INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 6, June 2026 | ISSN: 2456-4184 | IJNRD.ORG

Comprehensive Analysis of Machine Learning Algorithms for Cyber Hate Detection on Social Media Platform

Aniruddha S Holey *, Prof. Dr. Swati S Sherekar*
** Department of CSE, SGB Amravati University, Amravati, Maharashtra, India;*

Abstract: Cyber bullying and hate mongering are serious problems on social media platforms. With 4.9 billion people using social media worldwide and 398 million in India, unpleasant behavior such as online abuse, hate speech, trolls, and harassment is becoming more and more of a concern. Particularly, professional women and girls at educational institutions face numerous challenges. There are various approaches that researchers are focusing on to alleviate this issue, such as Proactive, randomized sampling, Robust, Automated detections, Machine learning, Deep learning, and session-based approaches. The increasing demand for secure online communication and digital safety makes this research area essential. Well-known social media sites like Facebook, Instagram, Twitter (X) and YouTube are frequently used for communication and information sharing, but they are sometimes abused to disseminate unpleasant and hateful content. As a result, identifying and stopping hate speech and cyber bulling has grown to be a significant challenge in the domains of digital forensics, artificial intelligence and cyber security. The study of natural language processing and machine learning as a foundation for detection and prevention of cyber bullying and religious content on social media platforms is proposed in this research. Lastly, we identify and talk about the future directions for research in Machine Learning and Natural Language Processing to combat cyber bullying on social media platforms.

IndexTerms **- Cyber Bulling, Cyber Hate, Cyber-Harassment, Social Media, Machine Learning**

I. INTRODUCTION

The expansion of the internet has led to a significant increase in social media use. On social media sites like Facebook, Twitter, YouTube, vine, Wikipedia talk pages, Reddit, Instagram, Pinterest, TikTok, Snap chat, and other platforms. Because there are billions of users on social media these days, there is a decline in in-person communication and a rise in the sharing of thoughts via online portals and social media platforms. Author Fatma Elsafoury specifies the definition [1] "Hate speech is a language that attacks on diminishes, that insights violence or hate against group based on a specific characteristics such as physical appearance, religion, descent, National or ethnic origin sexual orientation gender identity, or others and it can occurs with different linguistic style, even in subtle form or when humor is used." Cyber bullying is a form of cyber aggression that is defined as an intentional, harmful act to another person that takes place through online means and is characterized by imbalance of power between the individuals in the world and reputations of the act. [1] The two definitions of abuse language in cyber bullying that we have knowledge of are outlined as follows: hate speech is directed towards a specific group of humans who share particular features, such as race, behavior, physical appearance, sexuality, socioeconomic class, gender, ethnicity, disability, religion, intoxication, shallowness, etc., whereas abuse language is produced towards someone in particular. There are other ways for offensive behavior (Figure 1) to occur, such as online abuse, which is when someone gets hurt or distressed by using an online platform. Hate speech is defined as any online message that targets a group of people based on their gender, ethnicity, religion, race, or sexual orientation. Hatred of women is demonstrated by misogyny. One way that xenophobia manifests itself is as unreasonable hostility toward foreigners. A troll is someone who reacts erratically to what they see online. Cyber aggression, or online bullying, usually happens infrequently between peers. Cyber bullying is the deliberate use of digital tools or online platforms, such as forums, social networking sites, and smart phone applications to harm the target [2]. Artificial Intelligence, Machine Learning and Natural Language Processing approaches have been shown in recent research to be successful in identifying damaging content and cyber hate on social media platforms. While Ketsbaia et al presented a multistage ML and Fuzzy Logic technique to increase the accuracy of cyber hate identification, Arya et al. provided a multimodal framework utilizing contrastive language, image pre-training for hate speech recognition in memes [16, 17]. Additionally, Deep Learning and Transformer-based model have greatly improved the ability to detect hate speech and cyber bullying in online Social networks. Toktarova et al. assessed both ML and Deep Learning methods for hate speech detection in social media platforms; Obida et al. used deep learning algorithms for cyber bullying detection [18, 19]. The expanding relevance of transformer topologies in improving cyber hate detection systems is further demonstrated by recent developments like the MetaHate transformer framework [20].

This paper is divided into five sections: Section 2 summarizes the different types of cyber bullying, including hate or cyber bullying detection approaches, cyber bullying detection techniques, and machine learning model; Section 3 reviews the literature on machine learning-based detection methods; and Sections 4 and 5 analyze and discuss multiple detection methods. Section 6 wraps up the manuscript and outlines further projects.

figure 1: offensive behavior in social media

II. Cyber hate or bulling categories, detection approach and technique

The classification, methods, detection strategies and Machine learning models of cyber hate or bulling can be discussed in the section below

A. Cyber Hate or Bulling Categories

Cyber Hate or bullying is divided into twelve distinct categories. The purposeful and intentional use of technology to tell people they are not a part of the group and therefore their participation is not necessary is known as exclusion. When someone is denigrated, it is done by writing harsh, vulgar, hateful, cruel, or untrue statements about rumors about them and passing them off to others. Masquerading is the act of posing as someone else to spread harmful or destructive messages while making it seem as though the message originated with that person. For instance, breaking into a victim's email account and sending these messages right away [2]. At the very least, flame-blazing or battering involves users attacking one another physically and/or verbally. Organizing and providing in a driver-only urban setting via email, twitter, as well as social media sites and discussion boards [2, 5]. People often use capital letters to show their anger. There are a lot of burning textures that are harsh, nasty, and unjustifiable. Cyber stalking is the practice of using social media to harass, threaten, or stalk specific people, groups, or organizations [5]. Trolling or baiting is the deliberate posting of remarks that aim to incite conflict with other participants in the conversation. Similar to denigration, "outing" involves writing and disclosing embarrassing or humiliating personal information in public and calls for a close personal contact between the bully and the victims, either in person or on social media. This information may include accounts that you have heard from the victim or any private data you may have, such as address, passwords, and phone number [1, 2,5].

Threatening behavior based on an individual's age, gender, sexual orientation, and other factors is referred to as harassment, or harassment on social media. Sending brief messages that are menacing, violent, and full of threats is one example of a cyber-thread. Cat-fishing is the practice of building a false profile using someone else's details. These accounts, in particular, are created on email services or social networking sites [1]. Dissing someone is posting false information about them in an attempt to discredit them. Tricking someone into disclosing personal information or secrets. [1] Frapping is the practice of posting content on another person's internet account while deceiving others into thinking the original account holder posted it [1].

B. Cyber Hate or Bulling Detection Approaches

Recently, several sentiment-based techniques have been published to detect and identify abusive language. These techniques include machine learning, lexicon-based techniques and hybrid methods, as illustrated in figure 3. Supervised machine learning, unsupervised machine learning, semi-supervised learning, and deep learning are the methodologies that make up machine learning. With a set of pre-annotated training texts, supervised machine learning is used to automatically identify the features of classes or categories [3]. Uncovering and understanding the hidden structure in unintelligible data requires supervised machine learning. Semi-supervised learning is the development of an algorithm that takes benefit of the combination of labeled and unlabeled information to analyze the behavior of learning. Deep learning, a rapidly emerging field in machine learning, is inspired by artificial neural networks [5].

A dictionary is created using a vocabulary-based technique, and terms are then searched and tallied in text. The efficacy of the method in terms of classification may be limited by the use of domain-specific terms in dictionaries. A corpus-based technique incorporates sentiment levels and context into data-driven models, identifying new sentiment words and their polarity from a vast collection of centimeter-words with three different polarities. Word-hoard, WorldNet, Aretha, and other Lexicographical tools are utilized in the dictionary-based approach. Combining lexicon-based technique and machine learning approach results in a hybrid approach. [2]

figure 2: cyber hate or bulling categories

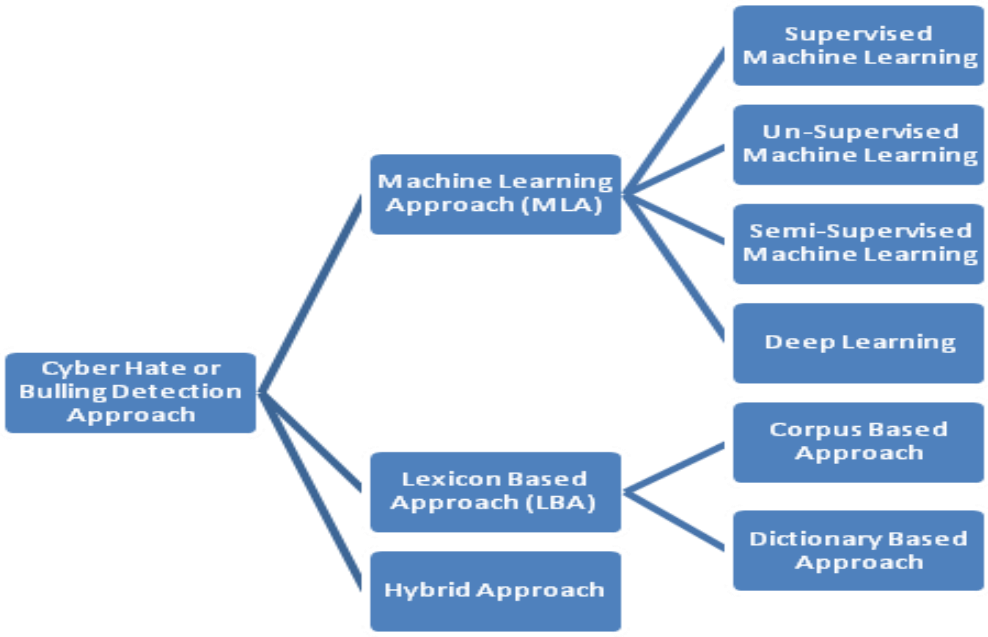

figure 3: cyber hate or bulling detection approaches

C. Cyber Hate or Bulling detection techniques

Text mining and analysis have grown in popularity and activity, and text analysis is based on the quantity of classes in these datasets. Bullying and cyber hatred tasks can be carried out in a binary or multiclass manner.

1) Binary Cyber hate or bulling classification: The detection of cyber bullying and hatred has been studied as a binary classification job, such as "cyber bullying vs. non-cyber bullying" or "hate vs. non-hate"[5].

2) Multi class cyber hate or bulling classification: Several researches have been conducted to categorize bullying and cyber hatred into various groups [2].

D. Machine Learning Models

Less dataset for training are available for the model that uses rule-based learning to classify the data. Cyber hate and bullying used seem small in quantity and scale, however deep learning models employed for their detection require a vast number of data points for training. The traditional machine learning paradigm, which excels at text categorization

tasks, is primarily employed in supervised learning models. Unconventional machine learning models offer a unique approach to feature engineering or labeled data sets [1, 4].

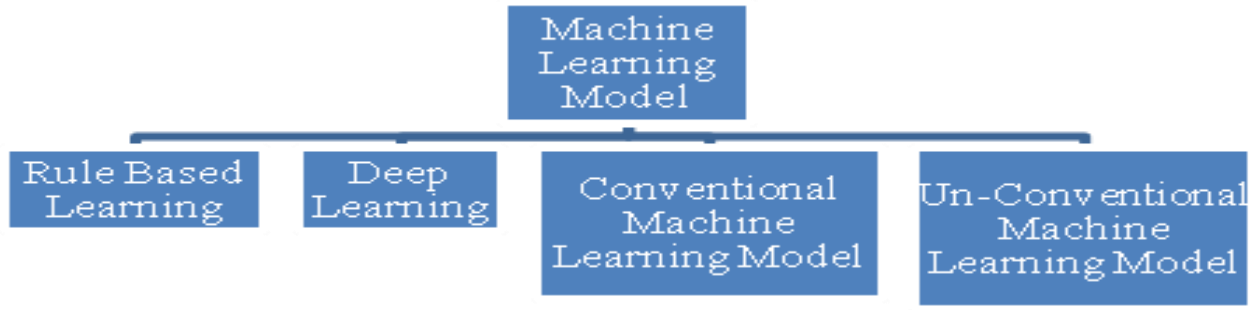

figure 4: machine learning model

III. Background/Literature Review

The utilization of social media by users has risen drastically worldwide, which has resulted in an increase in cyber bullying and hate crimes. Social media use by women and children is extremely detrimental, and as a result, there is more litigation on the site. We must stop harassment and xenophobia on social media platforms. Following are a few techniques that the researcher presented.

Sadiq et al. [6] presented a method for finding the cyber aggressive using the CNN, LSTM and Bi-LSTM algorithms in Machine Learning approach with the accuracy of 92 percent on Twitter Dataset troll. *Beyhan et al. [7]* developed a hate speech detection system using BERTurk using Transform architecture implemented on Istanbul Convention dataset and Refugee Dataset collected data, the system classification accuracy is 77% on Istanbul Convention dataset. This study does not include the URLs, Has-tags, Mentions, news related, media and Journalists account data in the dataset.

*Romim et al.[8]*uses a baseline experiment and Several deep learning as well as extensive pre-trained Bengali word embedding such as Word2Vec,FastTest, and BengFastText collected dataset on Facebook and Youtube Comments. The given result deep learning models perform well; SVM achieved good results with the accuracy of 87.5%. It can be challenging at times to comprehend the correct context of angry phrases in Bengali language through this research.

Karim et al [9] proposed the approach for detection of hate speech in Bengali Language. Political, personal, geographical, and religious hatred are the classification methods used to categorize the Bengali hate speech dataset. neural ensemble of various transformer-based neural architectures (e.g., multilingual BERTcased and uncased, XLM¬RoBERTa, and monolingual Bangla BERTbase), after which significant terms are identified using sensitivity analysis and layer-wise relevance propagation (LRP) to produce explanations that are understandable to humans. Threefold cross-validation tests yield F1 scores of 84%, 90%, 88%, and 88% for political, personal, geopolitical, and religious hates, respectively, when evaluations are conducted against a variety of machine learning (linear and tree¬ based models) and deep neural network (i.e., CNN, Bi¬LSTM, and Conv¬LSTM with word embeddings).

AlBayari & Abdallah et al [10] utilized their dataset to train the most fundamental classifiers like LR, SVM, RFC, and Multinomial Naive Bayes (MNB) for the identification of cyber hate speech or bullying. The SVM classifier is therefore a better option because it has a much higher F1-score value of 69% than the other classifiers.

Patil et al [11] examined the identification of hate speech in the regional Indian language of Marathi. The largest dataset of Marathi hate speech that is available to the public is the L3Cube-MahaHate Corpus, which they presented. Four fine-grained labels were applied to the dataset, which was collected from Twitter: None, Profane, Offensive, and Hate. More than 25,000 samples in the dataset have had the classifications manually assigned to them. They conducted tests using a variety of deep learning models, including transformer-based BERT, CNN, LSTM, and BiLSTM. With an accuracy of 80.3%, the BERT model performed better than the other models.

Atoum et al [12] gathered Datasets 1 and 2 from Twitter. The Twitter dataset includes both non-cyberbullying and cyber bullying tweets. They created and improved an effective technique that makes use of language models and sentiment analysis to identify cyber bullying in tweets. Using two twitter datasets, machine learning methods are analyzed and contrasted. Higher n-gram language models in CNN classifiers fared better than those in DT, RF, NB, and SVM. CNN classifiers have 93.62% and 91.03% accuracy on average.

Nabilah et al.[13]suggested a dataset of noxious remarks that had been gathered, handled, and tagged by hand. Data is collected from comments made by Indonesian users on social media sites including Instagram, Twitter, and Kaskus. These platforms provide multi-label features that make it possible to classify users into multiple groups. Defamation, radicalism, hate speech, and pornography are all present in the dataset. Pre-trained model trained for Indonesian to identify harmful sentence content in comments on Indonesian social media platforms. This work employed the Multilingual BERT (MBERT), IndoBERT, and Indo Roberta Small models to assess the classification outcomes and carry out a multi-label classification. The study's top results were obtained by 88.97%.

Moratanch N et al [14] propose a supervised machine learning strategy that uses three classification algorithms—random forest, SVM, and logistic regression—to identify and stop cyber bullying. Words related to cyber bullying are recognized with 92% accuracy of the random forest classifiers.
Yengejeh a et al [15] provide a machine learning model for the automatic detection of cyber bullying. This model makes use of classifiers such as logistic regression (LR), Multinomial Navie Bayes (MNB), k-nearest neighbor (KNN), and Extreme Gradient boosting (XGboost) on a dataset of textual data from Twitter. It shows that the XGboost classifier has the highest accuracy value for the TF-IDF and BOW properties.

IV. Analysis

Table 1 provides a summary of the literature review discussed in the previous section. It is based on parameters such as category, classes, social media platform data collected, cyber hate detection approach, dataset, algorithms, and tool s used by the researcher for implementation. It also considers the pros and cons of the approach and determines performance based on various parameters.

table 1: overview of scholarly research on machine learning-facilitated cyber bullying and hateful behavior detection

Author/ Year	Category	Classes (Number of Classes)	Social Network Platform	Approach	Algorithms	Datasets	Evaluation Matrix/ Parameters		Tools	Advantages	Diss-advantages
Sadiq et al. 2021 [6]	Trolling and Harassment	2 Cyber Aggressive, Non Cyber Aggressive	Twitter	MLA	CNN+ LSTM+ Bi-LSTM	Cyber Trolls	Accuracy Precision Recall F1 Score	92% 90% 90% 90%	Keras Python, DL Framework, Jupiter Notebook	----	----
Beyhan et al. 2021[7]	Trolling and Harassment	2 Country/Nationality, Race/Ethnicity (Binary)	Twitter	MLA	Transformer Architecture BERTurk	Trukish-HS-Dataset (Istanbul Convention dataset)	Accuracy F1 Score Precision Recall	76.70% 77.90% 78.20% 77.68%	---	----	This study not including the URLs, Has-tags, Mentions, news related, media and Journalists account data in dataset.
						Trukish-HS-Dataset (Refugee Dataset)	Accuracy F1 Score Precision Recall	73.80% 64.95% 70.19% 60.51%			
		3 Sexual orientation, Religion, Gender (Multi-Class)				Trukish-HS-Dataset (Istanbul Convention dataset)	Accuracy F1 Score Precision Recall	71.52%			
						Trukish-HS-Dataset	Accuracy F1 Score				

						(Refugee Dataset)	Precision Recall	72.34			
Romim et al.[8]	Trolling and Harassment	2 Hateful, Non Hateful	Facebook and Youtube	MLA (DLM)	SVM	Online Dataset	Accuracy	87.5 %	---	---	This study some time difficult to understand the proper context of aggressive words in Bengali Language.
					Word2Vec+ LSTM			83.85%			
					Word2Vec+ Bi-LSTM			81.52%			
					FastText + LSTM			84.3 %			
					FastText + Bi-LSTM			86.55%			
					BengFastText+LSTM			81%			
					BengFastText+Bi-LSTM			80.44%			
Karim et al 2021[9]	Trolling and Harassment	2 Hateful,Non-Hateful	Facebook, Youtube comment, Newspaper	MLA	LR	Bengali Hate Speech Dataset	F1-Score	67%	Scikit-learn, Keras, and PyTorch with TensorFlow backend, fastText for embedding.	----	This model contains limited amount of labeled data to train the model.
					NB			64%			
					SVM			66%			
					KNN			66%			
					RF			68%			
					GBT			68%			
					CNN			73%			
					Bi-LSTM			75%			
					Conv-LSTM			78%			
					Bangla BERT			86%			
					mBERT-cased			85%			
					XML-RoBERTA			87%			
					mBERT-uncased			86%			
					Ensamble			88%			

AL Bayari and Abdallah 2022[10]	Positive, Negative and Neutrial	2 Cyber bulling, Non-Cyber bulling	Instagram	MLA	MNB	Instagram based Benchmark Dataset	F1-Score (Accuracy)	66%	---	This is the first Arabic-language dataset on cyber bullying from Instagram.	---
					RF			67%			
					SVM			69%			
					LR			66%			
Patil et al.2022[11]	Trolling and Harassment	4 Hate, Offensive, Profane, None	Twitter	MLA	CNN	L3Cube-MahaHate Corpus large Marathi dataset	Accuracy	75.1 %	---	This is good Marathi dataset available by L3Cube.	---
					LSTM			75.1 %			
					Bi-LSTM			76.1 %			
					BERT			80.3 %			
Atoum 2023[12]	Trolling and Harassment	2 Cyber bulling, Non-Cyber bulling	Twitter	MLA	CNN classifier with higher N-gram language model	Twitter Dataset 1	Accuracy	93.62%	---	This model performs better than DT, RF, NB and SVM Machine learning classifier.	----
						Twitter Dataset 2		91.03%			
Nabilah et al. 2023 [13]	Trolling , Harassment and Flaming	4 Pornography, Hate Speech, Radicalism, Defamation	Instagram, Twitter , Kaskus	MLA	IndoBERT	Indonesian comments Instagram, Twitter, Kaskus Dataset	F1-Score	88.97%	Python Prog. Lang. Google Colab Pro	---	Model is useful only on Indonesian language dataset.
Moratamchn et al. 2023[14]	Cyber Bulling and Non Cyber Bulling	2 Addresses(Hashtag), Data Characteristic	Social Media Platform , Online Forum, Surveys	MLA (Supervised ML)	SVM	Online Dataset	Accuracy	62%	Python, Web Technology and FLASH Framework	This system provides better safety to women's from cyber bulling.	---
					LR			81%			
					RFA			92%			
						Twitter	Accuracy	0.96			

Yengejeh Amir et al. 2024 [15]	Cyber Bulling and Non Cyber Bulling	3 Race, Religion and Non Cyber bulling	Twitter	MLA	Bow + LR	Dataset on COVID-19 Pandemic	F1-Score	0.96	---	XGBoost is proper for Cyber Bulling detection.	---
							Accuracy	0.96			
					TF-IDF + LR		F1-Score	0.96			
							Accuracy	0.88			
					Bow + MNB		F1-Score	0.87			
							Accuracy	0.86			
					TF-IDF + MNB		F1-Score	0.85			
							Accuracy	0.86			
					Bow + KNN		F1-Score	0.86			
							Accuracy	0.46			
					TF-IDF + KNN		F1-Score	0.40			
							Accuracy	0.97			
					Bow + XGBoost		F1-Score	0.97			
							Accuracy	0.97			
					TF-IDF + XGBoost		F1-Score	0.97			

V. Discussion

In this research, we have developed several techniques for identifying cyber bullying and hate speech using a machine learning methodology. The majority of the social media platforms from which the dataset was gathered were Twitter and Facebook. Research is categorized based on the amount of classes that are available; binary and multi-class classifiers are two examples of the classes that are provided, and methods are categorized based on the classes that are present in the dataset. The four assessment matrices—Accuracy, Precision, Recall, and F1-Score—can be used to assess performance. The majority of approaches primarily compare the assessment matrix's accuracy and F1-score. We outline the author's suggested strategy with a performance algorithm based on Accuracy and F1-score given in Table 2.

table 2: performance of algorithms based on accuracy and f1-score

Author	Classes Classification	Algorithms	Evaluation Matrix	
Sadiq et al. 2021 [6]	Binary Classification	CNN + LSTM + Bi-LSTM	Accuracy	92%
Beyhan et al. 2021[7]		Transaction Architecture BERTurk Istanbul Convention Dataset	Accuracy	76.70%
Beyhan et al. 2021[7]		Transaction Architecture BERTurk Refugee Convention Dataset	Accuracy	73.80%
Romim et al.[8]		SVM	Accuracy	87.5%
Karim et al 2021[9]		Ensamble	F1-Score	88%
AL Bayari and Abdallah 2022[10]		SVM	F1-Score (Accuracy)	69%
Atoum 2023[12]		CNN	Accuracy	93.62%
Moratamch N et al. 2023[14]		RFA	Accuracy	81%
Beyhan et al. 2021[7]	Multi-Class Classification	Transaction Architecture BERTurk Istanbul Convention Dataset	Accuracy	71.52%
Beyhan et al. 2021[7]		Transaction Architecture BERTurk Refugee Convention Dataset	Accuracy	72.34%
Patil et al.2022[11]		BERT	Accuracy	80.3%
Nabilah et al. 2023 [13]		IndoBERT	F1-score	88.97%
Yengejeh Amir et al. 2024 [15]		BOW+LR	Accuracy	96%

The performance of algorithms on binary classification using various datasets on various social media platforms is displayed in *Table 2*. Three of the eight comparisons below do better than the rest. Between 87 to 94 percent of the three algorithms mentioned above— CNN + LSTM + Bi-LSTM, CNN, and SVM—are based on accuracy and Ensamble is based on F1-score 88%. A comparison of multi-class classification methods on various platforms and datasets is shown in *Table 2*. There are six comparisons in the provided figure, three of which work nicely. The accuracy of the three aforementioned algorithms BERT, BOW + LR, and TF-IDF+LR ranges from 80 to 96% and IndoBERT is 88.97% F1-score. In Table 3, several approaches are compared based on supported and non-supported criteria such as the dataset, social media platform, evaluation matrix, existing and suggested algorithms, and so on. As the table illustrates, the majority of datasets are used to locate hate speech on social media platforms in order to evaluate performance, primarily using the F1-score and accuracy.

table 3: comparative analysis of several approaches

Author	Dataset	Social Media Platform	Existing Algorithms	Proposed Algorithms	Evaluation Matrix			
					Accuracy	Precision	Recall	F1-Score
Sadiq et al. 2021 [6]	√	√	√	X	√	√	√	√
Beyhan et al. 2021[7]	√	√	√	X	√	√	√	√
Romim et al.[8]	√	√	√	X	√	X	X	X
Karim et al 2021[9]	√	√	√	√	√	X	X	√
AL Bayari and Abdallah 2022[10]	√	√	√	X	√	X	X	√
Patil et al.2022[11]				X		X	X	X

	√	√	√		√			
Atoum 2023[12]	√	√	√	X	√	X	X	X
Nabilah et al. 2023 [13]	√	√	**X**	√	×	X	X	√
Moratamch N et al. 2023[14]	√	√	√	X	√	X	X	X
Yengejeh Amir et al. 2024 [15]	√	√	√	X	√	X	X	√

Conclusion

In this paper, we examined the existing research on detecting hate speech through machine learning techniques applied to online platforms and various social media sites using binary or multi-class language datasets. Most authors utilize existing algorithms for identifying hate speech within the dataset. The significant body of work focusing on cyber hate or bullying detection has been conducted in the English language. Further investigation is essential to develop effective detection strategies for regional languages such as Marathi, Bengali, Hindi, Arabic, and others. In addition, the majority of annotation tasks currently require manual completion; however, this needs to transition to an automated process to enhance outcomes.

Acknowledgement

I would like to express sincere gratitude to the guide, faculty members and institution for their valuable support, guidance and encouragement throughout this research.

References

[1] Elsafoury, F., S. Katsigiannis, Z. Pervez and N. Ramzan. 2021. When the Timeline Meets the Pipeline: A Survey on Automated Cyberbullying Detection. IEEE Access, 9: 103541-103563. DOI: https://doi.org/10.1109/ACCESS.2021.3098979
[2] Sahana, V., et al. 2023. A Systematic Literature Review on Cyberbullying in Social Media: Taxonomy, Detection Approaches, Datasets and Future Research Directions. International Journal on Recent and Innovation Trends in Computing and Communication, 11(10): 406-430. URL: https://ijritcc.org
[3] Sultan, D., et al. 2023. A Review of Machine Learning Techniques in Cyberbullying Detection. Computers, Materials and Continua, 74(3): 5625-5640. DOI: https://doi.org/10.32604/cmc.2023.032412
[4] Yi, P. and A. Zubiaga. 2023. Session-Based Cyberbullying Detection in Social Media: A Survey. Online Social Networks and Media, 36: 100250. DOI: https://doi.org/10.1016/j.osnem.2023.100250
[5] Gamal, D., M. Alfonse, S.M. Jiménez-Zafra and M. Aref. 2023. Intelligent Multi-Lingual Cyber-Hate Detection in Online Social Networks: Taxonomy, Approaches, Datasets and Open Challenges. Big Data and Cognitive Computing, 7(2): 58. DOI: https://doi.org/10.3390/bdcc7020058
[6] Sadiq, S., A. Mehmood, S. Ullah, M. Ahmad, G.S. Choi and B.W. On. 2021. Aggression Detection through Deep Neural Model on Twitter. Future Generation Computer Systems, 114: 120-129. DOI: https://doi.org/10.1016/j.future.2020.07.034
[7] Beyhan, F., B. Çarık, A. İnanç, A. Terzioğlu, B. Yanikoglu and R. Yeniterzi. 2022. A Turkish Hate Speech Dataset and Detection System. Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, pp: 4177-4185. URL: https://aclanthology.org/2022.lrec-1.447
[8] Romim, N., M. Ahmed, H. Talukder and S. Islam. 2021. Hate Speech Detection in the Bengali Language: A Dataset and Its Baseline Evaluation. Proceedings of the International Joint Conference on Advances in Computational Intelligence, Singapore, pp: 457-468. DOI: https://doi.org/10.1007/978-981-16-2164-2_38
[9] Karim, M.R., S.K. Dey, T. Islam, S. Sarker, M.H. Menon, K. Hossain, M.A. Hossain and S. Decker. 2021. DeepHateExplainer: Explainable Hate Speech Detection in Under-Resourced Bengali Language. Proceedings of the 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), Porto, Portugal, pp: 1-10. DOI: https://doi.org/10.1109/DSAA53316.2021.9564187
[10] AlBayari, R. and S. Abdallah. 2022. Instagram-Based Benchmark Dataset for Cyberbullying Detection in Arabic Text. Data, 7: 83. DOI: https://doi.org/10.3390/data7060083
[11] Patil, H., A. Velankar and R. Joshi. 2022. L3Cube-MahaHate: A Tweet-Based Marathi Hate Speech Detection Dataset and BERT Models. Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022), Gyeongju, Republic of Korea, pp: 1-9. URL: https://aclanthology.org/2022.trac-1.1

[12] Atoum, J.O. 2023. Detecting Cyberbullying from Tweets through Machine Learning Techniques with Sentiment Analysis. In: Advances in Information and Communication. Arai, K. (Ed.), Springer Nature, Cham, Switzerland, pp: 25-38. DOI: https://doi.org/10.1007/978-3-031-28073-3_3
[13] Nabiilah, G.Z., S.Y. Prasetyo, Z.N. Izdihar and A.S. Girsang. 2023. BERT Base Model for Toxic Comment Analysis on Indonesian Social Media. Procedia Computer Science, 216: 714-721. DOI: https://doi.org/10.1016/j.procs.2022.12.186
[14] Moratanch, N. 2023. Cyber Bullying Scrutiny for Women Security in Social Media. International Research Journal of Modernization in Engineering Technology and Science (IRJMETS). DOI: https://doi.org/10.56726/IRJMETS34819
[15] Alipour Yengejeh, A. 2024. Combating Cyberbullying on Social Media: A Machine Learning Approach with Text Analysis on Twitter. Data Science and Data Mining, 15. URL: https://scholarworks.calstate.edu
[16] Arya, G., M.K. Hasan, A. Bagwari, N. Safie, S. Islam, F.R.A. Ahmed, A. De, M.A. Khan and T.M. Ghazal. 2024. Multimodal Hate Speech Detection in Memes Using Contrastive Language-Image Pre-Training. IEEE Access, 12: 22359-22374. DOI: https://doi.org/10.1109/ACCESS.2024.3361322
[17] Ketsbaia, L., B. Issac, X. Chen and S.M. Jacob. 2023. A Multi-Stage Machine Learning and Fuzzy Approach to Cyber-Hate Detection. IEEE Access, 11: 56046-56064. DOI: https://doi.org/10.1109/ACCESS.2023.3282834
[18] Obaida, M.H., S.M. Elkaffas and S.K. Guirguis. 2024. Deep Learning Algorithms for Cyberbullying Detection in Social Media Platforms. IEEE Access, 12: 76901-76915. DOI: https://doi.org/10.1109/ACCESS.2024.3406595
[19] Toktarova, A., D. Syrlybay, B. Myrzakhmetova, G. Anuarbekova, G. Rakhimbayeva, B. Zhylanbaeva and M. Kerimbekov. 2023. Hate Speech Detection in Social Networks Using Machine Learning and Deep Learning Methods. International Journal of Advanced Computer Science and Applications, 14(5). DOI: http://dx.doi.org/10.14569/IJACSA.2023.0140587
[20] Chapagain, S., S.M. Hamdi and S.F. Boubrahimi. 2025. Advancing Hate Speech Detection with Transformers: Insights from the MetaHate. arXiv Preprint arXiv:2508.04913. URL: https://arxiv.org/abs/2508.04913