Hate Speech and Cyber bullying Detection on Social Media: A Systematic Review of Methods, Datasets, and Challenges

Sheetal Gawande [1*] and Sharvari Govilkar[2]

Department of Computer Engineering, Pillai College of Engineering, New Panvel , Navi Mumbi , Maharastra , India

1sheetalp@mes.ac.in, 2sgovilkar@mes.ac.in

Abstract. Social Media is an integral part of everyday life, influencing shopping trends, business decisions, political discourse, and social awareness. These platforms are founded on the principle of freedom of expression, empowering people for opinions and ideas. On the darker side, this openness is exposed to Hate Speech, cyberbullying, which lead to serious real-world consequences such as psychological harm, social polarization, erosion of trust, and even violence. Just as responsible citizens contribute to physical safety, ensuring cyber safety and digital well-being within future smart cities is the roles of responsible netizens. This survey aims at the role of Natural Language Processing (NLP) to safeguard this digital landscape. The study systematically reviews research papers from IEEE, ACM, Elsevier, and other reputable sources, providing valuable insights, key challenges in this domain.

Keywords: Social Media, Natural Language Processing (NLP), Cyber Crime, Future Cities.

1 Introduction

Social Media like Faceook, X (formerly Twitter), YouTube, and Instagram have become an integral part of life. According to various studies, people are increasingly active on these platforms for multiple purposes, including business promotion, entertainment, political engagement, and social awareness, which has resulted in an enormous volume of user-generated textual data. With the advancement of technology, users are now able to create and share text, images, audio, and videos. According to Social Media Statistics[1], Worldwide 5.41 billion i.e 63.9% people use social media. According to a survey people spend 2 hours daily on Social Media, after every six seconds new users join social media, showing the influence of digital communication in modern society and revolution in future cities.

[1] https://backlinko.com/sociaal-media-users.

P. Goyal et al. (eds.), *Proceedings of the Conference on Technologies for Future Cities (CTFC 2025)*, Atlantis Highlights in Sustainable Development 7,
https://doi.org/10.2991/978-94-6239-650-0_6

The basic principle of freedom of expression which is considered as human right by law enforced by the United Nations[2] empowering individuals to express opinions, share ideas, and engage in dialogue across borders on social media.[7]. It has enabled the new forms of digital crimes which are varied in their objectives, and modus operandi, to be viewed as a "crime hyponym" in the digital ecosystem [53]. So to get insight this foundation is essential for robust model development for online threat detection. Fig.1 illustrates the categorization of various online threats that can happen through various Social Media.

Social media platforms have both advantages and disadvantages. It serves as powerful tools for information sharing, collaboration, community engagement, and public awareness, enabling citizens to stay connected and informed in real time. However, the same platforms also pose significant challenges, including the spread of misinformation, privacy breaches, and the negative impact on mental health [52].

The spread of false or misleading content on social media erodes public trust, damages the reputation of individuals and organizations, and can fuel extremism, polarization, panic, and fear within society[52]. This survey aims to (1)explore Natural Language Processing Techniques to enhance cyber safety, (2) transform textual data from online social media platforms into actionable insights for threat detection. (3)Aims to analyze existing research on NLP-driven approaches for identifying harmful content, such as cyberbullying and hate speech, challenges, and future directions toward building safer digital environments.

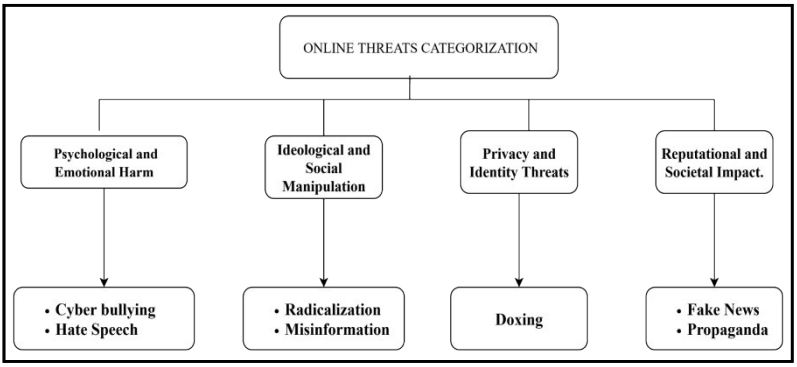

Fig. 1. Online Threats Categorization

Overall, in this study article, we try to cover the existing literature about cyberbullying and hate speech, the Methodology used, Multilinguality, ML model, DL Model, and NLP technology. The research publications from 2018 to 2025 are included. To maintain the quality, only papers with a minimum of two citations were selected. The literature search was conducted using the keywords cyberbullying, social media, Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL). These keywords were chosen to maintain relevance. The paper is organised as follows: Social Media, various threats from social media, and its impact on Future Cities are de-

[2] https://en.wikipedia.org/wiki/Freedom_of_speech

scribed in Section I , Hate Speech and Cyberbullying terminology in Section II , Section III content, and Social Media Content Reporting Policies. Taxonomical categorisation and detailed review of existing methodology are given in sections IV and V, respectively. Section VI concludes with a summary, Challenges, and future directions

2 Hate Speech and Cyberbullying Terminology

Abuse: Any action that causes harm to any person intentionally.
Sexism: Posting harmful content related to sex.
Racism: Treating or targeting some group of people based on their religion, cast, culture.
Cyberbullying: Cyberbullying includes posting negative, harmful, false, or mean content (personal information) about someone else on a digital platform.

3 Social Media Content Reporting Policies[7][3]

Facebook (Meta): Facebook has a different standard policy for content called Community Standards and standard policy for advertising called Advertising Standards. Core Violations (Community Standards) include the following standard: Violence and Criminal Behavior, Safety(Suicide, self-injury, bullying, harassment), Objectionable Content such as hate Speech, nudity/sexual activity, Integrity and Authenticity include Spam, false news (misinformation), impersonation, and fraud/scam.
X (formerly Twitter): Content may be subject to removal, labeling, or other enforcement actions for violating rules, which cover safety, privacy, authenticity, and Integrity.

4 Taxonomical Categorization of Existing Research

4.1 CyberBullying

Bullying, deliberately harming and humiliating others, on their physical appearance, body colour, cast, ethnicity [48]. and young adults are the target.[48], once restricted to the neighborhood and school, has now moved into the digital realm.[37].
As per the survey, College students experience cyberbullying on Social Media. 19% of students have been bullied, and 69% of students have witnessed cyberbullying. Now, the scenario is changed; the massive volume of data is generated on social media. Cyberbullying detection from such massive and unstructured data manually is not possible. To mitigate the psychological and social impact automatic detection mech-

[3] https://www.facebook.com/privacy/policy/

anism is essential. Researchers are working on it and trying to identify the optimal and effective solution for the same. Automatic detection mechanisms help flag, report harmful content to maintain safe online environments on various Social Media Platforms. In the presented survey, we studied the 23 research papers for cyberbullying detection.

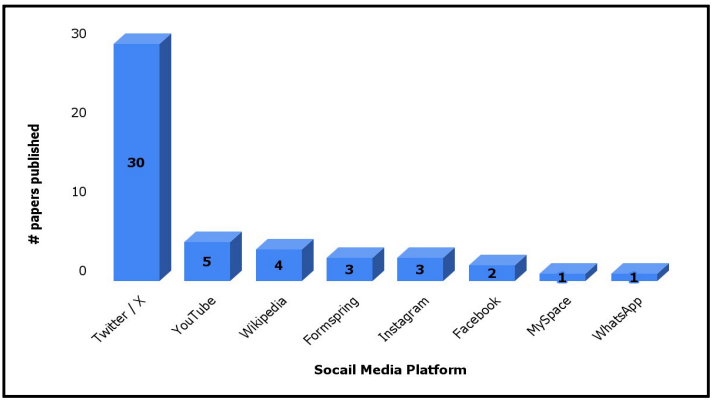

Fig. 2. Targeted Social Media

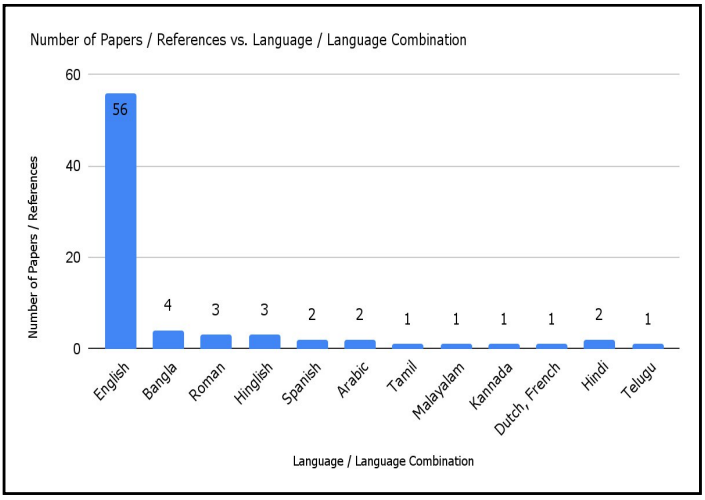

Fig. 3. Most Commonly Used Languages

4.1.1 Social Media Platform, Datasets and Languages

Researchers from various institutions collaborated to develop multiple datasets collected from different social media platforms in different languages. Several researchers collected the data and created datasets. Some of them used kaggle datasets. According to our study researchers used the X(Twitter) data sets with the English Language for experimentation [14,9] as shown in Fig.2 and Fig. 3, respectively. Several of these research groups worked on languages other than English. Researchers analyzed the

collected datasets and categorized words into multiple semantic and contextual classes. This classification helped in detecting cyberbullying by distinguishing abusive terms. Table 1 summarizes the datasets, platforms, languages, and classification types used by the researchers.

Table 1. Dataset used to detect Cyberbullying

Dataset Source (Platform)	Total Size (Number of Posts/Comments)	Language(s)	Categories
Twitter (Primary Source/Kaggle)[2]	47,692 tweets	English	Cyberbullying, and Not Cyberbullying.
Twitter (Kaggle)[15][39]	39,870 posts/comments	English	Multi-class: Religion, Age, Gender, Ethnicity, bullying, and Non-cyberbullying.
Twitter (Collected via snscrape)[11][37]	16,851 tweets	English	Binary: Cyberbullying (Yes) and Non-cyberbullying (No).
Twitter (Kaggle)[5]	26,835 tweets	English	Binary: Offensive and Non-offensive.
Twitter (Cited Research)[20]	~37,373 tweets	Global	Binary classification for cyberbullying detection.
Wikipedia Attack Dataset[33]	115,864 user comments	English	Binary (2 categories): Attack and Non-Attack
Wikipedia Web Toxicity Dataset[41]	159,689 comments	English	Binary: Toxic and Non-Toxic.
YouTube Comments [32]	5,000	Bangla	Binary (2 categories): Bullying and Non-Bullying.
YouTube Comments [33]	7,000	Romanized Bangla	Binary: Bullying and Non-Bullying.
Combined/Multilingual Dataset[37][38][4]	(15,307 English + 3,000 Hinglish)	English & Hinglish	Toxic (offensive) and Non-toxic (non-offensive).

4.2 Hate Speech

Different authors define hate speech differently. Broadly, it is one type of communication where offensive and harsh language targets a group of users based on different attributes like race, religion, gender, or ethnicity. Social media platforms are the backbone of this research because of the large network and diversity of text data in terms of text, images, video, and memes. For this survey we studied the 28 papers latest by 2025. In 2012, researchers collected the corpus from Yahoo! and American Jewish Congress (AJC) for the detection of hate speech for the first time. The template-based strategy used for feature extraction from the corpus and Support Vector Machine(SVM) , a ML model, was trained for classification.[2]. The sources consistently

indicate which platforms are most often targeted for collecting datasets used in hate speech detection research. Twitter is overwhelmingly cited as the most common platform for dataset collection due to data availability.[3] as shown in Table 2.

Table 2. Most Frequently used Social Media

Social Media	**Frequency in Sources**
Twitter (or X)[18]	Highest Frequency
Facebook[18][44]	High Frequency
YouTube[4]	Medium Frequency
Reddit[25]	Medium Frequency
Instagram[17]	Low Frequency

4.2.1 Multilingual Hate Speech Datasets.

As per the survey, researchers study 11 distinct languages, most of which are high resource languages and very few are low-resource languages. The following figure Fig.4.visualizes the statistics.

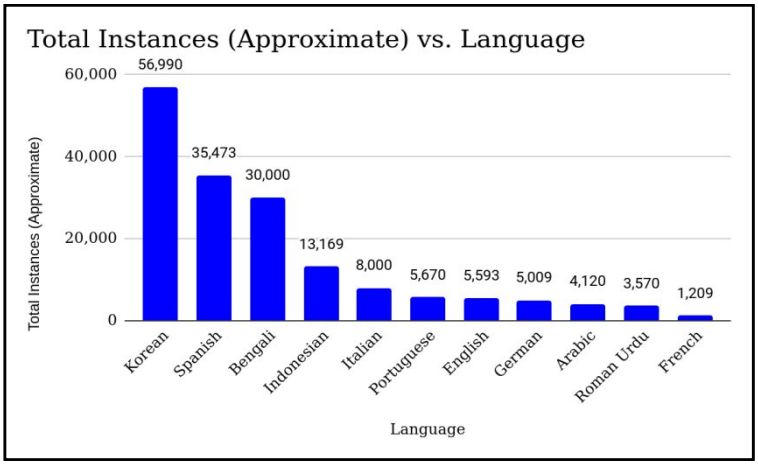

Fig. 4. . Language statistics for hate speech with data set instances.

4.2.2 Key Datasets for Major Languages.

The data set is the most common and efficient parameter for the research. The following table shows the details about the same

Table 3. Dataset collection for hate speech

Language	Dataset Name/Type	Total Instances (Approximate)	Study Focus
English[5][29]	MMHS150K(D1) Multimodal Twitter	150,000	Multimodal meme detection
English[6][31]	Founta et al. (FT) Twitter	80,000	Abuse detection
Telugu[16][33]	Monolingual Corpus Twitter	38,000	Low-resource language detection
Spanish[4]	MTLHateCorpus 2023 Twitter	35,473	Fine-grained classification, intensity
English[3][4]	Davidson et al. (DV) Twitter	24,802	Hate, offensive, neither
Arabic[4]	Arabic Hate-Speech (AHS) Twitter	9,352	Arabic BERT-Mini Model evaluation
Bengali[40]	Bengali Comments Dataset Facebook	7,425	Multi-class hate speech
Indonesian[4]	Pratiwi et al. (PR) Instagram	835	Harassing content detection

5 Technology Aspect

5.1 Natural Language Processing

NLP is the most powerful technologies for analyzing and understanding textual data. Through various preprocessing techniques such as tokenization, normalization, lemmatization, and stemming, NLP helps in cleaning and structuring raw text, enabling efficient analysis and downstream applications such as sentiment analysis, hate speech detection, and threat identification. Multi-lingual NLP enables the processing and understanding of text in multiple languages like English, Marathi, Hindi, Spanish, and Urdu, often using a single unified architecture. This is achieved by collecting the corpus from various languages and feeding it to the ML or DL model. Code mix and code

switching where two or more languages are used in communication e.g.Example: "Party खूप mast thi yaar", next time we should go together. Multimodal NLP refers to the intersection of natural language processing (NLP) with other data or modalities, such as images, videos, audio, and sensor data. Memes are shared by users on social media, which contains such a type of modality

5.2 Machine Learning

Machine Learning (ML), gives more accurate results. Support Vector Machine (SVM), Logistic Regression, Decision Tree, and Random Forest are employed to enhance the system's accuracy and performance

5.3 Deep Learning

To achieve better results and perform deeper analysis, Deep Learning (DL) models are often utilized. DL algorithms such as Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNN) help enhance the overall accuracy and performance of the system.

We did a technological survey covering Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) techniques to detect harmful content on social media. Various researchers worked on these techniques and got satisfying results. Table 4 describing the detailed summary of the technical survey across various studies.

Table 4. Technology Survey

Technique Category/Paradigm	Specific Technique/Model (NLP/ML/DL)	Description/Key Function
NLP: Feature Extraction & Representation		
Traditional Text Features	TF-IDF (Term Frequency-Inverse Document Frequency)[2],[16][3],[17],[4],[5]	Quantifies the importance of words; widely used for weighting unigrams and bigrams as numerical feature vectors
	N-grams (Word and Character)	Captures sequences of adjacent items (words or characters). Character n-grams often provide better performance.
	Bag-of-Words (BoW)[30][38][39][40][41][15]	Uses word frequencies collected from training data as features, but ignores word sequence and semantic content
Linguistic & Lexical Features	Part-of-Speech (PoS) Tags[14]	Detects the syntactic function and grammatical positions of words to capture context and relational patterns, significantly enhancing model performance
	Typed Dependencies (TD)[33]	Extracts dependency relationship labels to capture grammatical relationships
	Dictionaries/Lexical Resources[22][32]	Uses lists of specific keywords, derogatory terms, slang, and insults (e.g., Hatebase) as features or for data collection
Word/Sentence Embeddings	Word2Vec, GloVe, FastText[38][39]	Non-contextual word embedding methods used to enrich input features by representing words as vectors capturing semantic relationships

	ELMo (Embeddings from Language Model)	Deep contextualized word representation model pre-trained on large corpora, capturing grammar and semantics
	LASER (Language-Agnostic Sentence Representations)	Provides a uniform, language-independent vector embedding for sentences across multiple languages, effective in low-resource multilingual contexts.
	LaBSE (Language-Agnostic BERT Sentence Embeddings)	Used for textual feature extraction in multimodal models.
Contextual Features	Sentiment/Emotion Analysis[28]	Features derived from sentiment, often indicating negative emotions associated with hate speech
	Topic Modeling (LDA, BERTopic)[30]	Used to identify and compare underlying discursive themes. BERTopic uses UMAP for dimensionality reduction and HDBSCAN for clustering
	Readability Scores (FKGL, FRE)	Measures used to quantify the quality of a document
Preprocessing/Normalization	Tokenization, Stemming/Lemmatization, Stop-word/Punctuation Removal	Cleans text data; crucial step before feature extraction. Lemmatization and lowercasing presented high performance
	Contraction Expansion / Emoji Mapping	Converts abbreviations (e.g., "ain't") and visual language (emoticons/emojis) to text to reduce noise and clarify meaning
ML: Classical Machine Learning		
Linear Classifiers	Support Vector Machine (SVM) / Linear SVC (LSVC)	Highly used classifier, effective for binary classification.

		Linear SVM showed high accuracy in Arabic classification
	Naïve Bayes (NB) [	Commonly used classifier, particularly effective in text classification tasks
	Logistic Regression (LR)[45]	Used frequently, providing decent outcomes in HSD; often combined with embedding techniques like LASER
Tree-Based Models	Decision Trees (DT) [46]	Explored for classification tasks
	Random Forest (RF)[43]	An ensemble technique often used as a strong classifier
	XGBoost (Extreme Gradient Boosted Decision Trees)[42]	Ensemble algorithm noted for high performance across multiple social media platforms
DL: Neural Networks (Shallow/Recurrent/Convolution)		
Core Architectures	Multi-Layer Perceptron (MLP)[15]	Used for text categorization, capable of handling complex nonlinear feature relationships.
	Recurrent Neural Networks [17](RNN)[40]	Captures sequential patterns and temporal structure in text data
	Long Short-Term Memory (LSTM)[35]	Variant of RNN that excels at capturing long-range dependencies in text; commonly used for sequence modeling
	Bidirectional LSTM (BiLSTM)[15]	Processes input in both directions, demonstrating improved sensitivity to context and semantic nuances, often outperforming unidirectional LSTM

	Convolutional Neural Networks (CNN)[35]	Effective for feature extraction and capturing local patterns/semantics in text through convolutional filters
	Hybrid DL Models (e.g., CNN + LSTM, CNN + GRU)[35]	Combines architectures to capture both local patterns and sequential context
DL: Transformer Models (LLMs)		
Base Transformer Models	BERT (Bidirectional Encoder Representation from Transformers)[16][23][30]	Breakthrough model pretrained on large corpora, deeply bidirectional, used for fine-tuning specific NLP tasks
	ROBERTa (Robustly Optimized BERT Pretraining Approach)[24]	A robust BERT variant.
	DistilBERT[48][49]	A lightweight, distilled version of BERT, used in multilingual evaluation
	DeBERTa v3-large	Advanced transformer model fine-tuned for prediction tasks
Language-Specific LLMs	ABMM (Arabic BERT-Mini Model)	Specialized BERT version optimized for Arabic HSD on Twitter, with eight encoders
	AlBERTo (Italian BERT)	Fine-tuned BERT model used for Italian HSD and evaluating diachronic robustness.
Multilingual/Indic LLMs	mBERT (Multilingual BERT)[32]	Used for resource-constrained languages like Telugu.
Multimodal Architectures	CLIP, UNITER, BERT (MTL Framework)[27]	Multi-task learning framework combining Contrastive Language Image Pretraining (CLIP), UNiversal Image-TExt Representation Learning (UNITER), and BERT for hateful meme detection

According to the survey the workflow of the automatic content detection framework for hate speech and cyberbullying detection which were followed by most researchers is illustrated in Fig. 5. Researchers consider the text as well as image as an input from various datasets in different languages like English, Hindi, Bangali, Urdu, Marathi in their initial stages of research. For text preprocessing and feature extraction researchers applied techniques like TF-TDF, Word2Vector, NER. Different ML and DL models were applied in combination to get more accurate results.

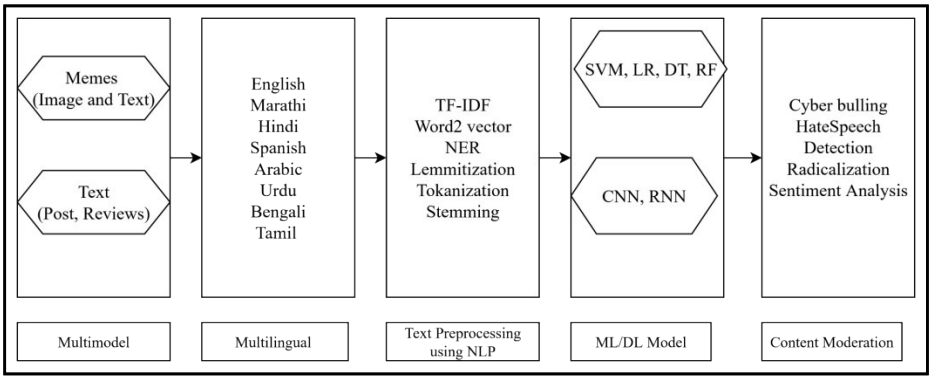

Fig. 5. Workflow of the automatic content detection

6 Model Evaluation

Researchers design the machine learning model, or Deep Learning model, for the automation of CyberBullying and hate speech detection from various social media platforms. The most frequently cited evaluation parameters across the sourcesareAccuracy [33][24][15][40],Precision[33][24][43][2][3][15], Recall[33][24][43][2][3][15][40, and F1-Score[33][24][43][2][3][15][40]. These models are computed on test data or on cross-validation data. Some researchers evaluated their model on all metrics, some evaluated only of F-1 score, precision, and recall. The overall model evaluation approach is language-centric; code-switching, code-mixing are missing.

7 Discussion

This section elaborates on the implications of our findings across important dimensions—social media platforms, datasets, languages, and technological aspects. During the reviewing of these dimensions collectively, we highlight the effectiveness, challenges, and future potential of research in this domain**.**

7.1 Social Media Platforms

The findings highlight each platform has distinct content formats, and linguistic features For example, platforms such as Twitter/X provide short, real-time text updates, whereas Facebook,YouTube offer longer posts, These variations affect the performance of NLP and machine-learning models, requiring platform-specific feature engineering and preprocessing strategies.

7.2 Datasets

Datasets play a crucial role in the detection of cyberbullying and hate speech. A wide variety of datasets exist, ranging from large collections of long-form text and messages to microblogging content from platforms such as Twitter, Facebook, and Reddit. As observed in the review, researchers have experimented with multiple datasets, each differing in size, structure, and contextual richness. One of the major challenges identified in the study is selecting an appropriate dataset that accurately represents real-world abusive behaviour. Additionally, classifying the data into distinct and meaningful categories like cyberbullying, harassment, threats, hate speech, or offensive. Overlapping linguistic patterns and annotation inconsistencies remains a challenge.

7.3 Languages

Language plays a pivotal role in the effectiveness of cyberbullying and hate-speech detection systems. Most existing research relies heavily on English datasets, As highlighted in the study, cyberbullying and hatespecch Words vary significantly across languages due to differences in grammar, slang. Code-Mixiging and code-switching remains challenges. Limited availability for low-resource languages became the barrier for the research. Appropriate use of language-specific NLP tools will help to improve the accuracy.

7.4 Technology Aspect

Technological advancements strongly impact the effectiveness of cyberbullying and hate-speech detection. Research gave preference to deep-learning and transformer-based models over traditional machine-learning models. However, challenges remain, handling multilingual, code-mixed, high computational needs, Efficient and adaptable

NLP technologies will perform well across diverse platforms and languages is a key finding from the study.

8 Conclusion

This article carried out a survey on automatic detection mechanisms for cyberbullying and hate speech content on social media. The evaluation in digital communication has made Social Media platforms vulnerable to various forms of cyber threats that are embedded in natural languages. Social media is the only means of communication and use for the spread of information in future cities. The most Popular social media is X (Twitter), the English language datasets used most frequently by the researcher are the key findings from this survey. The researchers used NLP, ML, and DL models in combination to achieve more accuracy and better performance. Researchers also mentioned the different challenges, like a)the unavailability of datasets in low-resource languages, especially Indian languages like Marathi, Panjabi, and Konkani. b) To work on Code Mix and Code Switch content c) annotations in different Indian Languages. Advanced technology like Zero-shot Learning, which performs tasks without explicit examples, will give a more accurate result and accelerate the automation mechanism in the future. Researchers can take this as an opportunity to work on this study.

9 References

1. Prashant Kapil, Asif Ekbal,(2025) A transformer based multi task learning approach to multimodal hate speech detectionNatural Language Processing Journal by Elsevier,2025
2. Fawzya Ramadan Sayed, Eman Hassan Elnashar, Fatma A. Omara,(2025) Cyberbullying detection in social media using natural language processing", Scientific African 28 (2025) e02713, homepage: www.elsevier.com/locate/sciaf.
3. Amira Ghenai, Zeinab Noorian, Hadiseh Moradisani, Parya Abadeh,Caroline Erentzen, Fattane Zarrinkalam,(2025) Exploring hate speech dynamics: The emotional, linguistic, and thematic impact on social media users", Information Processing and Management 62 (2025) 104079
4. Atul Kumar Srivastava, Mitali Srivastava, Sanchali Dasa, Vikas Jaina,Tej Bahadur Chandraa,(2025) Leveraging Deep Learning for Comprehensive Multilingual Hate Speech Detection 4th International Conference on Evolutionary Computing and Mobile Sustainable Networks, Procedia Computer Science 252 (2025) 832–840,
5. Sahrish Khan, Rabeeh Ayaz Abbasi, Muddassar Azam Sindhu, Sachi Arafat,Akmal Saeed Khattak, Ali Daud d, Mubashar Mushtaq, "Predicting the victims of hate speech on microblogging platforms", https://doi.org/10.1016/j.heliyon.2024.e40611,2025
6. Ronghao Pan, José Antonio García-Díaz , Rafael Valencia-García, "Spanish MTLHateCorpus 2023: Multi-task learning for hate speech detection to identify speech type, target, target group and intensity", https://doi.org/10.1016/j.csi.2025.103990, 2025
7. Deepawali Sharma,Tanusree Nath, Vediaka Gupta, Vivek Kumar Singh, "Hate Speech Detection Research in South Asian Languages: A Survey of Tasks, Datasets and Methods", ACM Trans. Asian Low-Resour.Lang. Inf. Process. 24, 3, Article 25 (March 2025), 44 pages. https://doi.org/10.1145/3711710

8. Mikelk K. Nguejio, Saurav Aryal, Marcellin Atemkeng, Gloria Washinton, Danda Rawat, “Decoding Fake News and Hate Speech: A Survey of Explainable AI Techniques” , ACM Comput. Surv. 57, 7, Article 169 (February 2025), 37 pages. https://doi.org/10.1145/3711123
9. P Vivekananth, Navneet Sharma,(2025) Detecting Cyberbullying in Social Media: An NLP-Based Classification Framework,INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY
10. Alessandra Teresa Cignarella, Anastasia,Giachanou, Elslefever,(2025) A Survey on Stereotype Detection in Natural Language Processing ,ACM 1557-7341/2025/10-ART https://doi.org/10.1145/3770754
11. Hasanov, Seppo Virtanen, Virtanen, Antti Hakkaala, Jouni Isoaho, (2024) Application of Large Language Models in Cybersecurity: A Systematic Literature Review by in IEEE
12. Deepawali Sharma, Vivek Kumar Singh, Akash Singh,(2024)THAR- Targeted Hate Speech Against Religion: A high-quality Hindi-English Codemixed Dataset with the Application of Deep lEarning Models for Automatic Detection”, ACM 2573-9522/2024/03-ARThttp://dx.doi.org/10.1145/3653017, 2024
13. Obaloluwa Ogundairo,Peter Broklyn, (2024) Natural Language Processing for Cybersecurity Incident Analysis, Journal of Cyber Security.
14. Emma Oye, P Peace, Jane Owen,(2024) The Role of Natural Language Processing in Cybersecurity,, ACM,2024.
15. Nikitha GS, Amritasri Shenoyy, K Chaturya, Latha JC, Janani Shree M,(2024) Detection of Cyberbullying Using NLP and Machine Learning in Social Networks for Bi-Language,,International Journal of Scientific Research & Engineering Trends Volume 10, Issue 2024
16. Y. Jeevan Nagendra Kumar, Rohith Reddy Vanapatla, Vamshi Krishna Pinamoni Jaswanth Kandukuri, Muntather Almusawi, Aravinda K, Lavish Kansal and Ravi Kalra, “Detecting cyberbullying in social media using text analysis and ensemble techniques” , E3S Web of Conferences 507, 01069 (2024) https://doi.org/10.1051/e3sconf/202450701069
17. Namit Khanduja,Nishant Kumar,Arun Chauhan,(2024) Telugu language hate speech detection using deep learning transformer models: Corpus generation and evaluation, by Systems and Soft Computing from Elsevier,
18. Yasmine M. Ibrahim, Reem Essameldin and Saad M. Darwish, “An Adaptive Hate Speech Detection Approach Using Neutrosophic Neural Networks for Social Media Forensics” ,Creative Commons Attribution 4.0 International License, DOI: 10.32604/cmc.2024.047840
19. Anchal Rawat, Santosh Kumar, Surender Singh Samant, “Hate speech detection in social media: Techniques, recent trends, and future challenges” , https://doi.org/10.1002/wics.1648, 2024
20. Julia Heine, Kristina Schaaff, and Tim Schlippe. 2024. Investigating Text and Image Features for the Detection of Cyberbulling. In 2024 8th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2024), es. https://doi.org/10.1145/3711542.3711545
21. Umar Farooq,Parvinder Singh,Surinder Singh Khurana Munish Kumar,(2023) Detection of content-based cybercrime in Roman Kashmiri using ensemble learning,,Springer,25 September
22. Vinaya Kulkarni, Vrushali Baghwat, Anjana Patil, Shubhangi Kumari, Suneha deep kour,(2023) A System to Identify Threats on Social Media Conversations and Providing Preliminary Legal Actions, IJRAR December

23. Zeya Lwin Tun, and Daniel Birks ,Lwin Tun and Birks,(2023)Supporting crime script analyses of scams with natural language processing. Crime Science (2023),https://doi.org/10.1186/s40163-022-00177-w.
24. Lanqin Yuan, Tianyu Wang,Gabriela Ferraro, Hanna Suominen,Marian-Andrei Rizoiu,(2023) Transfer learning for hate speech detection in social media, Journal of Computational Social Science (2023) 6:1081–1101 https://doi.org/10.1007s42001-023-00224-9, 2023
25. Aigerim Toktarova, Dariga Syrlybay, Bayan Myrzakhmetova, Gulzat Anuarbekova, Gulbarshin Rakhimbayeva,Balkiya Zhylanbaeva,Nabat Suieuova, Mukhtar Kerimbekov (2023),Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods,,(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 14, No. 5, 2023
26. Chesta Sofat,Divya Bansal, (2023) RadScore: An Automated Technique to Measure Radicalness Score of Online Social Media Users,Cybernetics and Systems An International Journal Volume 53, 2023.
27. Dr. T. Raghunadha Reddy, B. Madhubala, G. Varshini, S. K. Fayaz, (2023)A Deep Learning Approach for Author Profiling using Word Embeddings,,International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 11 Issue V May Available at www.ijraset.com
28. Nishant Vishwamitra, Keyan Guo, Song Liao Jaden Mu ,Zheyuan Ma,Long Cheng, Ziming Zhao, Hongxin Hu, (2023)Understanding and Analyzing COVID-19-related Online Hate Propagation Through Hateful Memes Shared on Twitter IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)
29. Francimaria R. S. Nascimento, George D. C. Cavalcanti, and Ma´rjory Da Costa-Abreu,(2023) Exploring Automatic Hate Speech Detection on Social Media: A Focus on Content-Based Analysis, SAGE Open April-June 2023: 1–19 The Author(s) 2023 DOI: 10.1177/21582440231181311 journals.sagepub.com/home/sgo
30. Devansh Mody, YiDong Huang, Thiago Eustaquio Alves de Oliveira,(2023) A curated dataset for hate speech detection on social media text Data in Brief 46 (2023) 108832,homepage: www.elsevier.com/locate/dib.
31. [30]Malik Almaliki, Abdulqader M. Almars, Ibrahim Gad and El-Sayed Atlam,(2023)ABMM: Arabic BERT-Mini Model for Hate-Speech Detection on Social Media", 2023, 12, 1048. https://doi.org/ 10.3390/electronics12041048
32. Joseph B. Walthe(2022)Social media and online hate,r,sciences Direct,
33. Md. Tofael Ahmed, Maqsudur Rahman, Shafayet Nur,Abu Zafor Muhammad Touhidul Islam,Dipankar Das,(2022)Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts TELKOMNIKA Telecommunication Computing Electronics and ControlVol. 20, No. 1, February 2022, pp. 89~97 ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v20i1.18630.
34. Stephen Afrifaa, and Vijayakumar Varadarajan, "Cyberbullying Detection on Twitter Using Natural Language Processing and Machine Learning Techniques", International Journal of Innovative Technology and Interdisciplinary Sciences, ISSN:2613-7305 Volume 5, Issue 4, pp. 1069-1080, 2022.
35. Megha Chaudhary,Sachin Vashistha,Divya Bansal (2022)Automated Detection of Anti-National Textual Response to Terroristic Events on Online Media,Cybernetics and Systems An International Journal Volume 53,
36. Chahat Raj, Ayush Agarwal, Gnana Bharathy,Bhuva Narayan and Mukesh Prasad,(2021) Cyberbullying Detection: Hybrid Models Based on MachineLearning and Natural Lan-

guage Processing Techniques.Electronics 2021, 10, 2810. https://doi.org/10.3390/electronics10222810.

37. Julia´n Rami´rez Sa´nchez, Alejandra Campo-Archbold, Andre´s Zapata Rozo,Daniel D´ıaz-Lo´pez,Javier Pastor-Galindo, Fe´lix Go´mez Ma´rmol,and Julia´n Aponte Dıaz, Hindawi,(2021)Uncovering Cybercrimes in Social Media through Natural Language Processing, Volume 2021, Article ID 7955637, 2021.
38. Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J. Wisniewski, and Munmun De Choudhury.2021. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 325 (October 2021), 34 pages. https://doi.org/10.1145/3476066
39. Manuel F. López-Vizcaíno, Francisco J. Nóvoa, Victor Carneiro, Fidel Cacheda(2021), "Early detection of cyberbullying on social media networks",Future Generation Computer Systems 118 (2021) 219–229
40. Mayur Gaikwad, Swati Ahirrao,Shraddha Phansalkar, and Ketan Kotecha,(2021)Online Extremism Detection: A Systematic Literature Review With Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools,,IEEE Access, 2021,
41. Amit Kumar Das, Abdullah Al Asif, Anik Paul, and Md. Nur Hossain,(2021) "Bangla hate speech detection on social media using attention-based recurrent neural network", Journal of Intelligent Systems 2021; 30: 578–591
42. Tommy K.H. Chan a, Christy M.K. Cheung b, Zach W.Y. Lee (2021) Cyberbullying on social networking sites: A literature review and future research directions, ,sciencedirect,2021.
43. Nanlir Sallau and Wan Mohd Nazmee Wan Zainon ,(2021) Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review, Digital Object Identifier 10.1109/ACCESS.2021.3089515,2021
44. Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile,and Viviana Patti,(2020) Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media, Appl. Sci. 2020, 10, 4180; doi:10.3390/app10124180, www.mdpi.com/journal/applsci
45. Muhammad U. S. Khan; Assad Abbas; Attiqa Rehman; Raheel Nawaz(2020), "HateClassify: A Service Framework for Hate Speech Identification on Social Media", IEEE Internet Computing (Volume: 25, Issue: 1, 01 Jan.-Feb. 2021) **Page(s):** 40 - 49
46. Zewdie Mossie, Jenq-Haur Wang,(2020)Vulnerable community identification using hate speech detection on social media", Information Processing & Management Volume 57, Issue 3, May 2020, 102087 https://doi.org/10.1016/j.ipm.2019.102087 Cited by (142), 2020
47. Chahat Raj, Ayush Agarwal, Gnana Bharathy, Bhuva Narayan, and Mukesh Prasad,(2021) Cyberbullying Detection: Hybrid Models Based on Machine Learning and Natural Language Processing Techniques, Electronics 2021, 10, 2810.https://doi.org/10.3390/electronics 10222810

48. Johnson, N. F.; Leahy, R.; Johnson Restrepo, N.; Velasquez, N.; Zheng, M.; Manrique, P.; Devkota, P.; Wuchty, S. (2019). "Hidden resilience and adaptive dynamics of the global online hate ecology"
49. Stevie Chancellor, Eric P.S Baumer, and Munmun De Choudhury.(2019) “Who is the “Human” in HumanCentered Machine Learning: The Case of Predicting Mental Health from Social Media”. Proc. ACM Hum.-Comput.Interact. 3, CSCW, Article 147 (November 2019), 32 pages. https://doi.org/10.1145/3359249
50. Sweta Agrawal, Amit Awekar,(2018) “Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms”, arXiv:1801.06482v1 [cs.IR] 19 Jan 2018.
51. Mengfan Yao,Charalampos Chelmis.Daphney–Stavroula Zois(2028),Cyberbullying Detection on Instagram with Optimal Online Feature Selection, 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

52. Michela Del Vicario, Alessandro Bessi,Fabiana Zollo,Fabio Petronic,Antonio Scala,Guido Caldarelli H. Eugene Stanley,and Walter Quattrociocchi(2015)The spreading of misinformation online, 2015
53. Yuksel Ekinci, Shubhankar Dam, Georgia Buckle, (2025) The Dark Side of Social Media Influencers: A Research Agenda for Analysing Deceptive Practices and Regulatory Challenges", Psychology & Marketing, 2025; 42:1201–1214 1201 of 1214 https://doi.org/10.1002/mar.22173
54. Brett Drury, Samuel Morais Drury, Md Arafatur Rahman, Ihsan Ullah,(2022) A social network of crime: A review of the use of social networks for crime and the detection of crime, Online Social Networks and Media 30 (2022) 100211