Grenze International Journal of Engineering and Technology, Jan Issue

Hate Speech Detection on Social Media using Machine Learning and Deep Learning: A review

Usman[1] and S.M.K. Quadri[2]
[1-2]Jamia Millia Islamia, New Delhi, India
Email: usman.mca.du.2014@gmail.com, quadrismk@jmi.ac.in

***Abstract*— Online toxic narratives may cause disputes and hurt virtual communities. Social networking sites are used in modern culture to share ideas and feelings. However, it occasionally can result in hate speech. Hate speech is offensive targeting someone or a group of people because of one or more of their origins, races, genders, religions, impairments, or other traits. It is a serious problem since it can have terrible results. In this paper, we have systematically investigated the various research works based on machine learning and deep learning techniques for text-based offensive speech classification. This research also offers empirical facts on the inherent features of hateful speech and assists in identifying future research.**

***Index Terms*— Deep learning, Hate speech, Machine learning, Offensive speech.**

I. INTRODUCTION

Hate speech may propagate quickly on media platforms, owing to preconceptions or disagreements between groups inside and beyond nations. Twitter prohibits threatening or harassing individuals based on race, sexuality, religious practice, or other factors. YouTube also filters violent or hateful material based on age, ethnic background, and disability. Hate speech is often investigated in relation to internet radicalization or criminal activity [1], but it's also investigated in other circumstances. Web - based information distribution provides considerable impetus to explore automated hate speech detection [2]. Hate speech identification helps reduce crime and defend people's values. This research is essential because continual conflicts alter reality and dehumanise the assaulted Ukrainians. Studies demonstrate an upsurge in online hatred towards China, including racist and offensive propaganda blaming individuals for COVID-19. A decreased percentage of hate speech minimises cyberbullying, which impacts social serenity [3] and information security [4].

Hate speech remains an issue despite numerous research. Due of the diversity and complexity of hate speech classes, machine learning algorithms and humans have trouble recognising it. Hate involves strong prejudice. Figure 1 depicts hate speech. Hate speech is an offensive element between aggressive and abusive language. Gender - based, homosexualism, and religious hatred target groups or genders. The figure indicates that idea separation is complex [5]. Homophobic and r Racist tweets are more liable to be categorised as hateful speech than other harmful material [5]. You can't generalise whether an offensive text constitutes hate speech [6]. The growth of brief texts, like those found on Twitter, poses issues to the standard bag-of-words paradigm, resulting in sparse data owing to lack context [7]. Due to categorization and specifications, several datasets [8] may vary, causing machine learning algorithms comparisons difficult.

The literature indicated various literature review studies, although most focused on one subject or were outdated [5, 9–12]. [5] builds a general metadata framework for hate speech categorization using semantic analysis and

Grenze ID: 01.GIJET.10.1.57

fuzzy logic. The [9] research focused on hate speech connected to cybercrime and global legal frameworks, but not Twitter or machine learning datasets. [10] cites mostly legal texts that define hate speech for criminal punishment. The [11] study focused on hate speech geography, social media platform variety, and qualitative or quantitative research methodologies. This study reviews hate speech ideas, methodologies, and datasets to inform academics about state-of-the-art investigations in hate speech identification. The following is an overview of the paper: The methodology of the study is discussed in Section 2. The discussion of the results and the content analysis may be found in Section 3. The difficulties encountered in research are discussed in Section 4, along with possible future possibilities. Section 5 then presents an illustration of the findings reached.

Users of social media express themselves, debate issues, and interact with others. These platforms feature objectionable or hazardous material, making them unsafe. Offensive language, radicalism, religious prejudice, sports aggression, and hateful speech on media platforms. These people feel confident in their right to express themselves openly. Here came the challenge of reconciling free speech with offensive material. So the challenge becomes, how to limit hate speech while protecting free speech on social media. To use abusive language is to "express oneself in a way that includes abusive/dirty words or phrases, either orally or in writing" [13]. There is now a formal global definition of hateful speech. In today's culture, particularly on social media, this phrase is ubiquitous. It's a term used to describe how some people feel about other people because of their differences.

II. Hate Speech Detection Using Machine Learning Models

Machine learning is asub-field of artificial intelligence (AI) and computer science called machine learning focuses on using data and algorithms to simulate how humans learn, gradually increasing the accuracy of the system. The rapidly expanding discipline of data science includes machine learning as a key element. Algorithms are trained using statistical techniques to produce classifications or predictions and to find important insights in data mining projects. The decisions made as a result of these insights influence key growth indicators in applications and enterprises, ideally.

In Table I, we have summarized a number of research publications together with their methodologies and findings.

Table I. Detailed Review Of Several Existing Methods Based On Machine Learning For Hate Speech Classification

Paper	Dataset	Approach	Findings
Raza et al., 2022 [36]	Urdu tweets	Transfer learning	Via experimental data, we demonstrate that BERT, distil-BERT and xlm-roberta are all capable of achieving satisfying F1-scores of 0.68, 0.70, and 0.69, respectively, using transfer learning.
William et al., 2022 [35]	CrowdFlower	SVM	After extensive testing, it was shown that 79 percent accuracy could be achieved using just bigram characteristics.
Nauros et al., 2021 [37]	Facebook and YouTube comments in Bengali language	BengFastText, FastText and Word2Vec	In spite of the fact that all of the deep learning models did very well, the experiment demonstrated that SVM produced the best result with 87.5% accuracy.
Moin Khan et al., 2021 [38]	Urdu tweets	Evaluated the efficacy of five supervised learning methods, including one based on deep learning, for identifying hate speech.	With an F1 score of 0.906, Logistic Regression surpassed all other methods, including deep learning methods, at both levels of categorization.
Aljarah et al., 2020 [19]	A collection of tweets in Arabic that were racist or linked to media, terrorism or sports.	T-IDF, TF, RF with BoW, DT, NB and SVM	Using TF-IDF, RF achieved the greatest results (91% accuracy), making it the most effective model.
Silva & Roman, 2020 [20]	Tweets in Portuguese language	Text representation using N-Gram, BoW, NB, LR, MLP and SVM	When one of the earlier studies that accomplished the same objective using LSTM was assessed to the findings of the models utilised, SVM was found to be roughly 22% more effective. The training process of classic models may be more efficient.
Da Costa Abreau & Araujo De Souza, 2020 [21]	Used an existing dataset consisting of tweets and classified as either hateful speech, offensive language, or neutral language.	Naïve Bayes and Linear SVM	In the case of text classification, Naive Bayes performs well. With imbalanced data, linear SVM performance drops significantly.

Mubarak et al., 2020 [18]	Arabic tweets	SVM, LR, AdaBoost, GNB, RF and DT	SVM performs the best.
Mulki et al., 2019 [22]	L-HSAB dataset having tweets in Arabic language	N-Gram and TF with NB and SVM	In terms of F1-score, accuracy, precision and recall, both models performed well.
Haddad et al., 2019 [23]	T-HSAB (Tunisian dataset)	SVM and NB	Both models performed well, prevailing over the binary categorization. Across all metrics, Model NB performed best.
Ibrohim & Budi, 2018 [24]	Indonesian tweets	Random forest decision tree, SVM, NB	When comparing the F1 score among the three classes, NB outperformed SVM and RFDT by 70.06 percent, and when comparing the two classes, it outperformed them by 86.43 percent.
Fauzi & Yuniarti, 2018 [25]	Indonesian tweets	SVM, RF, Maximum entropy, KNN and NB	The categorization efficiency of collected data may be enhanced by using the ensemble model.
Ozel et al., 2017 [26]	Tweets and Instagram comments in Turkish language.	KNN, NBM, Decision tree and SVM	Naive Bayes Multinomial is the most effective approach for classification, with an accuracy of 84%.
Abozinadah et al., 2015 [27]	Arabic tweets	DT, SVM and NB	According to the findings, the NB model performed the best with an accuracy of 90%.

A. TF/IDF Models

Identifying hate speech often involves using the word frequency (TF) in conjunction with the Inverse Document Frequency (IDF). Input text statistics and the precision of terms in the corpus are taken into account by the TFIDF, which gives more weight to less commonly used words. Examples of hateful speech on Twitter include the repeated use of offensive hashtags like "#BanIslam,","#whitegenocide," and "#whoriental," among others. As a result, an association rule system [20] was developed using hashtags, lexicons and POS (part-of-speech) tags to categorise hateful speech.

B. Lexicon-Based Methods

When it comes to identifying hate speech, lexicon-based approaches rely on pre-existing keyword sets (lexicons) that have been established by writers or culled from the literature. Frenda et al. [14] used SVM classifiers to preprocess two datasets containing sexist and misogynistic tweets. The Automatic Misogyny Identification (AMI) the AMI EvalIta [16], IberEval [15] and the SRW [17] were the datasets utilised by Frenda et al. To help weight lexical aspects, Information Gain [14] was applied to a list of 690 lexicon terms added to the TFIDF for words and characters. Hashtags, slang, acronyms, sexual terminology, terms pertaining to women, and words having to do with the human body are all part of their vocabulary. The accuracy of their model is 0.76. Using the English dataset from the AMI shared issue at EVALITA2018, they use the TFIDF of tweets, sentence embeddings, XGboost and Glove Bow over Logistic Regression. The best accuracy they could get was 0.704% using LR, which made it the most successful approach.

III. HATE SPEECH DETECTION USING DEEP LEARNING MODELS

As a kind of artificial neural network, deep learning models are becoming more popular. These kinds of models are the backbone of the data mining and categorization industries. These models are trained to recognise text patterns. It is clear from Figure2 that the optimal selection of a deep learning model is contingent on the number of layers hidden and the method used to represent the target attributes. In the following, we will examine a few recent literature studies in this area.

A. Feedforward Neural Network

The information in a feeding neural network moves in just one way. As data moves from one layer to the next, it enters at the input layer and leaves at the output layer via hidden layer if there exists. This category of nodes does not form any closed loops. An important feature of a particular neural network architecture is the use of multilayers, together with values and functions calculated in the forward direction of the data flow.

B. Recurrent Neural Network (RNN)

Callbacks and sessions in an RNN are controlled by the processing units in this network architecture. In this way, the output of one layer is fed into the input of the layer below it. In a network, the current layer is typically the sole layer, therefore each consecutive loop gets fed information from the previous layer. The present output is based

on the network's past states, which are stored in a memory. Multiple inputs may be used to generate multiple results in an RNN [13].

C. Long term short Memory (LSTM)

Similar to a Feedforward Neural Network, this one does not keep track of which data points served as the network's principal inputs. Long-term, short-term memory treatment is employed for this issue. The Long Short-Term Memory (LSTM) neural network model is widely used. When training a network in reverse feedforward mode, LSTM is used to suppress states. While training an LSTM, it is necessary to provide memory blocks in which to keep track of the model's prior states. Memory cells may be found inside each memory block. These cells contain the network's temporal states. Input, storage, and output gates allow cells to govern the passage of signals by controlling their timing and strength. Whether information is being read, written, or saved, it is all under the watchful eye of these gates [13].

D. Convolutional Neural Network (CNN)

The field of deep neural networks known as Convolutional Neural Networks (CNNs) finds widespread use in areas such as image processing, video identification, medical picture classification, and natural language processing. They use a multilayer feedforward neural network. Each stage—convolution, nonlinearity, and pooling—makes up the whole [15].

- *Convolution*: at this point, the input undergoes a convolution, a mathematical procedure. The input is processed via a convolution filter to generate feature maps, which are then passed on to the next layer.
- *Nonlinearity*: As part of this process, we will attempt to include nonlinear features into the network. In cases when a nonlinear process, such as ReLu, is applied.
- *Clustering*: It is a process that uses a function like average clustering or maximum clustering to reduce the number of features in a feature map [14].

In Table II, we have summarized a number of research publications together with their methodologies and findings based on deep learning.

TABLE II. DETAILED REVIEW OF SEVERAL EXISTING METHODS BASED ON DEEP LEARNING FOR HATE SPEECH CLASSIFICATION

Paper	Dataset	Approach	Findings
Pradeep Kumar Roy et al., 2022 [42]	Malayam and Tamil mixed dataset	To better identify hateful speech and abusive words on social media, it is recommended to use an ensemble model that combines the results of deep learning and transformer-based models.	Experimental results from the suggested framework achieved 0.933 and 0.802 in weighted F1-score, which is superior to the state-of-the-art.
Qasim Mehmood et al., 2022 [41]	Islamophobic tweets	The suggested method utilises a one-dimensional Convolutional Neural Network and Long Short-Term Memory network inspired classifier for feature extraction.	Over 90% accuracy is observed when the suggested method is tested on a dataset consisting of 1290 pre-processed internet tweets.
Tapan Kumar Das et al., 2021 [40]	Mixed language English-Odia from Facebook	RF, SVM and NB	For both binary and ternary classes, the SVM with word2vec-based models outperformed the NB and RF alternatives. The outcome demonstrates that the tertiary models performed better than the binary ones in differentiating among hate speech and other types of communication.
Areej Al-Hassan et al., 2021 [39]	Arabic tweets	CNN, LSTM and GRU	When compared to the SVM model, all four deep learning models perform better at identifying malicious tweets. The average recall of the deep learning algorithms is 75%, whereas that of the SVM is just 74%. The total performance of detection is enhanced by integrating a layer of CNN to LTSM, from 72% precise to 73% F1 score and 75% recall.
Alshalan et al., 2020 [28]	Arabic tweets	CNN	With an F1-score of 79% and an accuracy of 83%, CNN performed well.

Abuzayed & Elsayed, 2020 [29]	Arabic tweets	15 unique models, ranging from the traditional to the neural.	The neural learning model with an F1-score of 73% was the top performer.
Aref et al., 2020 [30]	Sunnis and Shiites-related tweets	Emedding techniques(word2vec and FastText)	CNN combined with word2vec produced the highest accuracy, 79%. CNN with FastText had the best recall at 69%. RNN combined with Fasttext produced the highest F1-score (52%) possible.
Faris et al., 2020 [19]	Arabic tweets	RNN, CNN, SVM, DT, complement NB and RF LSTM and CNN	The categorization did a good job of separating hateful and neutral tweets. Word2vec performed higher than its competitors, scoring 66.564% on accuracy and 71.688% on the F1 scale.
Ibrohim et al., 2019 [31]	Indonesian tweets	LSTM with word embedding	When combined with word embedding, they discover that LSTM can effectively classify texts in both English and Indonesian. The model's F1 score was 83.68 percent, which is equivalent to 19.44 percent.
Chaudhari et al., 2019 [32]	English tweets	LSTM and hybrid CNN	The user-ids of these people allowed them to be singled out. The report included the users' whereabouts, the amount of people that follow them, and their own personal details.
Mathur et al., 2018 [33]	English-Hindi offensive tweets	Ternary trans CNN	Ternary Trans-CNN performed well on the HEOT dataset. The model attained an F1 Score of 71% and accuracy of 83.90%.
Gambäck & Sikdar, 2017 [34]	Tweets	CNN	CNN did best in F1 with word2vec, which got a score of 78.3%.

IV. Challenges In Hate Speech Detection

Despite the availability of several techniques, there are obstacles in the detection of abusive language or hate speech. One of the biggest problems is that different people use different terms, and a word that one person could consider aggressive or hostile might not be in the same context when another person uses it. Also, it's possible that emoji-expressed speech might reveal how its listener interprets the text. A issue with identifying the content of text may arise, for instance, if it seems as though the text does not include any term that may signify abusive or hate speech, yet it may involve emoji expressing it. The most significant difficulty is the existence of several varieties of the same dialect, such as the Gulf dialect and Levantine dialect. While certain words might be considered offensive or nasty terms in a language or in a particular region, they may be categorised as a regular word for another dialect or region.

V. Conclusion

Social media platforms provide an inclusive environment wherein individuals are afforded the opportunity to actively engage in discussions and share their thoughts, while simultaneously accommodating others who choose to passively consume the content. Due to the unrestricted nature of opinion-sharing on social media platforms, it is of utmost importance to exercise careful scrutiny when engaging with the content disseminated therein. It is imperative to closely monitor the frequency with which these media outlets employ hate speech and other forms of derogatory language. This study presents a comprehensive evaluation of the advancements made in the automated identification of hate speech within textual content, focusing on the current developments. The findings indicate that machine learning techniques, specifically Support Vector Machines (SVM), along with various forms of Term Frequency-Inverse Document Frequency (TF-IDF) features, were initially favoured by researchers. However, with the progression of deep learning technology, there have been significant enhancements in the analysis of hate speech. In the future, we will conduct a comprehensive analysis and comparative evaluation of hate speech detection methodologies across many nations.

References

[1] Theodosiadou, O.; Pantelidou, K.; Bastas, N.; Chatzakou, D.; Tsikrika, T.; Vrochidis, S.; Kompatsiaris, I. Change point detection in terrorism-related online content using deep learning derived indicators. Information 2021, 12, 274.

[2] Mondal, M.; Silva, L.A.; Benevenuto, F. A measurement study of hate speech in social media. In Proceedings of the HT 2017—28th ACM Conference on Hypertext and Social Media, Prague, Czech Republic, 4–7 July 2017; pp. 85–94.

[3] Sanoussi, M.S.A.; Xiaohua, C.; Agordzo, G.K.; Guindo, M.L.; al Omari, A.M.M.A.; Issa, B.M. Detection of Hate Speech Texts Using Machine Learning Algorithm. In Proceedings of the 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 26–29 January 2022; pp. 266–273.
[4] Forestiero, A. Metaheuristic algorithm for anomaly detection in Internet of Things leveraging on a neural-driven multiagent system. Knowl. Based Syst. 2021, 228, 107241.
[5] Ayo, F.E.; Folorunso, O.; Ibharalu, F.T.; Osinuga, I.A. Machine learning techniques for hate speech classification of twitter data: State-of-The-Art, future challenges and research directions. Comput. Sci. Rev. 2020, 38, 100311.
[6] Strossen, N. Freedom of speech and equality: Do we have to choose. JL Pol'y 2016, 25, 185.
[7] Comito, C.; Forestiero, A.; Pizzuti, C. Word embedding based clustering to detect topics in social media. In Proceedings of the 2019 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2019, Thessaloniki, Greece, 14–17 October 2019; pp. 192–199.
[8] MacAvaney, S.; Yao, H.R.; Yang, E.; Russell, K.; Goharian, N.; Frieder, O. Hate speech detection: Challenges and solutions. PLoS ONE 2019, 14, e0221152.
[9] Chetty, N.; Alathur, S. Hate speech review in the context of online social networks. Aggress. Violent Behav. 2018, 40, 108–118.
[10] Paz, M.A.; Montero-Díaz, J.; Moreno-Delgado, A. Hate Speech: A Systematized Review. SAGE Open 2020, 10, 3022.
[11] Matamoros-Fernández, A.; Farkas, J. Racism, Hate Speech, and Social Media: A Systematic Review and Critique. Telev. New Media 2021, 2, 205–224.
[12] Fortuna, P.; Bonavita, I.; Nunes, S. Merging datasets for hate speech classification in Italian. CEUR Workshop Proc. 2018, 2263.
[13] Ibrohim, M. O., & Budi, I. (2018). A dataset and preliminaries study for abusive language detection in Indonesian social media. Procedia Computer Science, 135, 222-229.
[14] Frenda, S.; Ghanem, B.; Montes-Y-Gómez, M.; Rosso, P. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. J. Intell. Fuzzy Syst. 2019, 36, 4743–4752
[15] Fersini, E.; Rosso, P.; Anzovino, M. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. IberEval@SEPLN 2018, 2150, 214–228.
[16] Fersini, E.; Nozza, D.; Rosso, P. Overview of the evalita 2018 task on automatic misogyny identification (ami). EVALITA Eval. NLP Speech Tools Ital. 2018, 12, 59.
[17] Waseem, Z.; Hovy, D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of the NAACL Student Research Workshop, San Diego, CA, USA, 16–17 June 2016; pp. 88–93.
[18] Mubarak, H., Rashed, A., Darwish, K., Samih, Y., & Abdelali, A. (2020). Arabic offensive language on twitter: Analysis and experiments. arXiv preprint arXiv:2004.02192
[19] Aljarah, I., Habib, M., Hijazi, N., Faris, H., Qaddoura, R., Hammo, B., ... & Alfawareh, M. (2020). Intelligent detection of hate speech in Arabic social network: A machine learning approach. Journal of Information Science, 0165551520917651.
[20] Silva, A., & Roman, N. (2020, October). Hate Speech Detection in Portuguese with Naïve Bayes, SVM, MLP and Logistic Regression. In Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional (pp. 1-12). SBC.
[21] Da Costa Abreu, M., & Araujo De Souza, G. (2020). Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata.
[22] Mulki, H., Haddad, H., Ali, C. B., & Alshabani, H. (2019, August). LHSAB: a levantine Twitter dataset for hate speech and abusive language. In Proceedings of the Third Workshop on Abusive Language Online (pp. 111-118).
[23] Haddad, H., Mulki, H., & Oueslati, A. (2019, October). T-HSAB: a tunisian hate speech and abusive dataset. In International Conference on Arabic Language Processing (pp. 251-263). Springer, Cham.
[24] Ibrohim, M. O., Sazany, E., & Budi, I. (2019, March). Identify abusive and offensive language in indonesian twitter using deep learning approach. In Journal of Physics: Conference Series (Vol. 1196, No. 1, p. 012041). IOP Publishing.
[25] Fauzi, M. A., & Yuniarti, A. (2018). Ensemble method for indonesian twitter hate speech detection. Indonesian Journal of Electrical Engineering and Computer Science, 11(1), 294-299.
[26] Özel, S. A., Saraç, E., Akdemir, S., & Aksu, H. (2017, October). Detection of cyberbullying on social media messages in Turkish. In 2017 International Conference on Computer Science and Engineering (UBMK) (pp. 366-370). IEEE.
[27] Abozinadah, E. A., Mbaziira, A. V., & Jones, J. (2015). Detection of abusive accounts with Arabic tweets. Int. J. Knowl. Eng.-IACSIT, 1(2), 113-119.
[28] Alshalan, R., Al-Khalifa, H., Alsaeed, D., Al-Baity, H., & Alshalan, S. (2020). Hate Detection in COVID-19 Tweets in the Arab Region using Deep learning and Topic Modeling. Journal of Medical Internet Research. (Preprint).
[29] Abuzayed, A., & Elsayed, T. (2020, May). Quick and simple approach for detecting hate speech in arabic tweets. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection (pp. 109-114).
[30] Aref, A., Al Mahmoud, R. H., Taha, K., & Al-Sharif, M. HATE SPEECH DETECTION OF ARABIC SHORTTEXT.
[31] Ibrohim, M. O., Sazany, E., & Budi, I. (2019, March). Identify abusive and offensive language in indonesian twitter using deep learning approach. In Journal of Physics: Conference Series (Vol. 1196, No. 1, p. 012041). IOP Publishing.
[32] Chaudhari, S., Chaudhari, P., Fegade, P., Kulkarni, A., & Patil, R. (2019, November 11). Hate Speech And Abusive Language Suspect Identification And Report Generation. International Journal of Scientific & Technology Research. https://www.ijstr.org/finalprint/nov2019/Hate-Speech-And-Abusive-Language-SuspectIdentification-And-Report-Generation.pdf

[33] Mathur, P., Shah, R., Sawhney, R., & Mahata, D. (2018, July). Detecting offensive tweets in hindi-english code-switched language. In Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media (pp. 18-26).
[34] Gambäck, B. and Kumar Sikdar, U., 2017. Using Convolutional Neural Networks To Classify Hate-Speech.
[35] P. William, R. Gade, R. e. Chaudhari, A. B. Pawar and M. A. Jawale, "Machine Learning based Automatic Hate Speech Recognition System," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 315-318, doi: 10.1109/ICSCDS53736.2022.9760959.
[36] R. Ali, U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg. Hate speech detection on twitter using transfer learning. Computer Speech & Language, 74:101365, 2022.
[37] Romim, N., Ahmed, M., Talukder, H., Saiful Islam, M. (2021). Hate Speech Detection in the Bengali Language: A Dataset and Its Baseline Evaluation. In: Uddin, M.S., Bansal, J.C. (eds) Proceedings of International Joint Conference on Advances in Computational Intelligence. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-16-0586-4_37
[38] Khan, Muhammad Moin and Shahzad, Khurram and Malik, Muhammad Kamran, Hate Speech Detection in Roman Urdu, 2021, Association for Computing Machinery, New York, NY, USA,vol 20, issn 2375-4699, https://doi.org/10.1145/3414524
[39] Al-Hassan, A., Al-Dossari, H. Detection of hate speech in Arabic tweets using deep learning. Multimedia Systems 28, 1963–1974 (2022). https://doi.org/10.1007/s00530-020-00742-w
[40] Mohapatra SK, Prasad S, Bebarta DK, Das TK, Srinivasan K, Hu Y-C. Automatic Hate Speech Detection in English-Odia Code Mixed Social Media Data Using Machine Learning Techniques. Applied Sciences. 2021; 11(18):8575. https://doi.org/10.3390/app11188575
[41] Mehmood, Q., Kaleem, A., Siddiqi, I. (2022). Islamophobic Hate Speech Detection from Electronic Media Using Deep Learning. In: Djeddi, C., Siddiqi, I., Jamil, A., Ali Hameed, A., Kucuk, İ. (eds) Pattern Recognition and Artificial Intelligence. MedPRAI 2021. Communications in Computer and Information Science, vol 1543. Springer, Cham. https://doi.org/10.1007/978-3-031-04112-9_14
[42] Pradeep Kumar Roy, Snehaan Bhawal, Chinnaudayar Navaneethakrishnan Subalalitha, Hate speech and offensive language detection in Dravidian languages using deep ensemble framework, Computer Speech & Language, Volume 75, 2022, 101386, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2022.101386.