A Survey on Identification of Hate Speech on Social Media Post

Archika Jain
Department of Computer Engineering
Suresh Gyan Vihar University
Jaipur, India
archikaagarwal@gmail.com

Sandhya Sharma
Department of Electronics & Communication
Suresh Gyan Vihar University
Jaipur, India
sandhya.sharma@mygyanvihar.com

2022 3rd International Conference on Computing, Analytics and Networks (ICAN) | 978-1-6654-9944-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAN56228.2022.10007283

***Abstract* - In the age of social media and mobile internet, the development of autonomous methods for online detection of hate speech or abusive language is vital for societal and community empowerment because social media is accessible worldwide via the internet. Anybody may simply attack someone or a group who adheres to a different culture or ideology on social media. While everyone has the freedom to express their own opinions, it should not be destructive, and everyone has the right to be free to say anything. To learn about hate speech, a review process include a large number of research articles which were published in the period of year 2017 to year 2022. After an exhaustive review process, we have found the common findings, strengths, weaknesses, gaps and solution approaches & their results.**

Keywords - hate speech, tweet dataset, social media, machine learning, and accuracy

I. INTRODUCTION

Hate speech is defined as public speech that expresses hate or encourages violence towards a person or group. It is based on sexual orientation, religion, race and sex. Hate speech is often defined as expressions of hatred or disparagement directed at an individual or a group. Hate speech is defined differently in different countries.

There has been a lot of discussion about free speech, hate speech, and hate speech laws. Hate speech is defined as speech, gestures, conduct, writing, or displays that incite violence or prejudicial actions against a group or individuals based on their membership in the group, or disparage or intimidate a group or individuals based on their membership in the group, according to the laws of some countries. A group may be identified by the law based on particular criteria. In some countries, hate speech is not a legal term. Furthermore, the hate speech is legally protected in many countries, including the United States. In some countries, a victim of hate speech may seek reparation through civil or criminal law.

There is a large scale of other language datasets are missing for detecting the hate speech. Also audio-video dataset, image dataset are not available. So incorporate with this predict the hate speech from multi-model dataset and identify the linguistic features.

For example, the Nazi swastika, the Confederate Battle Flag (of the Confederate States of America), and pornography have all been considered hate speech by a variety of people and groups.

A. Definition of Machine Learning:

Machine learning is a branch of artificial intelligence. The basic goal of machine learning is to understand the structure of data and fit that data into models so that humans can understand and utilize it.

Artificial intelligence (AI) and computer science's machine learning field focuses on using data and algorithms to mimic how people learn, progressively increasing the accuracy of its predictions.

B. Machine Learning Key Elements:

- Data Set
- Algorithms
- Models
- Feature Extraction
- Training

C. Types of Machine Learning

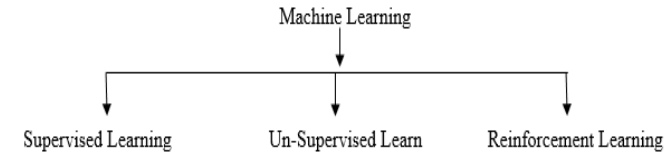

Fig. 1. Types of machine learning

a) Supervised Learning: For training the data machine learning algorithm is used [1]. Train the dataset from the larger dataset.

b) Un-Supervised Learning: Unsupervised machine learning has the advantage of being able to cope with untagged data. Labels are employed in supervised learning to assist the computer in making judgments that determine the strength of any association between two data items [1].

c) Reinforcement Learning: Considers how people benefit from data in their everyday lives. It incorporates a trial-and-error system that learns from various settings. Non-favorable outcomes are discouraged or punished, whereas positive outcomes are encouraged or reinforced. Reinforcement learning places the algorithm in a work context with an interpreter and a reward system, based on the psychology idea of conditioning [1].

D. Application of Machine Learning:

Machine learning is a phrase in today's technology that is rapidly gaining acceptance. We utilize machine learning in our daily lives through Google Maps, Google Assistant, Alexa, and other such services, even if we aren't aware of it. Some of the most common real-world Machine Learning applications are as follows:

a) Image recognition: One of the most popular uses of machine learning is image identification. Image recognition is used in face detection and automatic tagging.

b) Speech recognition: It is used to convert the voice instructions into text.

c) Traffic prediction: It is used to find out the best aand fastest route on the basis of traffic conditions.

d) Product recommendations: Many e-commerce and entertainment businesses, like Amazon, Netflix, and others, employ machine learning to propose products to customers. Because of machine learning, after we conduct a product search on Amazon, we begin to see advertisements for related items while using the same browser to explore the internet.

e) Self-driving cars: The most well-known manufacturer, Tesla, is developing self-driving cars. Using an unsupervised learning method, it teaches driving-related motor vehicle models to recognise people and other entities.

f) Junk mail and worm filtering: Every new email we receive is classified as either critical, normal, or spam right away. Machine learning is the technology that enables us to get critical information. Our spam box is full with spam emails, and our mailbox is full of crucial symbols.

g) Virtual personal assistant: Among the virtual personal assistants available are Google Assistant, Alexa, Crotona, and Siri. As the name indicates, they aid us in finding information using our voice commands.

h) Online fraud detection: By identifying fraud transactions, machine learning makes our online transactions safer and more secure. There are several ways for a fraudulent transaction to occur when we execute an online transaction.

i) Stock market trading: Because there is always a chance of stake money fluctuations in the stock market, a machine learning is utilized to predict stock market trends.

j) Medical diagnosis: As a result, thanks to important developments in medical technology.

k) Automatic language translation: Nowadays, visiting a new area and not knowing the language is not an issue.

Utmost Contribution of our paper is we have deeply studied a large number of research article from the year 2017 to year 2022 and identified the findings, strengths, weaknesses and gaps on the basis of literature review.

II. Literature Review

A large number of research articles, processing methods, and techniques published in the period of year 2017 to year 2022.

A. Common Findings in "Hate Speech Detection Techniques":

It is based on the findings of a large number of research articles.

Common techniques used by researchers are:
Random Forest (RF) decision Tree
Sentiment analysis
Natural Language Processing (NLP) preprocessing tools
Natural Language Processing
Supervised learning and unsupervised learning
Latent Semantic Analysis (LSA)
Deep learning (Bi-LSTM and Bi-GRU)
Support vector machine algorithm
Naïve Bayes algorithm
Logistic Regression
Random Forest

Common classifiers used by researchers are:
Abusive language classifier
Logistic Regression (LR) Classifier
Binary and ternary classification
Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) classifiers
Support Vector Machines (SVM)
Random Forest Decision Tree (RFDT)

Common features used by researchers are:
Word n-gram feature
Character level
TF-IDF
Negative sentiment
Syntactic-based features
Character n-gram
Bag-of-word feature
Word skip-gram features

Findings of reviews are as follows: German refugee crisis unable to deal with problematic comments [2]. Make an effort to create a more comprehensive vocabulary of hate speech patterns that may be used in conjunction with a unigram lexicon to identify offensive and hostile online messages [3]. In the development of the testing dataset and feature selection procedure, hybridization techniques that employ more complicated sentence structures [4]. Data annotation is the most difficult task rather than collection [5]. For generation of polarity score in each tweet a sentiment analysis model was implemented [6]. Imbalanced learning on more advanced feature engineering and classification models [7]. It's a difficult assignment, to say the least. Both plain English and Sinhala comments can be found in Singlish comments [8]. Linguistic features will perform [9]. Further improved with the accuracy of detecting hate words in online content [10]. Roman Urdu data collected from YouTube [11]. Lack of data sources from which Sinhala text containing hate speech can be extracted [12]. Improvements can be done by considering other dependency parser and explore other traversing tree approaches to extract syntactic n-grams [13]. Used the proportion of dataset using sampling or down sampling methods, and try to contextual embedding methods [14]. Up sampling or down sampling techniques is used for unbalanced datathat is the based on language models [15].

B. Research Work Reviewed Strengths and Weaknesses:

a) Strengths: Fig. 2 has shown the strengths of many research articles. In this various techniques like RDT & n-gram [16], F-score [7], LR [17], CNN [18], RETINA [19] and LR [20] has been used for F-measure and Binary classificaation [3], Ternary classification [3], ROS [3], CNN [4] and LSTM [4] has been used for accuracy.

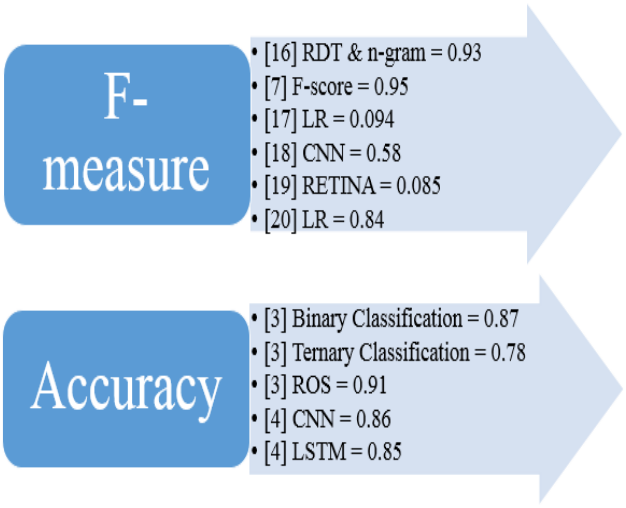

Fig. 2. Strengths

b) Weaknesses: Fig. 3 has shown the weaknesses of many research articles like low precision value [21], data annotation is difficult [5], generate polarity score for each tweet [6], lack of data for sinhala text [12] and overall F-score remains low [22].

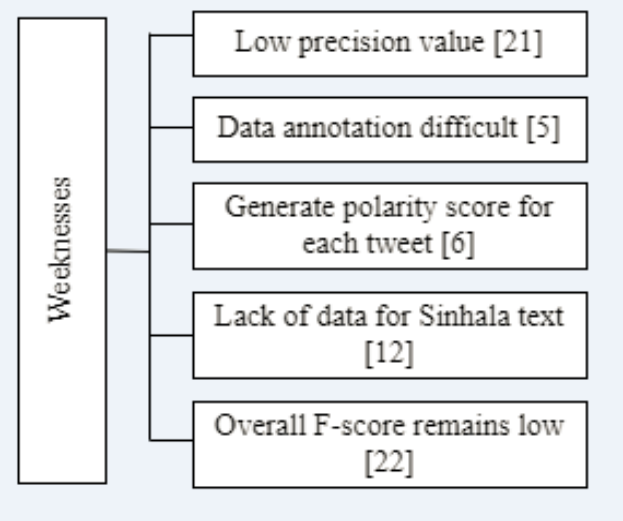

Fig. 3. Weaknesses

c) Gaps in the Published Research: Some gaps are identified while reviewing the papers.

- Absence of Large-scale German, French, Spanish, Russian and Hindi language dataset for hate speech detection.
- Identify the role of linguistic features in hate speech detection.
- Predict hate speech from the Visual and Textual features of the speech shared over Social Media.

C. Solution Approaches & their Results:

Table 1 define the different solution aapproaches and resultss on basis of literature review of a large number of research articles.

In Fig. 4 we have highest dataset i.e. 80000 Sinhala hate speech and lowest dataset i.e. 1000 tweets.

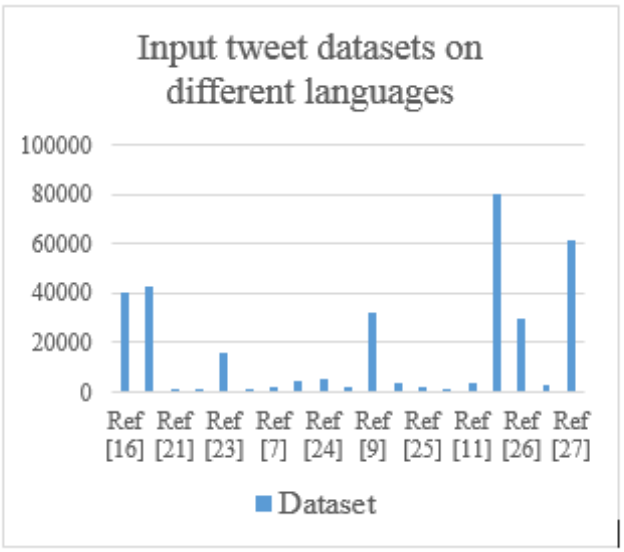

Fig. 4. Input Tweet datasets on different languages

This Fig. 5 contain accuracy, precision, recall and f1-score.

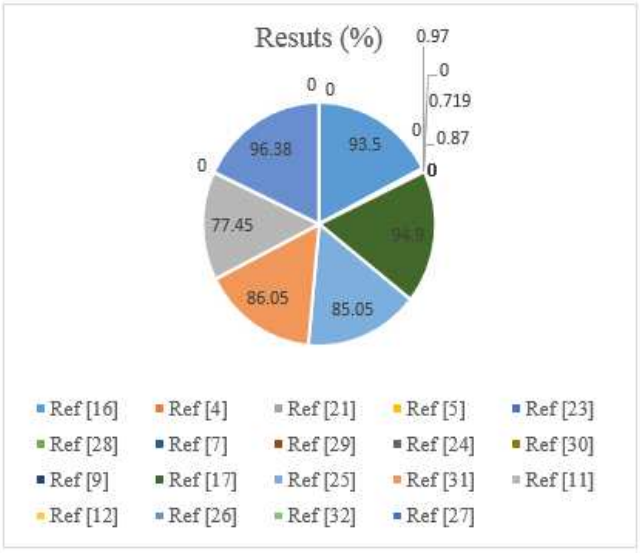

Fig. 5. Overall performance of the output parameters

TABLE I. DIFFERENT SOLUTION APPROACHES & RESULTS

Data Set	Techniques	Classifiers	Feature extraction	Result (%)
40000 Indonesian language tweets [16]	Machine learning	Naïve Bayes, SVM, Bayesian LR	Word n-gram feature	F-measure = 0.93
42365 Tweets [4]	Machine learning	Convolutional Neural Network (CNN) and LSTM	Word-level, n-gram level and Character level	Accuracy (CNN) = 0.86 and (LSTM) = 0.85
1288 Tweets [21]	Machine learning	SVM	TF-IDF	Precision =0.97
1000 Local English [5]	Machine learning	Naïve Bayes	TF-IDF	F-score = 0.719
16000 Tweets [23]	Deep learning , Fusion approach	LR), RF, SVM	N-gram feature	Precision = 0.056, recall = 0.035, F1-score = 0.046
1400 Indonesian tweets [28]	Machine learning	SVM	TF-IDF	Accuracy = 0.87
2000 Imbalance tweets [7]	Machine learning	SVM, LR and Naïve Bayes	Character n-gram	Accuracy = 0.91 F1-Score = 0.95
4587 Toxic South African tweets [29]	Machine learning	SVM	Word and character n-gram features	Accuracy = 0.91 and F1-score = 0.94
5678 Swedish [24]	Machine learning	Baseline classifier	TF-IDF	Accuracy = 0.74, Precision = 0.73, Recall = 0.73 and F1-score = 0.74
2400 Tweets [30]	Latent Dirichlet Allocation (LDA) and Self-Organizing Maps (SOM)	SVM, MNB	TF-IDF	Log-likelihood score=895399.1204, perplexity score 1290.9525 & sparsity score 0.17290
31962 Twitter dataset and 49159 Training dataset [9]	Machine learning	LR	Bag of words and TF-IDF	Accuracy (BOW) = 0.94 and (TFIDF) = 0.94
3567 Indonesian tweets [17]	Machine learning	LR, NB, and RFDT	TF-IDF and word bigrams	F-measure = 0.94
2000 Tweets [25]	Machine learning	CNN	N-grams and Bag-of-Words	Accuracy = 0.85
1000 Tweets [31]	Machine learning	GBDT	TF-IDF	Accuracy = 0.86
4000 Roman Urdu YouTube comments [11]	Synthetic Minority Oversampling Technique (SMOTE)	SVM	N-grams and TF-IDF	Accuracy = 0.77
80000 Sinhala hate speech [12]	Deep learning	CNN	N-gram	Accuracy = 0.83, F1 score = 0.67
30,000 Tweets [26]	Deep learning	CNN	DBOW and DMM	Accuracy = 0.96
3189 Tweets Hindi offensive text (HOT) [32]	Machine learning	LR, RF, Bi-LSTM and CNN	TF-IDF	F1-score (LR)=0.83, F1-score (RF)=0.75. F1-score (Bi-LSTM) = 0.860, F1-score (CNN) =0.83.
61396 Tweets [27]	Machine learning	CAT Boost, Gradient Boost and Random forest	Character n-grams and Word n-grams	Accuracy = 0.89, F1 = 0.87 and AUC = 0.88
16135 Tweets [33]	Machine learning	LSTM with Google news embedding	Word embedding	Accuracy = 0.81
24782 Tweets [34]	Lexicon based machine learning	NLP	TF-IDF	Accuracy = 0.80
1223 Tweets [35]	Machine learning	LSA	-	Precision = 0.67, Recall = 0.76 and Accuracy = 0.57
Out of 2,285 detecting misogynistic in the Urban dictionary [36]	Deep learning and machine learning	Bi-LSTM, Bi-GRU and RF	-	Accuracy (Bi-GRU) = 0.93, Sensitivity (Bi-LSTM) = 0.92 and Specificity (RF) = 0.96
25000 Tweets [37]	Machine learning	SVM	TF-IDF	Accuracy = 0.82
415,844 Tweets [38]	Deep learning	RNN	Word embedding	F1-score (GRU) = 0.85 and F1-score (LSTM) = 0.76
9925 Tweets in Dataset A and 31926 Tweets in Dataset B [39]	Deep learning	Bi-RNN	Doc2Vec (Dataset A) and Word2Vec (Dataset B)	For Dataset A, Precision = 0.53, Recall = 0.40 and F1-score = 0.46, For Dataset B, Precision = 0.76, Recall = 0.65 and F1-score = 0.70
1339 Tweets [40]	Machine learning	SVM and NB	TF-IDF	Accuracy (SVM) = 0.70 and Accuracy (NB) = 0.72

III. Conclusion & Future Scope

A review of a large number of research articles in the field of hate speech detection using machine learning is used to analyses and categories diverse hate speech in order to investigate and identify existing challenges and the scope of work. When using various machine learning algorithms, classifiers, and features, we found various research gaps that need to be filled. In this we explore various hate speech issues and find their solution approaches with results. The goal of these approaches and tactics is to reduce the inefficiencies associated with hate speech while enhancing its accuracy, precision, recall, and F1 score.

In future, we will apply machine learning classifiers to increase accuracy and work on various languages such as Hindi, German, French, and Hinglish languages etc.

References

[1] O. Oriola and E. Kotzé, "Evaluating machine learning techniques for detecting offensive and hate speech in south African tweets," in IEEE Access, vol. 8, pp. 21496-21509, 2020.

[2] M. Niemann, "Abusiveness is non-binary: five shades of gray in German online news-comments," IEEE 21st Conference on Business Informatics (CBI), vol. 1, pp. 11-20, 2019.

[3] H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate speech on Twitter: a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," in IEEE Access, vol. 6, pp. 13825-13835, 2018.

[4] C. Abderrouaf and M. Oussalah, "On online hate speech detection effects of negated data construction," IEEE International Conference on Big Data (Big Data), pp. 5595-5602, 2019.

[5] N. D. T. Ruwandika and A. R. Weerasinghe, "Identification of hate speech in social media," 18th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 273-278, 2018.

[6] T. Febriana and A. Budiarto, "Twitter dataset for hate speech and cyberbullying detection in Indonesian language," International Conference on Information Management and Technology (ICIMTech), vol. 1, pp. 379-382, 2019.

[7] H. Rathpisey and T. B. Adji, "Handling imbalance issue in hate speech classification using sampling-based methods," 5th International Conference on Science in Information Technology (ICSITech), pp. 193-198, 2019.

[8] H. M. S. T. Sandaruwan, S. A. S. Lorensuhewa and M. A. L. Kalyani, "Sinhala hate speech detection in social media using text mining and machine learning," 19th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 1-8, 2019.

[9] G. Koushik, K. Rajeswari and S. K. Muthusamy, "Automated hate speech detection on Twitter," 5th International Conference on Computing, Communication, Control and Automation (ICCUBEA), pp. 1-4, 2019.

[10] S. Gupta, S. Lakra and M. Kaur, "Study on BERT model for hate speech detection," 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1-8, 2020.

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

[12] S. W. A. M. D. Samarasinghe, R. G. N. Meegama and M. Punchimudiyanse, "Machine learning approach for the detection of hate speech in sinhala unicode text," 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 65-70, 2020.

[13] N. A. Abdul Aziz, M. Aizaini Maarof and A. Zainal, "Hate speech and offensive language detection: a new feature set with filter-embedded combining feature selection," 3rd International Cyber Resilience Conference (CRC), pp. 1-6, 2021.

[14] A. Marpaung, R. Rismala and H. Nurrahmi, "Hate speech detection in Indonesian Twitter texts using bidirectional gated recurrent unit," 13th International Conference on Knowledge and Smart Technology (KST), pp. 186-190, 2021.

[15] G. Singh, and G. K. Sethi, "Automatic land cover classification using learning techniques with dynamic features," International Journal of Innovative Technology and Exploring Engineering, 8(8S3), pp. 499–503, 2019.

[16] I. Alfina, R. Mulia, M. I. Fanany and Y. Ekanata, "Hate speech detection in the Indonesian language: a dataset and preliminary study," International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp. 233-238, 2017.

[17] D. Elisabeth, I. Budi and M. O. Ibrohim, "Hate code detection in Indonesian tweets using machine learning approach: a dataset and preliminary study," 8th International Conference on Information and Communication Technology (ICoICT), pp. 1-6, 2020.

[18] M. Beatty, "Graph-based methods to detect hate speech diffusion on twitter," IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 502-506, 2020.

[19] S. Masud, S. Dutta, S. Makkar, C. Jain, V. Goyal, A. Das and T. Chakraborty "Hate is the new infodemic: a topic-aware modeling of hate speech diffusion on twitter." IEEE 37th International Conference on Data Engineering (ICDE), pp. 504-515, 2021.

[20] J. Sachdeva, K. K. Chaudhary, H. Madaan and P. Meel, "Text based hate-speech analysis," International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 661-668, 2021.

[21] H. Şahi, Y. Kılıç and R. B. Saglam, "Automated detection of hate speech towards woman on Twitter" 3rd International Conference on Computer Science and Engineering (UBMK), pp. 533-536, 2018.

[22] G. Singh, G. K. Sethi, and S. Singh, "Performance analysis of deep learning classification for agriculture applications using sentinel-2 data", in IEEE Internet Computing, vol. 1393, 2021.

[23] M. Sajjad, F. Zulifqar, M. U. G. Khan and M. Azeem, "Hate speech detection using fusion approach," International Conference on Applied and Engineering Mathematics (ICAEM), pp. 251-255, 2019.

[24] J. Fernquist, O. Lindholm, L. Kaati and N. Akrami, "A study on the feasibility to detect hate speech in Swedish," IEEE International Conference on Big Data (Big Data), pp. 4724-4729, 2019.

[25] M. Khan, A. Abbas, A. Rehman and R. Nawaz, "Hate classify: a service framework for hate speech identification on social media" in IEEE Internet Computing, vol. 25, pp. 40-49, 2021.

[26] L. Ketsbaia, B. Issac and X. Chen, "Detection of hate tweets using machine learning and deep learning," IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 751-758, 2020.

[27] K. A. Qureshi and M. Sabih, "Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text," in IEEE Access, vol. 9, pp. 109465-109477, 2021.

[28] U. A. N. Rohmawati, S. W. Sihwi and D. E. Cahyani, "SEMAR: an interface for Indonesian hate speech detection using machine learning," International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), pp. 646-651, 2018.

[29] O. Oriola and E. Kotzé, "Automatic detection of toxic south African tweets using support vector machines with n-gram features," 6th International Conference on Soft Computing & Machine Intelligence (ISCMI), pp. 126-130, 2019.

[30] Y. Saini, V. Bachchas, Y. Kumar and S. Kumar, "Abusive text examination using latent dirichlet allocation, self-organizing maps and k means clustering," 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1233-1238, 2020.

[31] M. H. Abdurrahman, B. Irawan and C. Setianingsih, "A review of light gradient boosting machine method for hate speech classification on twitter," 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), pp. 1-6, 2020.

[32] Rahul, V. Gupta, V. Sehra and Y. R. Vardhan, "Ensemble based hinglish hate speech detection," 5th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1800-1806, 2021.

[33] J. Sachdeva, K. K. Chaudhary, H. Madaan and P. Meel, "Text based hate-speech analysis," International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 661-668, 2021.

[34] R. Martins, M. Gomes, J. J. Almeida, P. Novais and P. Henriques, "Hate speech classification in social media using emotional analysis," 7th Brazilian Conference on Intelligent Systems (BRACIS), pp. 61-66, 2018.

[35] I. M. Ahmad Niam, B. Irawan, C. Setianingsih and B. P. Putra, "Hate speech detection using latent semantic analysis (LSA) method based on image," International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), pp. 166-171, 2018.

[36] T. Lynn, P. T. Endo, P. Rosati, I. Silva, G. L. Santos and D. Ging, "A comparison of machine learning approaches for detecting misogynistic

speech in urban dictionary," International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), pp. 1-8, 2019.
[37] E. Ombui, L. Muchemi and P. Wagacha, "Hate speech detection in code-switched text messages," 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pp. 1-6, 2019.
[38] E. Sazany and I. Budi, "Hate speech identification in text written in Indonesian with recurrent neural network," International Conference on Advanced Computer Science and information Systems (ICACSIS), pp. 211-216, 2019.
[39] L. Jiang and Y. Suzuki, "Detecting hate speech from tweets for sentiment analysis," 6th International Conference on Systems and Informatics (ICSAI), pp. 671-676, 2019.
[40] S. Ahammed, M. Rahman, M. H. Niloy and S. M. M. H. Chowdhury, "Implementation of machine learning to detect hate speech in Bangla language," 8th International Conference System Modeling and Advancement in Research Trends (SMART), pp. 317-320, 2019.