International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

A MULTI-STAGE MACHINE LEARNING AND FUZZY APPROACH TO CYBER-HATE DETECTION

[1] K. Kalyani, [2] R.Divya
[1]Assistant Professor,[2]Student
Department of MCA
Sree Chaitanya College of Engineering, Karimnagar

ABSTRACT

The rapid growth of online social platforms has increased the prevalence of cyber-hate, including hate speech, abusive language, offensive content, and discriminatory expressions targeting individuals or groups based on race, religion, gender, ethnicity, or other social characteristics. Such harmful content negatively impacts social harmony, mental well-being, and online community engagement. Traditional keyword-based detection methods often struggle to accurately identify the complex linguistic patterns and contextual nuances associated with cyber-hate. This paper presents a multi-stage machine learning and fuzzy logic approach for cyber-hate detection. The proposed framework employs Natural Language Processing techniques for text preprocessing, feature extraction, and semantic analysis, followed by machine learning algorithms for initial classification of online content. A fuzzy inference system is subsequently utilized to handle uncertainty, ambiguity, and varying degrees of hate intensity within textual data. The multi-stage architecture enhances detection accuracy by combining the predictive capabilities of machine learning with the interpretability and flexibility of fuzzy logic. Experimental analysis demonstrates that the proposed approach effectively identifies cyber-hate content, reduces false classifications, and improves overall detection performance. The developed framework provides an intelligent and scalable solution for promoting safer and more inclusive online communication environments.

I. 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. Numerous machine learning models, including Multinomial Naive Bayes and Logistic Regression, have been used in combination with Particle Swarm Optimisation, Genetic Algorithm, and other Bio-Inspired Optimisation techniques. The optimal feature selection subset that more accurately

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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 desired solution to language challenges; ii) It addresses reasoning and delivers perspectives that are closer to the precise sentiment values. The structure of the paper is as follows: Related studies are presented in Section II, datasets are described in Section III, and framework-related information is provided in Section IV. Sections V–VII delineate the framework's three phases. The technological configuration and setting used for the report's tests are detailed in Section VIII. Results and discussion are presented in Section IX, and finally The article is concluded in Section X.

II. LITERATURE SURVEY

"Detection of cyberbullying on social media through machine learning,"

A. Mohammed, Z. Emad, E. Amer, M. Nashaat, M. Ahmed, J. Hani,

Cyberbullying, a kind of bullying via electronic messaging, has evolved as a result of the exponential growth in social media users. Bullies might use social networks as a rich environment in which to target victims with assaults. Finding appropriate measures to identify and stop cyberbullying is vital given the effects it has on its victims. By identifying the bullies' linguistic patterns, machine learning may assist create a model that can automatically identify instances of cyberbullying. A supervised machine learning method for identifying and stopping cyberbullying is presented in this study. Bullying behaviours are trained and identified using a variety of classifiers. According to the assessment of the suggested method on a cyberbullying dataset, SVM scores 90.3 and Neural Network performs better, achieving 92.8% accuracy. Additionally, on the same dataset, NN performs better than other classifiers of comparable effort.

"Modelling the detection of cyberbullying via text,"

H. Lieberman, R. Reichart, and K. Dinakar,

With more and more teenagers acknowledging that they have experienced cyberbullying as a victim or witness, the problem has reached frightening proportions. This societal threat has been made worse by anonymity and the absence of effective oversight in the technological medium. A victim is more prone to internalise comments or articles that touch on delicate subjects that are personal to them, which often has catastrophic results. We break down the detection issue as a whole into sub-problems of text categorisation and sensitive topic identification. We test a variety of binary and multiclass classifiers on a dataset of 4500 YouTube comments. For individual labels, we find that binary classifiers perform better than multiclass classifiers. Our results demonstrate that developing individual topic-sensitive classifiers helps address the identification of textual cyberbullying.

"Identifying cyberbullying: Search terms and methods,"

L. Edwards, A. Garron, K. Reynolds, and A. Kontostathis,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

The vocabulary employed in cyberbullying is closely examined in this article. We use a selection of posts from Formspring.me as our corpus. On the social networking site Formspring.me, individuals may ask each other questions. Teenagers and young adults are its target audience, and the site has a lot of stuff about cyberbullying—between 7% and 14% of the articles we looked at included such content. In this article, two outcomes are reported. The purpose of our first tests was to gain insight into the words that cyberbullies use and the context in which they are employed. The most popular phrases for cyberbullying have been determined, and queries that may be used to find cyberbullying material have been created. At rank 100, five of our queries had an average accuracy of 91.25%. We expanded on this work in our second series of tests by identifying cyberbullying using a supervised machine learning technique. In addition to finding another querying method that successfully assigned scores to postings from Formspring.me, the machine learning studies also found other phrases that are associated with material related to cyberbullying. There is a significant density of cyberbullying material in the postings with the highest ratings.

"Identifying instances of harassment on Web 2.0,"

L. Edwards, B. D. Davison, A. Kontostathis, Z. Xue, L. Hong, and D. Yin

Web-based communities and applications have grown and changed as a result of Web 2.0. These communities provide venues for cooperation and information exchange. Additionally, they allow for improper online behaviour, such harassment, when individuals intentionally offend other members of a virtual community by posting unpleasant remarks. Online harassment detection is a novel and difficult endeavour, and few systems presently make an effort to address this issue. In this study, we identify harassment using a supervised learning strategy. Our method makes use of the contextual, sentiment, and content elements of documents. The experimental findings presented here demonstrate that our methodology outperforms a number of baselines, including Term Frequency-Inverse Document Frequency (TFIDF) techniques, by a considerable margin. When emotion and contextual feature traits are added to TFIDF, it becomes possible to identify online abuse.

"Enhanced detection of cyberbullying through the use of gender data,"

R. Ordelman, D. Trieschnigg, F. D. Jong, and M. Dadvar

Friendships, relationships, and social communication are all changing as a consequence of the development of social networks, and new definitions seem to be relevant. It's possible to have hundreds of "friends" without ever meeting them. Concurrent with this shift, there is mounting evidence that kids and teenagers are bullying others online using social media apps. Current research on cyberbullying detection has mostly ignored the traits of the players engaging in cyberbullying in favour of concentrating on the conversation's substance. According to social research on cyberbullying, a harasser's written language changes depending on their characteristics, such as gender. In this work, we trained a gender-specific text classifier using a support vector machine model. We showed that a classifier's discriminating ability to identify cyberbullying is enhanced when gender-specific linguistic variables are taken into consideration.

III. SYSTEM ANALYSIS AND DESIGN

EXISTING SYSTEM

Many countries have passed legislation to combat cyberbullying in reaction to the widespread increase of cyber hatred. For

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

instance, the Malicious Communications Act of 1988 has been implemented in the United Kingdom [3]. After a conviction, this legislation prescribes punitive measures, such as a maximum six-month jail sentence and a monetary fine for the perpetrator. Additionally, the perpetrator may face criminal charges under the Harassment Act of 1997 if the victim experiences fear or distress as a result of the offender's online actions. In a similar vein, the Canadian legal system uses a variety of preventive strategies to combat cyberbullying, such as jail time, device seizure, and restitution for the harmed parties. The possible charges against the offenders of cyberbullying are determined by the gravity of the occurrence; these accusations may include criminal harassment, threats, intimidation, public promotion of hate, and offence against the person and reputation [3].

In order to combat instances of cyberbullying, some states in the US have passed legislation imposing a range of punitive penalties, such as monetary fines and jail time. On the other hand, several governments have not yet provided a clear and thorough explanation of the legal frameworks governing cyberbullying.

Researchers have been inspired to create automated systems that would identify and handle the issue of online hatred since there is a dearth of laws addressing it globally. The big data age we are presently living in has made it feasible to access and analyse vast amounts of data on people and their societies, which were previously impossible to do. OSNs like Facebook, Instagram, and Twitter may provide data for research, including social influences, community, content, and link prediction.

Disadvantages

- Conventional machine learning models, which are most often employed in the categorisation of online hatred, are not utilised in an existing system.
- Logistic regression, which is more accurate and efficient, has never been utilised in an existing system.

PROPOSED SYSTEM

Four distinct hybrid models were suggested after an assessment of the Machine Learning classifiers' performance. While the other two models (NB-Fuzzy-PSO and NBFuzzy-GA) were intended to improve the outcomes of Multinomial Naive Bayes, the first two models (LG-Fuzzy-PSO and LG-Fuzzy-GA) were created to directly improve the results of Logistic Regression.

• LG-Fuzzy-GA combines Genetic Algorithm with Fuzzy Logic and Logistic Regression; LG-Fuzzy-PSO combines Particle Swarm Optimisation with Logistic Regression and Fuzzy Logic.

• NB-Fuzzy-PSO: Combines Multinomial Naïve Bayes, fuzzy logic, and particle swarm optimisation.

NB-Fuzzy-GA combines Genetic Algorithm with Multinomial Naive Bayes and Fuzzy Logic. Machine learning classifiers like Logistic Regression and Multinomial Naive Bayes had relatively low complexity before using Particle Swarm Optimisation (PSO) and Genetic Algorithms (GA). Both classifiers are simple and efficient methods for dealing with situations involving binary classification.

A linear model called logistic regression (LR) uses the logistic function, often known as the sigmoid function, to estimate the likelihood of a certain class or event. It determines which coefficients are best for reducing the difference between expected probability and actual classes. The amount of samples and features in the dataset affects the matrix multiplication and

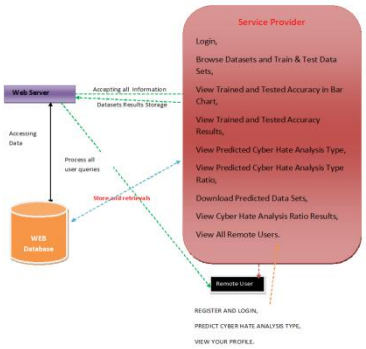
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

inversion operations utilised during training, which are the main basis of the LR algorithm's calculation.

Advantages

- Three different steps define the suggested framework. Preprocessing is the first step that eliminates noise from the datasets.
- In the second step, machine learning classifiers were constructed utilising bio-inspired optimisation approaches such genetic algorithms and particle swarm optimisation.
- Based on the machine learning confidence ratings obtained in the second step, fuzzy logic was employed in the final stage.

SYSTEM ARCHITECTURE



IV. IMPLEMENTATION

Modules

Service Provider

The Service Provider must use a working user name and password to log in to this module. Following a successful login, he may do various tasks including browsing datasets and training and testing datasets. View Results of Trained and Tested Accuracy, View Trained and Tested Accuracy in Bar Chart, View the Job Title Identification Type Ratio, View the Predicted Job Title Identification Type, Get Predicted Data Sets here. View All Remote Users and Job Title Identification Type Ratio Results.

View and Authorize Users

The administrator may see a list of all registered users in this module. Here, the administrator may see the user's information, like name, email, and address, and they can also grant the user permissions.

Remote User

A total of n users are present in this module. Before beginning any actions, the user needs register. Following registration, the user's information will be entered into the database. Following a successful registration, he must use his password and authorised user name to log in. Following a successful login, the user will do tasks such as registering and logging in, predicting the job title and identification type, Examine your profile.

ALGORITHIMS

Naïve Bayes

The supervised learning technique known as the "naive bayes approach" is predicated on the straightforward premise that the existence or lack of a certain class characteristic has no bearing on the existence or nonexistence of any other feature.

However, it seems sturdy and effective in spite of this. It performs similarly to other methods of guided learning. Numerous explanations have been put forward in the literature. We emphasise a representation bias-based explanation in this lesson. Along with logistic regression, linear discriminant analysis, and linear SVM (support vector machine), the naive bayes classifier is a linear classifier. The technique used to estimate the classifier's parameters (the learning bias) makes a difference.

Although the Naive Bayes classifier is commonly used in research, practitioners who want to get findings that are useful do not utilise it as often. On the one hand, the researchers discovered that it is very simple to build and

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

apply, that estimating its parameters is simple, that learning occurs quickly even on extremely big datasets, and that, when compared to other methods, its accuracy is rather excellent. The end users, however, do not comprehend the value of such a strategy and do not get a model that is simple to read and implement.

As a consequence, we display the learning process's outcomes in a fresh way. Both the deployment and comprehension of the classifier are simplified. We discuss several theoretical facets of the naive bayes classifier in the first section of this lesson. Next, we use Tanagra to apply the method on a dataset. We contrast the outcomes (the model's parameters) with those from other linear techniques including logistic regression, linear discriminant analysis, and linear support vector machines. We see that the outcomes are quite reliable. This helps to explain why the strategy performs well when compared to others. We employ a variety of tools (Weka 3.6.0, R 2.9.2, Knime 2.1.1, Orange 2.0b, and RapidMiner 4.6.0) on the same dataset in the second section. Above all, we make an effort to comprehend the outcomes.

Random Forest

Random forests, also known as random decision forests, are ensemble learning techniques that build a large number of decision trees during training for tasks like regression and classification. The class chosen by the majority of trees is the random forest's output for classification problems. The mean or average forecast of each individual tree is given back for regression tasks. The tendency of decision trees to overfit to their training set is compensated for by random decision forests. Although random forests are less accurate than gradient enhanced trees, they often perform better than choice trees. However, their performance may be impacted by data peculiarities.

Tin Kam Ho[1] developed the first algorithm for random decision forests in 1995 by using the random subspace technique, which in Ho's definition is a means of putting Eugene Kleinberg's "stochastic discrimination" approach to classification into practice.

Leo Breiman and Adele Cutler created an algorithm extension and filed for a trademark in 2006 for "Random Forests" (owned by Minitab, Inc. as of 2019).The extension builds a set of decision trees with controlled variance by combining Breiman's "bagging" concept with random feature selection, which was initially proposed by Ho[1] and then separately by Amit and Geman[13].

Businesses often employ random forests as "blackbox" models since they need minimal setup and provide accurate forecasts across a variety of inputs.

SVM

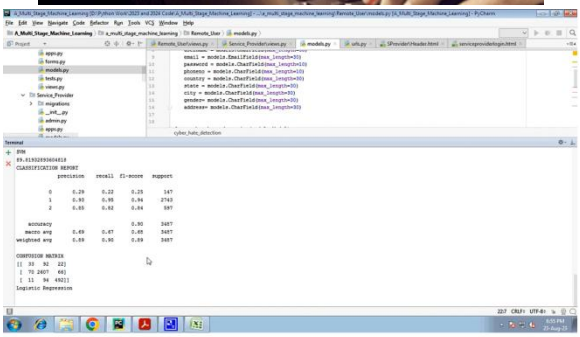
The goal of a discriminant machine learning approach in classification problems is to identify a discriminant function that can accurately predict labels for newly acquired instances based on an independent and identically distributed (iid) training dataset. A discriminant classification function takes a data point x and assigns it to one of the several classes that are part of the classification job, in contrast to generative machine learning techniques that call for calculations of conditional probability distributions. Discriminant techniques are less effective than generative approaches, which are mostly used when prediction entails the identification of outliers. However, they need less training data and processing resources, particularly when dealing with a multidimensional feature space and when just posterior probabilities are required. Finding the equation for a multidimensional surface that optimally divides the various classes in the

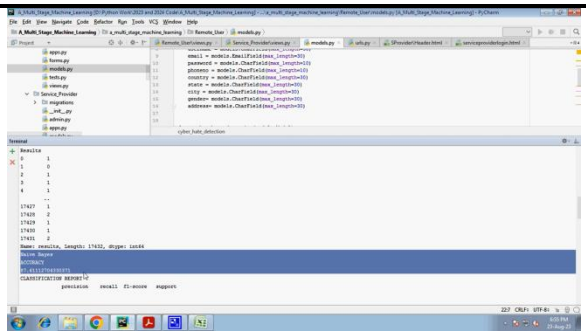
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

feature space is the geometric equivalent of learning a classifier.

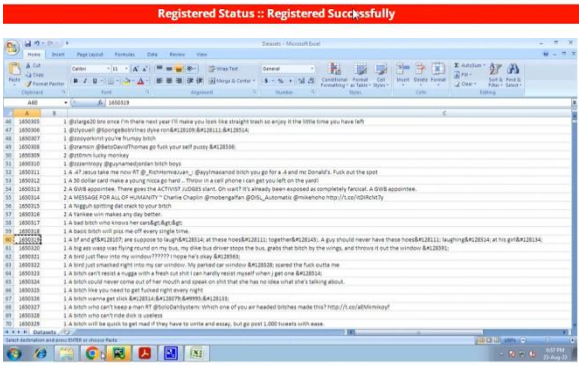
SVM is a discriminant approach that, unlike genetic algorithms (GAs) or perceptrons, which are both often used for classification in machine learning, always returns the same optimum hyperplane value since it solves the convex optimisation issue analytically. The initialisation and termination criteria have a significant impact on the solutions for perceptrons. While the perceptron and GA classifier models are distinct every time training is started, training yields uniquely specified SVM model parameters for a given training set for a certain kernel that converts the data from the input space to the feature space. The only goal of GAs and perceptrons is to reduce training error, which will result in several hyperplanes satisfying this criterion.

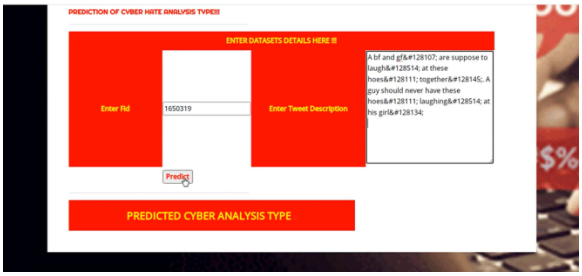
V. SCREENSHOTS

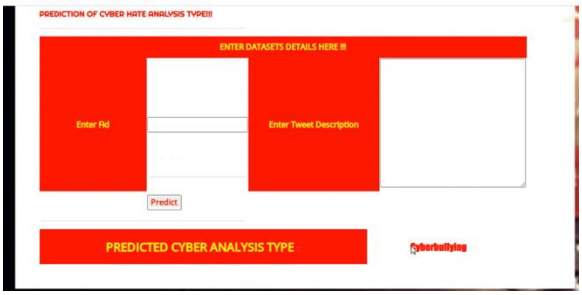


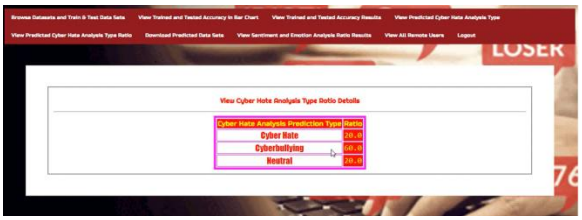


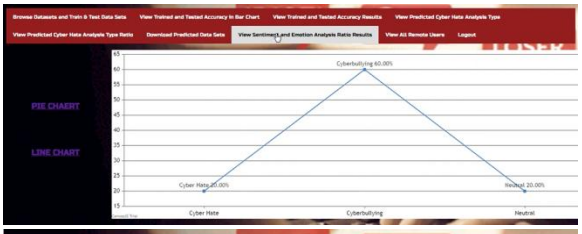
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper











VI. CONCLUSION

This study presented a multi-stage machine learning and fuzzy logic framework for cyber-hate detection in online environments. The proposed approach integrates Natural Language Processing, machine learning classification, and fuzzy inference mechanisms to accurately identify and assess hate-related content across digital platforms. By combining machine learning's predictive capabilities with fuzzy logic's ability to manage uncertainty and linguistic ambiguity, the system effectively detects varying levels of cyber-hate and offensive language. Experimental observations indicate that the proposed framework achieves improved detection accuracy, enhanced

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

classification reliability, and reduced false positive rates compared to conventional detection approaches. Furthermore, the multi-stage architecture enables better interpretation of contextual information and supports adaptive content moderation strategies. The proposed system contributes to safer online communities, improved digital communication standards, and more effective monitoring of harmful content. Future research may focus on integrating transformer-based language models, multilingual hate speech detection, explainable artificial intelligence techniques, and real-time social media monitoring systems to further enhance cyber-hate detection performance and online platform safety.

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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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.