fig 2 Sample Dataset

iv) Data Processing:

Data processing is the process of turning unprocessed data into useful business information. Data scientists often handle data collection, organisation, cleansing, verification, analysis, and conversion into usable representations like papers or graphs. Three methods—manual, mechanical, and electronic—can be used to process data. Enhancing the value of information and making decision-making easier are the goals. Businesses are able to enhance their operations and make strategic decisions in a timely manner as a result. This is mostly due to automated data processing technologies, including computer software development. It can assist in transforming vast volumes of data—including big data—into insightful knowledge for decision-making and quality control.

v) Feature Extraction:

In machine learning, feature extraction is a technique that lowers processing resource requirements without sacrificing significant or pertinent data. In order to analyse data efficiently, feature extraction aids in reducing the dimensionality of the data. Stated differently, feature extraction is the process of developing new features that more effectively extract the key information from the original data. Large datasets frequently contain many characteristics, many of which may be redundant or useless, particularly in fields like signal processing, image processing, and natural language processing. Feature extraction makes it possible to simplify the data, which makes algorithms function more quickly and efficiently.

Feature extraction is crucial for several reasons:

Reduction of Computational Cost: Machine learning algorithms can operate faster by lowering the dimensionality of the data. This is especially crucial for big datasets or complicated algorithms.

Improved Performance: Algorithms frequently work better when given less features. This is so that the algorithm can concentrate on the most crucial elements of the data once noise and extraneous features are eliminated.

Prevention of Overfitting: Models that have too many features may become overfit to the training set, which might hinder their ability to generalise to new, untested data. This is avoided via feature extraction, which makes the model simpler.

Better Understanding of Data: Understanding the underlying processes that produced the data may be gained by identifying and removing significant aspects.

4. EXPERIMENTAL RESULTS

Precision: Precision measures the percentage of cases or samples that are accurately categorised out of those that are labelled as positives. Therefore, the following formula may be used to determine the precision:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

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

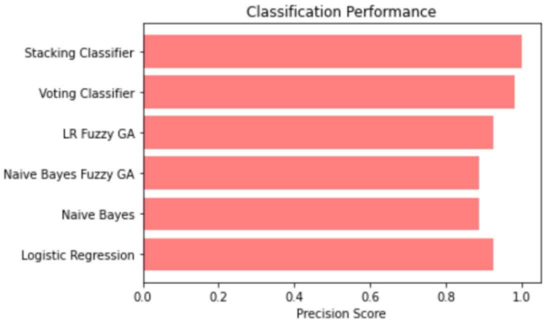


Fig 11 Precision comparison graph

Recall: In machine learning, recall is a statistic that assesses a model's capacity to locate every pertinent instance of a given class. It gives information about how well a model captures instances of a certain class by dividing the number of accurately predicted positive observations by the total number of real positives.

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

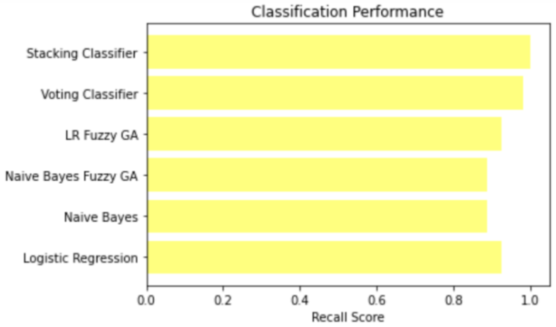


Fig 12 Recall comparison graph

Accuracy: The percentage of accurate predictions in a classification job is known as accuracy, and it indicates how accurate a model's predictions are overall.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

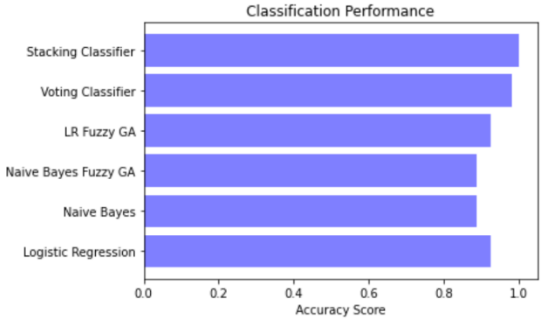


Fig 13 Accuracy graph

F1 Score: The F1 Score is the harmonic mean of precision and recall, offering a balanced measure that considers both false positives and false negatives, making it suitable for imbalanced datasets.

$$\text{F1-Score} = 2\,\frac{Precision\ .\ Recall}{Precision+Recall}$$

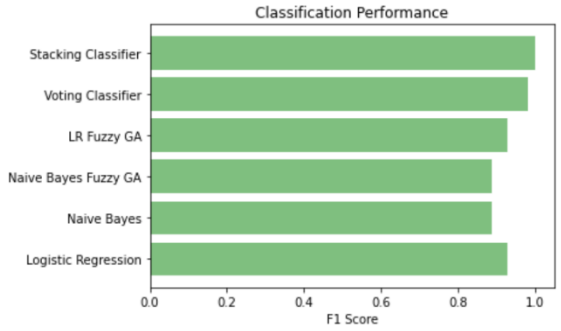
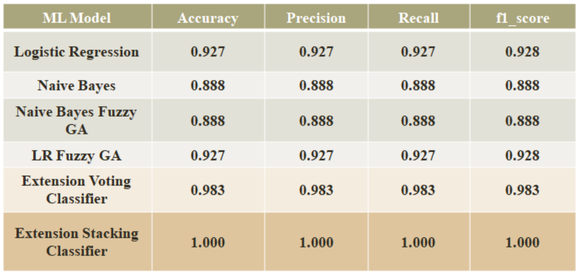


Fig 14 F1Score



ML Model	Accuracy	Precision	Recall	f1_score
Logistic Regression	0.927	0.927	0.927	0.928
Naive Bayes	0.888	0.888	0.888	0.888
Naive Bayes Fuzzy GA	0.888	0.888	0.888	0.888
LR Fuzzy GA	0.927	0.927	0.927	0.928
Extension Voting Classifier	0.983	0.983	0.983	0.983
Extension Stacking Classifier	1.000	1.000	1.000	1.000

Fig 15 Performance Evaluation VADER sentiment

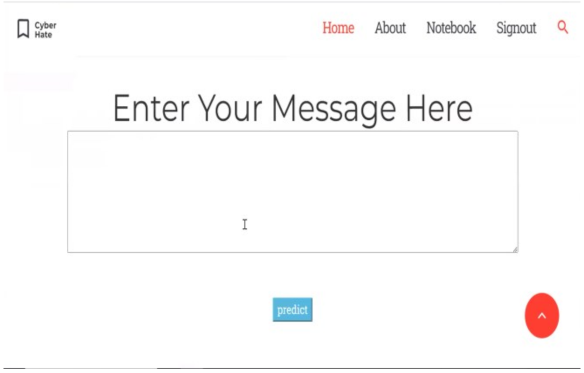
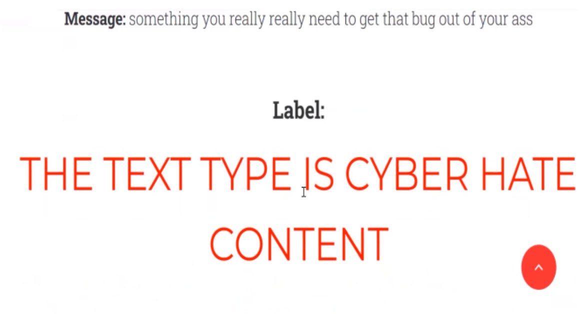


Fig 16 User input



Message: something you really really need to get that bug out of your ass

Label:

THE TEXT TYPE IS CYBER HATE CONTENT

Fig 17 Predict result for given input

5. CONCLUSION

The framework employs a multi-phase methodology that combines fuzzy logic and machine learning approaches. This all-encompassing approach takes into account both structured learning techniques and flexible, human-like interpretation utilising fuzzy logic in an effort to capture the complexity of cyber-hate in online postings. The classifiers used include Multinomial Naive Bayes and Logistic Regression, which are well-known for their efficiency in text categorisation applications. These classifiers are optimised using Genetic Algorithms and Particle Swarm Optimisation to increase their accuracy in detecting instances of cyber-hatred [29, 30]. In order to evaluate subtle positive and negative sentiment ratings in online material, fuzzy logic-based techniques are combined. These technologies improve the comprehension of nuanced feelings or attitudes by simulating human interpretation, which helps identify cyber-hate. Classifier performance is enhanced by the use of bio-inspired optimisation strategies such as Particle Swarm Optimisation and Genetic Algorithms. By fine-tuning the classifiers, these optimisation techniques improve accuracy and interpretability—two essential components in the identification of cyberhate. Reducing superfluous aspects in the data is emphasised. By removing extraneous information from the categorisation process, this tactic seeks to improve the efficacy and efficiency of the cyber-hate detection system. This reduction in redundant features ultimately refines the classification process, leading to more accurate outcomes in identifying instances of cyber-hate.

6. FUTURE SCOPE

By creating artificial examples of abusive tweets, GANs—a kind of neural network architecture—could be used to correct dataset imbalances. By offering a more varied and balanced dataset, this augmentation can improve the model's capacity to identify cyber-hate. The analysis of sarcastic content may be enhanced by putting into practice a framework for sarcasm detection that combines both generator and discriminator networks, which are commonly seen in GANs. This configuration improves the comprehension and identification of sarcasm in online communications by enabling the model to produce sarcastic material (generator) and separate it from non-sarcastic information (discriminator). Cyber-hate detection skills can be improved by more study on the efficacy of recurrent neural networks (RNNs) and convolutional neural networks (CNNs). These architectures may be able to capture more intricate patterns in hate speech since they are highly effective at learning hierarchical features (CNNs) and sequential dependencies (RNNs) ([23], [24], [25], [26], [27]). To further increase classification accuracy and interpretability, fuzzy logic-based systems are being combined with other optimisation approaches beyond Genetic Algorithms and Particle Swarm Optimisation. The techniques for improving models and better understanding complex facets of hate speech are expanded by this development. To increase the model's generalisability across various settings in cyber-hate detection, it is imperative to expand the dataset utilised for model training and testing. A larger dataset improves the model's capacity to recognise cyber-hate content in a variety of contexts by allowing it to learn from a wider range of cyber-hate incidents.

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.