A Comprehensive Review of Cyberbullying Detection Techniques and Datasets on Social Media Platforms

Yeshodha R
Aditya Institute of Management Studies and Research, Bangalore, Karnataka, India

Abstract—Cyberbullying has emerged as a critical issue in today's digital society, particularly among adolescents who extensively use social media for communication and information exchange. Unfortunately, some users exploit these online platforms to harass, humiliate, or intimidate others through posts, messages, or digital interactions, resulting in severe psychological and emotional consequences for victims. Although numerous studies have investigated the detection, prevention, and mitigation of cyberbullying, the problem persists due to evolving online behaviors and linguistic complexities. This study presents a comprehensive systematic review of existing research on cyberbullying detection from 2018 to 2024. It examines state-of-the-art datasets, methodologies, and technologies—ranging from traditional machine learning models to deep learning and large language models (LLMs). The paper identifies key challenges, research gaps, and opportunities for future exploration, offering insights and recommendations to enhance the accuracy, fairness, and scalability of cyberbullying detection systems across diverse social media platforms. The review analyzed 27 papers using the PRISMA framework, revealing that hybrid deep learning models (CNN-BiLSTM) achieve the highest performance (93.4% accuracy, 92.8% F1-score), while dataset imbalance and English-language dominance (74.1%) remain critical challenges.

Index Terms—Cyberbullying Detection, Social Media Analysis, Machine Learning, Deep Learning, Natural Language Processing, Large Language Models, Online Harassment, Hate Speech Detection, Sentiment Analysis, Dataset Imbalance.

I. INTRODUCTION

A. Background and Motivation

THE digital revolution has fundamentally transformed how individuals communicate, share information, and build communities. Social media platforms such as Facebook, Twitter, Instagram, YouTube, and TikTok have become integral to daily life, connecting billions of users worldwide. As of 2024, approximately 4.95 billion people actively use social media platforms, representing 61% of the global population, with Facebook leading at 3.05 billion users [1]. This unprecedented connectivity has created new opportunities for social interaction and information dissemination.

However, this digital expansion has enabled the proliferation of cyberbullying—a form of harassment that occurs through digital devices and online platforms [2]. Cyberbullying involves repeated, aggressive digital behavior across various platforms, targeting individuals using devices such as mobile phones and computers [3]. Unlike traditional bullying, cyberbullying can occur 24/7, reach vast audiences instantly, and leave permanent digital footprints that amplify harm to victims.

The impact of cyberbullying on victims is severe and multifaceted. Research indicates that victims experience increased rates of depression (41%), anxiety (37%), suicidal thoughts (26%), self-harm (19%), eating disorders (16%), and substance abuse (15%) [4]. Adolescents aged 9 to 17 are particularly vulnerable, with the COVID-19 pandemic's shift to virtual interactions intensifying cyberbullying instances by nearly 70% [5].

B. Problem Statement and Research Gap

Despite extensive research and development of numerous detection systems, cyberbullying remains a persistent challenge. Current detection approaches face several critical limitations:

1) Evolving Tactics: Cyberbullies continuously develop new methods involving obfuscated words, misspellings, and coded languages that evade detection systems [6]. The dynamic nature of online

communication, including slang, emojis, and cultural references, makes pattern recognition increasingly complex.

2) Language Bias: Most cyberbullying detection algorithms are developed for high-resource languages like English, Chinese, French, and German, causing bias in detecting cyberbullying in low-resource languages like Swahili, Bengali, and Marathi [7]. This linguistic imbalance affects accuracy, precision, and fairness across diverse global communities.

3) Data Scarcity: Collection and labeling of data is cost and time-consuming due to platform restrictions [8]. Only Twitter allows researchers to clone unlabeled data via API for academic use without significant restrictions compared to other platforms with tight data access policies.

4) Multimodal Challenges: Many algorithms are developed to classify instances based on text only, which cannot capture all forms of cyberbullying including multimodal content [9].

5) Platform Diversity: Social media platforms are changing in terms of data availability, features, and user interfaces, making it difficult to develop multipurpose automatic detection systems [10].

C. Research Objectives and Contributions

This systematic review aims to provide comprehensive analysis of cyberbullying detection methodologies from 2018 to 2024. The specific objectives are:

1. To conduct a systematic review of existing studies concerning methods used in detecting cyberbullying on social media platforms
2. To compare and analyze existing technologies, approaches, datasets, and evaluation metrics
3. To identify gaps in current research and propose effective solutions
4. To examine linguistic diversity and cultural considerations across different regions and languages
5. To evaluate performance of various detection approaches including ML, DL, traditional methods, and LLMs

The key contributions of this paper include:

- Comprehensive taxonomy of cyberbullying detection approaches covering ML, DL, traditional methods, and LLMs
- Detailed analysis of publicly available datasets, including sources, sizes, languages, and annotation quality
- Comparative performance evaluation using standard metrics
- Identification of critical challenges including dataset imbalance, linguistic diversity, and real-time detection limitations
- Recommendations for future research directions including multilingual detection, multimodal analysis, and ethical considerations

II. RESEARCH METHODOLOGY

A. Review Protocol (PRISMA Framework)

This systematic review employs the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework [11] to ensure transparency and replicability. The review process is categorized into three stages: identification, eligibility, and inclusion.

1) Identification Stage: The identification phase involved comprehensive searches across IEEE Xplore Digital Library, SpringerLink, ScienceDirect, ACM Digital Library, MDPI, and other sources including Google Scholar and arXiv. The search covered publications from January 2018 to December 2023. A total of 279 records were identified from database searching, with an additional 87 records from other sources, resulting in 366 initial records.

2) Search Strategy: The search employed combinations of keywords: primary terms ("cyberbullying detection", "online harassment"), method-related terms ("machine learning", "deep learning", "NLP"), platform-related terms ("Twitter", "Facebook", "Instagram"), linguistic terms ("multilingual", "low-resource languages"), and behavior-related terms ("hate speech", "toxic comments", "abusive language").

3) Eligibility Stage: After removing 160 duplicate records, 119 abstracts and titles were screened. Inclusion criteria encompassed studies focusing on cyberbullying detection methodologies using ML, DL, traditional methods, or LLMs; using social media data; with clear methodology and evaluation metrics. Exclusion criteria included competition reports without methodology, studies unrelated to cyberbullying, and non-English publications. After screening, 69 full-text articles were assessed, with 42 excluded for not meeting criteria.

4) Inclusion Stage: The final selection resulted in 27 relevant papers included in the systematic review.

B. Data Extraction and Classification

For each included study, information was systematically extracted including: bibliographic information, methodology (detection approach and algorithms), datasets (source, languages, size), features (extraction methods, embeddings), evaluation metrics (accuracy, precision, recall, F1-score), performance results, limitations, and future directions.

Data was categorized into four main approaches: Machine Learning (SVM, Naive Bayes, Random Forest, K-NN, Logistic Regression); Deep Learning (CNN, RNN, LSTM, BiLSTM, GRU, hybrid models); Traditional Methods (rule-based, keyword filtering, character matching); and Large Language Models (BERT, GPT, RoBERTa, DistilBERT).

Timeline Analysis: The chronological distribution reveals a rising trend from 2018 (3 papers) to 2023 (10 papers), indicating increasing research interest. Source Distribution: 48.1% came from other sources, 22.2% from IEEE Xplore, 11.1% each from SpringerLink and MDPI, with smaller percentages from ScienceDirect (3.7%) and ACM (3.7%).

III. CYBERBULLYING: CONTEXT AND IMPACT

A. Definition and Characteristics

Cyberbullying is characterized by repeated, aggressive digital behavior occurring across platforms like social media, gaming, and messaging, targeting individuals using digital devices [12], [13]. Unlike traditional face-to-face bullying, cyberbullying leverages technology to inflict harm through digital channels, making it pervasive and difficult to escape.

Key distinguishing features include: anonymity enabling aggressive behavior, 24/7 accessibility extending beyond traditional boundaries, wide reach spreading content to large audiences, permanence of digital content, and difficulty in supervision by parents and educators [14], [15].

The primary forms include harassment (repeated offensive messages), hate speech (prejudice-based content), online harassment (persistent multi-channel targeting), impersonation (fake profiles), cyberstalking (intense harassment causing fear), flaming (angry messages), outing (sharing private information), doxing (publishing identifying information), denigration (spreading rumors), exclusion (intentional isolation), and trolling (inflammatory content) [16]-[18].

B. Causes and Contributing Factors

Cyberbullying stems from various factors [19], [20]: individual factors (physical appearance, academic achievement, sexuality, financial status, psychological traits); social and cultural factors (religion, cultural background, political views, entertainment preferences); technological factors (cyber syndrome, platform design, digital literacy); and environmental factors (peer influence, lack of supervision, normalization of aggressive behavior).

C. Impact on Individuals and Society

Psychological Impacts: Depression (41% of victims), social anxiety (37%), suicidal thoughts (26%), self-harm (19%), eating disorders (16%), substance abuse (15%), social media withdrawal (32%), and profile deletion (28%) [21]-[23].

Academic and Social Impacts: Decreased academic performance, reduced school attendance, withdrawal from activities, strained relationships, difficulty maintaining friendships, and reduced self-esteem [24], [25].

Societal Implications: Increased healthcare costs, reduced workplace productivity, normalization of online aggression, legal challenges, and erosion of trust in online spaces [26], [27].

Vulnerable Populations: Adolescents are particularly vulnerable during critical developmental periods. Other vulnerable groups include individuals with

disabilities, minority communities, LGBTQ+ youth, those with pre-existing mental health conditions, and individuals experiencing concurrent offline bullying [28], [29].

IV. DETECTION FRAMEWORK AND PROCESS

A. Data Acquisition

Dataset preparation includes data source identification, acquisition, and distribution for training and testing. Social media APIs facilitate data collection: Twitter API (Twitter4J, Tweepy) most commonly used; Facebook Graph API with strict controls; Instagram Graph API for public content; YouTube Data API for comments; and REST API for generic access [30]. Web scraping is employed when API access is limited, but must comply with platform terms and legal requirements.

Ethical considerations include adherence to platform terms of service, user privacy protection (GDPR, CCPA), IRB approvals, informed consent, and anonymization protocols [31].

B. Preprocessing Pipeline

Data preprocessing prepares data for algorithm inputs. Steps include: data cleaning (removing URLs, duplicates, spam, handling missing values); tokenization (breaking text into tokens using word-level, subword, or character-level approaches); stop words removal (eliminating common words); normalization (lowercase conversion, contraction expansion, spelling standardization, emoji handling); stemming and lemmatization (reducing words to root forms); and data labeling (binary or multi-class categorization) [32], [33].

C. Feature Extraction

Feature extraction significantly affects algorithm performance [34]. Traditional methods include: TF-IDF (measuring word importance), Bag of Words (frequency counts), N-grams (word sequences), PoS tagging (grammatical roles), semantic features (meaning-based), topic modeling (LDA), PCA (dimensionality reduction), and sentiment analysis (emotional tone).

Word embeddings provide distributed representations: Word2Vec (Skip-gram, CBOW), GloVe (co-occurrence statistics), FastText (character n-grams handling misspellings), ELMo (contextualized embeddings), BERT (context-aware bidirectional), and Farasa (Arabic-specialized) [35], [36].

D. Model Training and Evaluation

Model learning trains algorithms to classify cyberbullying instances. The training process involves: data splitting (70/15/15 or 80/10/10 for train/validation/test), hyperparameter tuning, cross-validation, and class imbalance handling (SMOTE, undersampling, weighted loss) [37].

Evaluation metrics include: accuracy (overall correctness), precision (reducing false positives), recall (reducing false negatives), F1-score (harmonic mean), and ROC-AUC (discriminative ability) [38].

V. DETECTION APPROACHES: COMPARATIVE ANALYSIS

A. Machine Learning Approaches

Machine learning enhances cyberbullying detection by extracting features and improving precision [39]. Supervised algorithms include SVM (optimal hyperplanes), Naive Bayes (probabilistic classifier), Random Forest (ensemble trees), K-NN (similarity-based), Logistic Regression (linear model), Decision Trees (hierarchical decisions), and Gradient Boosting (XGBoost, AdaBoost) [40]-[43]. Unsupervised approaches include K-Means clustering [44].

Performance Comparison: Analysis of machine learning approaches reveals that supervised methods outperform unsupervised ones [45], [46]. SVM average: 87.5% accuracy, 85.2% precision, 78.5% recall, 81.3% F1-score. Naive Bayes average: 89.1% accuracy, 86.8% precision, 82.1% recall, 84.3% F1-score. Random Forest: 85.3% accuracy, 81.9% F1-score. Ensemble Methods: 90.2% accuracy, 87.5% F1-score.

Strengths: Interpretability, efficiency, lower computational requirements, effectiveness with smaller datasets, and established theoretical foundations [47], [48].

Limitations: Manual feature engineering required, limited context understanding, difficulty capturing complex patterns, performance degradation with

unseen patterns, and high false positive rates [49], [50].

B. Deep Learning Approaches

Deep learning efficiently manages large datasets with automatic feature extraction [51]. CNNs treat text as one-dimensional sequences [52]. RNN variants include LSTM (addressing vanishing gradients), BiLSTM (bidirectional processing capturing past and future contexts), and GRU (simplified LSTM variant) [53]-[55]. Hybrid models combine architectures: CNN-LSTM, CNN-BiLSTM, CNN-BiGRU, and BiLSTM-BiGRU [56], [57]. Autoencoders (stacked, sparse, denoising) perform unsupervised learning [58]. GANs augment data for class imbalance [59].

Performance Comparison: BiLSTM average: 91.5% accuracy, 90.2% precision, 92.8% recall, 91.4% F1-score. CNN: 88.7% accuracy, 86.4% F1-score. LSTM: 87.2% accuracy, 87.2% F1-score. Hybrid models: 93.4% accuracy, 92.8% F1-score [60]-[63].

Strengths: Automatic feature extraction, capturing complex patterns, scalability, superior high-dimensional performance, sequential and contextual information capture, and better linguistic nuance handling [64], [65].

Limitations: Data dependency requiring large labeled datasets, computational intensity, black-box nature, overfitting risk, longer training times, and hyperparameter tuning difficulty [66], [67].

C. Traditional (Rule-Based) Methods

Traditional techniques include keyword filtering (offensive word lists), pattern matching (syntactic patterns like repeated punctuation, ALL CAPS), character percentage matching (offensive character analysis), and rule-based systems (expert knowledge heuristics) [68].

Performance: Average 96% accuracy, 71% precision, 73% recall, 97% F1-score on Ofcom (English) and TUKI (Swahili) datasets [69].

Strengths: Simplicity, speed, transparency, no training data required, and low computational requirements.

Limitations: Context insensitivity, high false positives, inability to detect subtle bullying, easy circumvention through obfuscation, limited adaptability, poor sarcasm handling, and scalability issues [70].

D. Large Language Models (LLMs)

LLMs leverage transformer architectures and massive pre-training [71]. BERT variants (Multilingual BERT, DistilBERT, ALBERT) use bidirectional context [72]. RoBERTa optimizes BERT training [73]. GPT models (GPT-3, GPT-3.5-turbo) use autoregressive prediction [74]. XLNet uses permutation-based modeling [75]. XLM-RoBERTa provides cross-lingual capabilities [76].

Transfer learning involves fine-tuning pre-trained models on cyberbullying datasets. Few-shot and zero-shot learning enable detection with minimal task-specific training [77], [78].

Performance Comparison: BERT variants: 90.5% accuracy, 89.7% precision, 88.4% recall, 89.0% F1-score. RoBERTa: 87.0% accuracy, 87.0% F1-score. GPT models: 83.8% accuracy, 75.4% F1-score [79]-[81].

Strengths: Superior context understanding, multilingual capability, transfer learning with limited data, state-of-the-art performance, adaptability, nuanced pattern capture, and zero/few-shot capabilities [82], [83].

Limitations: Computational cost, large model size, training time, bias concerns, limited interpretability, real-time constraints, and data privacy concerns [84], [85].

VI. DATASETS ANALYSIS

A. Overview of Public Datasets

Major Platform Sources: Twitter is most frequently used due to API accessibility (datasets ranging from 1,431 to 350,000 tweets) [86], [87]. Facebook datasets include 6,000 to 126,704 posts [88]. Instagram datasets range from 500 to 4,816,345 posts [89]. YouTube datasets include 6,000 to 92,000 comments [90]. Reddit datasets range from 1,431 to 47,000 posts [91]. Other platforms include FormSpring (500 to 99,544 posts), ASKfm (192,085 posts), Wikipedia (115,864 to 128,000 edits), Myspace (10,566 posts), Ofcom (200 posts), and TUKI (200 posts) [92]-[95].

Language Distribution: English dominates at 74.1%, followed by Bengali (7.4%), Arabic (7.4%), Hindi (3.7%), Swahili (3.7%), and Dutch (3.7%), highlighting the need for diverse linguistic representation [96].

B. Dataset Challenges

Class Imbalance: Severe imbalance affects model training [97]. Examples include Twitter dataset with 10,000:1 ratio (68.4% neither sexist nor racist, 19.5% sexist, 12.1% racist); FormSpring with 15:1 ratio (93.9% non-cyberbullying, 6.1% cyberbullying); Wikipedia with 8:1 ratio (88.3% non-cyberbullying, 11.7% cyberbullying) [98]-[100]. Impact includes bias toward majority class, poor generalization, misleading metrics, and threshold challenges.

Limited Low-Resource Language Coverage: Most datasets focus on high-resource languages, causing bias [101]. Challenges include lack of labeled data, limited linguistic resources, insufficient annotators, cultural differences, and code-switching [102].

Annotation Inconsistencies: Sources include subjective labeling, ambiguous definitions, cultural variations, low inter-annotator agreement, and annotation errors [103]. Solutions involve multiple annotator consensus, clear guidelines, training, quality control, and semi-automated tools [104].

Privacy and Ethical Concerns: Key concerns include user consent, re-identification risk, victim privacy, data sharing ethics, platform ToS compliance, and research ethics [105]. Best practices include anonymization, data aggregation, limited sharing, secure storage, and IRB approval.

C. Dataset Quality Assessment

Quality dimensions include representativeness, size and coverage, annotation quality, temporal relevance, platform diversity, and linguistic diversity [106]. Issues identified: limited dataset size (< 1,000 samples show poor generalization), inaccurate data (deviations from actual content), and lack of generalization (single-platform datasets don't transfer well) [107], [108].

Improvement strategies include data augmentation (synonym replacement, back-translation, GANs, paraphrasing), active learning (selecting informative samples), cross-platform validation, temporal validation, and multi-annotator approaches [109], [110].

VII. EVALUATION METRICS AND PERFORMANCE

A. Standard Metrics

Accuracy: Overall correctness = (TP + TN) / (TP + TN + FP + FN). Useful for balanced datasets but misleading with imbalance.

Precision: Proportion of correct positive predictions = TP / (TP + FP). Important when false positives are costly [111].

Recall: Proportion of actual positives identified = TP / (TP + FN). Important when missing actual bullying is costly.

F1-Score: Harmonic mean = 2 × (Precision × Recall) / (Precision + Recall). Balanced metric for imbalanced datasets [112].

Additional metrics include ROC-AUC (threshold-independent) and confusion matrix (detailed breakdown) [113].

B. Cross-Method Performance Comparison

TABLE I presents comprehensive performance comparison across different detection approaches.

Machine Learning: SVM (87.5% accuracy, 81.3% F1), Naive Bayes (89.1% accuracy, 84.3% F1), Random Forest (85.3% accuracy, 81.9% F1), Ensemble (90.2% accuracy, 87.5% F1) [114]-[116].

Deep Learning: BiLSTM (91.5% accuracy, 91.4% F1), CNN (88.7% accuracy, 86.4% F1), LSTM (87.2% accuracy, 87.2% F1), Hybrid CNN-BiLSTM (93.4% accuracy, 92.8% F1) [117]-[119].

Traditional: Rule-based (96% accuracy, 97% F1 but only 71% precision) [120].

LLMs: BERT (90.5% accuracy, 89.0% F1), RoBERTa (87.0% accuracy, 87.0% F1), GPT-3 (83.8% accuracy, 75.4% F1) [121]-[123].

C. Best-Performing Approaches

Overall Rankings: 1) Hybrid CNN-BiLSTM (93.4% accuracy, 92.8% F1); 2) BiLSTM (91.5% accuracy,

91.4% F1); 3) BERT variants (90.5% accuracy, 89.0% F1); 4) Ensemble ML (90.2% accuracy, 87.5% F1); 5) Naive Bayes (89.1% accuracy, 84.3% F1) [124]-[126].

By Task Type: Binary classification: CNN-BiLSTM (95.8% F1). Multi-class: BERT (89.7% F1). Multilingual: Multilingual BERT (88.5% F1). Real-time: Naive Bayes with TF-IDF (84.3% F1, <100ms latency) [127]-[129].

VIII. CHALLENGES AND LIMITATIONS

A. Language and Linguistic Challenges

Multilingual Detection: Limited training data for non-English languages, different linguistic structures, cultural context variations, code-switching, and lack of pre-trained models for low-resource languages [130], [131]. Impact includes bias toward English, inadequate protection for non-English communities, and difficulty transferring models.

Code-Switching: Users mix languages (Hinglish, Spanglish), creating tokenization difficulties, embedding limitations, grammar complexity, and contextual interpretation challenges [132].

Slang and Informal Language: Platform-specific slang, generational differences, rapid evolution, regional variations, abbreviations ("lol", "brb"), emoji usage, and meme references [133].

Sarcasm and Irony: Challenges include literal vs. intended meaning, context dependency, cultural variations, missing tone indicators, and subtle linguistic cues [134].

Obfuscated Text: Techniques include character substitution ("h8"), leetspeak ("l33t"), intentional misspellings, character insertion, and phonetic variations [135].

B. Technical Challenges

Real-Time Detection: Latency requirements (<1 second), computational limitations, scalability to millions of users, model complexity vs. speed trade-off, and batch vs. stream processing [136]. Impact includes delayed intervention and resource overload.

Model Interpretability: Deep learning as black boxes, difficulty explaining predictions, lack of transparency, debugging challenges, and limited domain knowledge incorporation [137]. Approaches include attention visualization, LIME, SHAP, and feature importance.

Scalability and Generalization: Processing billions of posts daily, memory constraints, distributed computing requirements, performance degradation on unseen platforms, temporal drift, and cross-platform variability [138], [139].

Transfer Learning: Domain mismatch, hyperparameter sensitivity, catastrophic forgetting, computational cost, and limited labeled fine-tuning data [140], [141].

C. Data-Related Challenges

Imbalanced Datasets: Bias toward majority class, poor minority recall, misleading metrics, and threshold challenges [142]. Mitigation includes SMOTE, undersampling, class weighting, ensembles, and synthetic data [143]-[145].

Low-Resource Language Scarcity: Insufficient data for languages beyond English [146]. Solutions include cross-lingual transfer, multilingual models, crowdsourced annotation, and active learning [147].

Annotation Quality: Inter-annotator variability, subjective interpretations, cultural misunderstanding, fatigue errors, and cost constraints [148]. Quality assurance includes consensus, expert review, guidelines, and calibration.

Privacy and Ethical Constraints: User consent, re-identification risks, sensitive content exposure, data breach vulnerabilities, and regulatory compliance (GDPR, CCPA) [149], [150].

IX. FUTURE DIRECTIONS AND RECOMMENDATIONS

A. Dataset Enhancement

Multilingual and Balanced Datasets: Expand coverage to low-resource languages (Swahili, Bengali, Marathi), balance classes, include diverse types, capture temporal evolution, and collect cross-platform data [151]. Strategies include international collaboration, crowdsourced collection, platform partnerships, multilingual teams, and standardized protocols.

Synthetic Data Generation: Use GANs, GBO, SSG, VAEs, back-translation, and paraphrasing to address

imbalance, expand datasets, create diversity, reduce costs, and improve robustness [152]-[154].

Automated Annotation: Leverage SSL models, semi-supervised learning, active learning, pre-trained models, and human-in-the-loop validation to reduce time and cost, scale datasets, ensure consistency, and enable iterative improvement [155], [156].

B. Advanced Detection Techniques

Ensemble and Hybrid Models: Stack ML algorithms, use voting-based ensembles, weighted combinations, hybrid CNN-BiLSTM architectures, and multi-task learning for improved accuracy, robustness, reduced overfitting, and better generalization [157], [158].

Cross-Lingual Transfer: Use Multilingual BERT, XLM-RoBERTa, zero-shot transfer, few-shot learning, language-agnostic representations, and pivot languages to reduce data requirements, enable faster deployment, share knowledge, and provide cost-effective coverage [159], [160].

Multimodal Detection: Analyze text, images (memes, screenshots), videos (content, comments), audio (voice, tone), and metadata (timestamps, engagement) [161]. Use early fusion (combine features), late fusion (combine predictions), attention-based fusion, and hierarchical fusion. Apply computer vision, OCR, video analysis, speech-to-text, and emoji analysis for comprehensive detection, reduced evasion, better context, and improved accuracy [162], [163].

C. Practical Implementation

Real-Time Content Moderation: Develop stream processing pipelines (Kafka, Flink), lightweight models, edge computing, caching, and scalable infrastructure [164]. Requirements include <100ms latency, high throughput, fault tolerance, and load balancing.

Platform Integration: Create API-based services, webhook notifications, browser extensions, mobile SDK integration, and automated flagging systems [165]. Use gradual rollout, A/B testing, monitoring, feedback loops, and continuous improvement.

User Reporting Tools: Design intuitive interfaces, provide context collection, enable anonymous reporting, offer real-time feedback, and establish appeal processes [166]. Support victims through resource links, safety tips, counseling access, community support, and follow-up communication.

D. Ethical Considerations

Privacy-Preserving Methods: Implement federated learning (train without centralizing data), differential privacy (add noise), homomorphic encryption (compute on encrypted data), secure multi-party computation, and on-device processing [167]. Benefits include user trust, regulatory compliance, reduced liability, and enhanced security.

Bias Mitigation: Use diverse training data, fairness-aware algorithms, regular bias audits, inclusive annotation teams, and adversarial debiasing [168]. Ensure equitable protection, reduce discrimination, build trust, meet ethical standards, and promote social justice.

Transparency and Accountability: Provide explainable AI, clear decision disclosure, user appeals, regular audits, and stakeholder engagement [169]. Foster trust, enable improvement, ensure oversight, meet regulations, and promote responsible AI.

X. CONCLUSION

This comprehensive review examined cyberbullying detection methodologies across social media platforms, focusing on machine learning, deep learning, traditional techniques, and large language models. The study analyzed 27 papers from 2018 to 2024 using the PRISMA framework, revealing significant insights into current capabilities and future directions.

Key Findings: Hybrid deep learning models (CNN-BiLSTM) achieve the highest performance (93.4% accuracy, 92.8% F1-score), followed by BiLSTM (91.5% accuracy, 91.4% F1) and BERT variants (90.5% accuracy, 89.0% F1). Machine learning approaches, particularly ensemble methods (90.2% accuracy) and Naive Bayes (89.1% accuracy), offer good performance with lower computational requirements. Traditional rule-based methods show high accuracy (96%) but suffer from low precision (71%), limiting practical utility.

Critical Challenges: Dataset imbalance remains pervasive, with majority-to-minority ratios reaching

10,000:1, severely impacting model performance. English dominates datasets (74.1%), creating significant bias against non-English communities. Low-resource languages (Swahili, Bengali, Marathi) receive minimal attention despite representing billions of speakers. Real-time detection capabilities remain limited, hindering timely intervention. Model interpretability challenges persist, particularly with deep learning and LLMs. Linguistic complexities including slang, sarcasm, code-switching, and obfuscated text continue to challenge detection systems.

Research Gaps: Limited multimodal detection capabilities, as most systems focus solely on text analysis. Insufficient cross-lingual transfer learning applications for low-resource languages. Lack of standardized evaluation frameworks and benchmarks. Limited real-world deployment studies and impact assessments. Inadequate attention to ethical considerations and bias mitigation. Scarcity of longitudinal studies examining temporal evolution of cyberbullying patterns.

Future Directions: Development of comprehensive multilingual datasets with balanced class distributions. Integration of multimodal analysis combining text, images, videos, and audio. Advancement of real-time detection systems with sub-second latency. Implementation of privacy-preserving techniques like federated learning. Enhancement of model interpretability through attention mechanisms and explainable AI. Application of cross-lingual transfer learning to expand language coverage. Establishment of ethical frameworks addressing bias, fairness, and accountability.

Call for Interdisciplinary Collaboration: Effective cyberbullying detection requires collaboration across computer science, psychology, sociology, linguistics, law, and education. Such collaboration ensures holistic solutions addressing technical, human, and societal dimensions.

Only through continued innovation and interdisciplinary collaboration can we create safer digital environments where all users can communicate without fear of harassment and harm.

ACKNOWLEDGMENT

The author would like to thank the researchers whose work was reviewed in this study, and the academic institutions and platforms that make cyberbullying detection research possible through data access and computational resources.

REFERENCES

[1] [1] R. Shewale, "Social media users and statistics in 2024," DemandSage, 2024. [Online]. Available: https://www.demandsage.com/social-media-users/

[2] [2] W. Cassidy, C. Faucher, and M. Jackson, "Cyberbullying among youth: A comprehensive review of current international research and its implications and application to policy and practice," *School Psychology International*, vol. 34, no. 6, pp. 575-612, Dec. 2013.

[3] [3] T. R. Soomro and M. Hussain, "Social media-related cybercrimes and techniques for their prevention," *Appl. Comput. Syst.*, vol. 24, no. 1, pp. 9-17, May 2019.

[4] [4] C. Nixon, "Current perspectives: The impact of cyberbullying on adolescent health," *Adolesc. Health Med. Ther.*, vol. 5, pp. 143-158, Aug. 2014.

[5] [5] M. Raj, S. Singh, K. Solanki, and R. Selvanambi, "An application to detect cyberbullying using machine learning and deep learning techniques," *SN Comput. Sci.*, vol. 3, no. 401, pp. 1-13, Sep. 2022.

[6] [6] P. Dedeepya et al., "Detecting cyber bullying on twitter using support vector machine," in *Proc. 3rd Int. Conf. Artif. Intell. Smart Energy (ICAIS)*, 2023, pp. 817-822.

[7] [7] P. Chindhuja, K. Darshini, M. Haritha, and J. Kowsalya, "Cyber bullying detection on social media network," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 4, pp. 4182-4186, Apr. 2023.

[8] [8] A. Desai, S. Kalaskar, O. Kumbhar, and R. Dhumal, "Cyber bullying detection on social media using machine learning," *ITM Web Conf.*, vol. 40, p. 03038, 2021.

[9] [9] N. Haydar and B. N. Dhannoon, "A comparative study of cyberbullying detection in social media for the last five years," *Al-Nahrain J. Sci.*, vol. 26, no. 2, pp. 47-55, Jun. 2023.

[10] [10] L. Seitz, "All the latest cyberbullying statistics for 2024," Broadband Search, 2024.

[Online]. Available: https://www.broadbandsearch.net/blog/cyber-bullying-statistics

[11] [11] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, no. 71, pp. 1-9, Mar. 2021.

[12] [12] PACER, "Cyberbullying: Definition," 2024. [Online]. Available: https://www.pacer.org/bullying/resources/cyberbullying/

[13] [13] UNICEF, "Cyberbullying: What is it and how to stop it," 2024. [Online]. Available: https://www.unicef.org/end-violence/how-to-stop-cyberbullying

[14] [14] J. Nambusi, "Cyberbullying: Effect on work place production," *Africa Int. J. Multidiscip. Res.*, vol. 2, no. 1, pp. 24-39, 2018.

[15] [15] P. C. Chukwuere, J. E. Chukwuere, and D. Adom, "The psychosocial effects of social media cyberbullying on students in selected African countries," *Acta Inform. Malaysia*, vol. 5, no. 2, pp. 62-70, Sep. 2021.

[16] [16] T. Mahlangu, "A review of automated detection methods for cyberbullying," *Digital Siege*, pp. 1-5, 2018.

[17] [17] H. Beghin, "The effects of cyberbullying on students and schools," Educational Research Report, 2020.

[18] [18] R. Slonje, P. K. Smith, and A. Frisén, "The nature of cyberbullying, and strategies for prevention," *Comput. Human Behav.*, vol. 29, no. 1, pp. 26-32, 2013.

[19] [19] R. Tas, N. Gistituati, and A. Ananda, "Cyberbullying in the digital age: A common social phenomenon," in *Proc. Int. Conf.*, 2020.

[20] [20] K. Varjas, J. Talley, J. Meyers, L. Parris, and H. Cutts, "High school students' perceptions of motivations for cyberbullying: An exploratory study," *West. J. Emerg. Med.*, vol. 11, no. 3, pp. 269-273, 2010.

[21] [21] J. H. Marco and M. P. Tormo-Irun, "Cyber victimization is associated with eating disorder psychopathology in adolescents," *Front. Psychol.*, vol. 9, pp. 1-7, Jun. 2018.

[22] [22] T. Vaillancourt, R. Faris, and F. Mishna, "Cyberbullying in children and youth: Implications for health and clinical practice," *Can. J. Psychiatry*, vol. 62, no. 6, pp. 368-373, Jun. 2017.

[23] [23] M. L. Ybarra, M. Diener-West, and P. J. Leaf, "Examining the overlap in internet harassment and school bullying: Implications for school intervention," *J. Adolesc. Health*, vol. 41, no. 6, pp. S42-S50, Dec. 2007.

[24] [24] G. Egeberg, S. Thorvaldsen, and J. A. Ronning, "The impact of cyberbullying and cyber harassment on academic achievement," in *Educational Research*, SensePublishers, 2016, pp. 183-204.

[25] [25] M. N. K. Karanikola, A. Lyberg, A. L. Holm, and E. Severinsson, "The association between deliberate self-harm and school bullying victimization and the mediating effect of depressive symptoms and self-stigma: A systematic review," *BioMed Res. Int.*, vol. 2018, pp. 1-37, 2018.

[26] [26] S. Batool, Rabia, and Yousaf, "Bullying in social media: An effect study of cyber bullying on the youth," *J. Commun. Stud.*, vol. 5, no. 1, pp. 45-58, 2017.

[27] [27] N. Alavi et al., "Relationship between bullying and suicidal behaviour in youth presenting to the emergency department," *J. Can. Acad. Child Adolesc. Psychiatry*, vol. 26, no. 2, pp. 70-77, 2017.

[28] [28] G. M. Abaido, "Cyberbullying on social media platforms among university students in the United Arab Emirates," *Int. J. Adolesc. Youth*, vol. 25, no. 1, pp. 407-420, Dec. 2020.

[29] [29] J. S. Hong, R. Navarro, and M. F. Wright, "Adolescent cyberbullying," IGI Global, pp. 1-22, Jul. 2024.

[30] [30] C. Canali, M. Colajanni, and R. Lancellotti, "Data acquisition in social networks: Issues and proposals," in *Proc. IEEE Workshop*, 2011, pp. 1-6.

[31] [31] M. Mancosu and F. Vegetti, "Collecting social media data," *Navigating Res. Data Methods*, vol. 1, no. 1, pp. 1-11, Jul. 2023.

[32] [32] E. Biswas, K. Vijay-Shanker, and L. Pollock, "Exploring word embedding techniques to improve sentiment analysis of software engineering texts," MSR Technical Report, 2019.

[33] [33] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: A

review," *EURASIP J. Wireless Commun.*, pp. 1-12, Dec. 2017.

[34] [34] M. Sarhan, S. Layeghy, N. Moustafa, M. Gallagher, and M. Portmann, "Feature extraction for machine learning-based intrusion detection in IoT networks," *Digit. Commun. Netw.*, vol. 10, no. 1, pp. 205-216, Aug. 2021.

[35] [35] F. Chollet, *Deep Learning with Python*. Shelter Island, NY: Manning Publications, 2018.

[36] [36] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN Comput. Sci.*, vol. 2, no. 6, pp. 1-20, Nov. 2021.

[37] [37] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1-16, Dec. 2019.

[38] [38] K. Kowsari et al., "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[39] [39] B. Kadam, "Cyberbullying detection using machine learning algorithms," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 5, pp. 1326-1328, May 2023.

[40] [40] N. A. Azeez, S. O. Idiakose, C. J. Onyema, and C. Van Der Vyver, "Cyberbullying detection in social networks: Artificial intelligence approach," *J. Cyber Security Mobility*, vol. 10, no. 4, pp. 745-774, 2021.

[41] [41] R. Pawar and R. R. Raje, "Multilingual cyberbullying detection system," in *Proc. IEEE Int. Conf. Electro Inf. Technol.*, May 2019, pp. 40-44.

[42] [42] A. M. Alduailaj and A. Belghith, "Detecting Arabic cyberbullying tweets using machine learning," *Mach. Learn. Knowl. Extr.*, vol. 5, no. 1, pp. 29-42, Mar. 2023.

[43] [43] C. VanHee et al., "Automatic detection of cyberbullying in social media text," *PLoS ONE*, vol. 13, no. 10, pp. 1-22, Oct. 2018.

[44] [44] K. E. Abdelfatah, G. Terejanu, and A. A. Alhelbawy, "Unsupervised detection of violent content in Arabic social media," in *Proc. CS IT Conf.*, Mar. 2018, pp. 1-7.

[45] [45] R. M. Rabii and M. M. Siraj, "Cyberbully detection using term weighting scheme and Naïve Bayes classifier," *Int. J. Innov. Comput.*, vol. 10, no. 1, pp. 35-39, May 2020.

[46] [46] C. Raj et al., "Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques," *Electronics*, vol. 10, no. 22, p. 2810, 2021.

[47] [47] M. H. U. Rahman, "Cyberbullying detection using natural language processing," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 5, pp. 5241-5248, May 2022.

[48] [48] K. S. Alam, S. Bhowmik, and P. R. K. Prosun, "Cyberbullying detection: An ensemble based machine learning approach," in *Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mobile Netw. (ICICV)*, Feb. 2021, pp. 710-715.

[49] [49] N. Yuvaraj et al., "Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification," *Comput. Electr. Eng.*, vol. 92, p. 107186, Jun. 2021.

[50] [50] M. M. Islam et al., "Cyberbullying detection on social networks using machine learning approaches," in *Proc. IEEE Asia-Pacific Conf. Comput. Sci. Data Eng. (CSDE)*, Dec. 2020, pp. 1-6.

[51] [51] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, M. Ahmed, and S. Islam, "A review on deep-learning-based cyberbullying detection," *IEEE Access*, vol. 11, pp. 45527-45550, May 2023.

[52] [52] M. Libina, G. Sasipriya, and V. Rajasekar, "An automatic method to prevent and classify cyber bullying incidents using machine learning approach," in *Proc. 8th IEEE Int. Conf. Sci. Technol. Eng. Math. (ICONSTEM)*, 2023, pp. 1-7.

[53] [53] S. Balakrishna, Y. Gopi, and V. K. Solanki, "Comparative analysis on deep neural network models for detection of cyberbullying on social media," *Ingeniería Solidaria*, vol. 18, no. 1, pp. 1-33, Jan. 2022.

[54] [54] R. Beniwal, S. Jha, S. Mehta, and R. Dhiman, "Cyberbullying detection using deep learning models in Bengali language," in *Proc. IEEE CONIT*, 2023, pp. 1-5.

[55] [55] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Syst.*, vol. 29, no. 4, pp. 1839-1852, Jun. 2023.

[56] [56] M. A. Al-Ajlan and M. Ykhlef, "Deep learning algorithm for cyberbullying detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 9, pp. 199-205, 2018.

[57] [57] S. Neelakandan et al., "Deep learning approaches for cyberbullying detection and classification on social media," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1-13, 2022.

[58] [58] A. Al-Marghilani, "Artificial intelligence-enabled cyberbullying-free online social networks in smart cities," *Int. J. Comput. Intell. Syst.*, vol. 15, no. 1, pp. 1-13, Dec. 2022.

[59] [59] A. Kiran and S. S. Kumar, "A comparative analysis of GAN and VAE based synthetic data generators for high dimensional, imbalanced tabular data," in *Proc. IEEE INOCON*, 2023, pp. 1-6.

[60] [60] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," *arXiv preprint*, pp. 1-12, Jan. 2018.

[61] [61] D. Chatzakou et al., "Weakly supervised cyberbullying detection using co-trained ensembles of embedding models," in *Proc. IEEE/ACM ASONAM*, 2018, pp. 479-486.

[62] [62] M. Dadvar and K. Eckert, "Cyberbullying detection in social networks using deep learning based models: A reproducibility study," *arXiv preprint*, 2022.

[63] [63] Y. Luo, X. Zhang, J. Hua, and W. Shen, "Multi-featured cyberbullying detection based on deep learning," in *Proc. IEEE ICCSE*, Aug. 2021, pp. 746-751.

[64] [64] A. Abulwafa, "A survey of deep learning algorithms and its applications," *Nile J. Commun. Comput. Sci.*, vol. 3, no. 1, pp. 28-49, 2022.

[65] [65] Nisha C. M. and N. Thangarasu, "Deep learning algorithms and their relevance: A review," *Int. J. Data Inform. Intell. Comput.*, vol. 2, no. 4, pp. 1-10, Dec. 2023.

[66] [66] M. Meenakshi, P. S. Babu, and V. Hemamalini, "Deep learning techniques for spamming and cyberbullying detection," in *Proc. IEEE ICNWC*, Apr. 2023, pp. 1-10.

[67] [67] U. Brandes, C. Reddy, and A. Tagarelli, "Are they our brothers? Analysis and detection of religious hate speech in the Arabic Twittersphere," in *Proc. IEEE/ACM ASONAM*, 2018, pp. 69-76.

[68] [68] G. Njovangwa and G. Justo, "Automated detection of bilingual obfuscated abusive words on social media forums: A case of Swahili and English texts," *Tanzania J. Sci.*, vol. 47, no. 4, pp. 1352-1361, Oct. 2021.

[69] [69] G. Njovangwa and G. Justo, "Automated detection of bilingual obfuscated abusive words on social media forums: A case of Swahili and English texts," *Tanzania J. Sci.*, vol. 47, no. 4, pp. 1352-1361, Oct. 2021.

[70] [70] M. Al-Ajlan and M. Ykhlef, "Firefly-CDDL: A firefly-based algorithm for cyberbullying detection based on deep learning," *Comput. Mater. Continua*, vol. 75, no. 1, pp. 19-34, 2023.

[71] [71] W. X. Zhao et al., "A survey of large language models," *arXiv preprint*, pp. 1-124, Mar. 2023.

[72] [72] D. Ottosson, "Cyberbullying detection on social platforms using large language models," Master's thesis, KTH Royal Institute of Technology, 2022.

[73] [73] B. Ogunleye and B. Dharmaraj, "The use of a large language model for cyberbullying detection," *Analytics*, vol. 2, no. 3, pp. 694-707, Sep. 2023.

[74] [74] K. L. Chiu, A. Collins, and R. Alexander, "Detecting hate speech with GPT-3," *arXiv preprint*, pp. 1-29, Mar. 2021.

[75] [75] Priya and S. Gupta, "Hate speech detection using OpenAI and GPT-3," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 12, no. 5, pp. 132-138, May 2022.

[76] [76] C. Morbidoni and A. Sarra, "Can LLMs assist humans in assessing online misogyny? Experiments with GPT-3.5," *arXiv preprint*, 2022.

[77] [77] T. H. Teng and K. D. Varathan, "Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches," *IEEE Access*, vol. 11, pp. 55533-55560, 2023.

[78] [78] P. K. Roy and F. U. Mali, "Cyberbullying detection using deep transfer learning," *Complex Intell. Syst.*, vol. 8, no. 6, pp. 5449-5467, Dec. 2022.

[79] [79] B. Ogunleye and B. Dharmaraj, "The use of a large language model for cyberbullying detection," *Analytics*, vol. 2, no. 3, pp. 694-707, Sep. 2023.

[80] [80] K. L. Chiu, A. Collins, and R. Alexander, "Detecting hate speech with GPT-3," *arXiv preprint*, pp. 1-29, Mar. 2021.

[81] [81] C. Morbidoni and A. Sarra, "Can LLMs assist humans in assessing online misogyny? Experiments with GPT-3.5," *arXiv preprint*, 2022.
[82] [82] A. Kumar and N. Sachdeva, "Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data," *Multimedia Syst.*, vol. 28, no. 6, pp. 2027-2041, Dec. 2022.
[83] [83] W. M. S. Yafooz, A. Al-Dhaqm, and A. Alsaeedi, "Detecting kids cyberbullying using transfer learning approach: Transformer fine-tuning models," in *Springer Proc.*, 2023, pp. 255-267.
[84] [84] T. H. H. Aldhyani, M. H. Al-Adhaileh, and S. N. Alsubari, "Cyberbullying identification system based on deep learning algorithms," *Electronics*, vol. 11, no. 20, pp. 1-19, Oct. 2022.
[85] [85] S. S. Aluru, B. Mathew, P. Saha, and A. Mukherjee, "A deep dive into multilingual hate speech classification," in *Lecture Notes in Computer Science*, vol. 12461, 2021, pp. 423-439.
[86] [86] N. A. Azeez, S. O. Idiakose, C. J. Onyema, and C. Van Der Vyver, "Cyberbullying detection in social networks: Artificial intelligence approach," *J. Cyber Security Mobility*, vol. 10, no. 4, pp. 745-774, 2021.
[87] [87] R. Pawar and R. R. Raje, "Multilingual cyberbullying detection system," in *Proc. IEEE Int. Conf. Electro Inf. Technol.*, May 2019, pp. 40-44.
[88] [88] B. Haidar, M. Chamoun, and A. Serhrouchni, "Multilingual cyberbullying detection system detecting cyberbullying in Arabic content," *arXiv preprint*, 2022.
[89] [89] I. Abishak, M. Kabilash, R. Ramesh, J. Sheeba, and D. Pradeep, "Unsupervised hybrid approaches for cyberbullying detection in Instagram," *Int. J. Eng. Res.*, vol. 10, no. 5, pp. 234-241, 2021.
[90] [90] A. M. Alduailaj and A. Belghith, "Detecting Arabic cyberbullying tweets using machine learning," *Mach. Learn. Knowl. Extr.*, vol. 5, no. 1, pp. 29-42, Mar. 2023.
[91] [91] C. Morbidoni and A. Sarra, "Can LLMs assist humans in assessing online misogyny? Experiments with GPT-3.5," *arXiv preprint*, 2022.
[92] [92] C. VanHee et al., "Automatic detection of cyberbullying in social media text," *PLoS ONE*, vol. 13, no. 10, pp. 1-22, Oct. 2018.
[93] [93] R. M. Rabii and M. M. Siraj, "Cyberbully detection using term weighting scheme and Naïve Bayes classifier," *Int. J. Innov. Comput.*, vol. 10, no. 1, pp. 35-39, May 2020.
[94] [94] C. Raj et al., "Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques," *Electronics*, vol. 10, no. 22, p. 2810, 2021.
[95] [95] G. Njovangwa and G. Justo, "Automated detection of bilingual obfuscated abusive words on social media forums: A case of Swahili and English texts," *Tanzania J. Sci.*, vol. 47, no. 4, pp. 1352-1361, Oct. 2021.
[96] [96] R. Albayari and S. Abdallah, "Instagram-based benchmark dataset for cyberbullying detection in Arabic text," *Data*, vol. 7, no. 7, pp. 1-11, Jul. 2022.
[97] [97] M. C. Babu and S. Pushpa, "Imbalanced dataset analysis with neural network model," in *Springer Proc.*, 2020, pp. 93-104.
[98] [98] S. Agrawal and A. Awekar, "Deep learning for detecting cyberbullying across multiple social media platforms," *arXiv preprint*, pp. 1-12, Jan. 2018.
[99] [99] B. Ogunleye and B. Dharmaraj, "The use of a large language model for cyberbullying detection," *Analytics*, vol. 2, no. 3, pp. 694-707, Sep. 2023.
[100] [100] C. Raj et al., "Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques," *Electronics*, vol. 10, no. 22, p. 2810, 2021.
[101] [101] T. Mahmud, M. Ptaszynski, J. Eronen, and F. Masui, "Cyberbullying detection for low-resource languages and dialects: Review of the state of the art," *Inf. Process. Manage.*, vol. 60, no. 5, p. 103454, Sep. 2023.
[102] [102] K. Maity, S. Saha, and P. Bhattacharyya, "Cyberbullying detection in code-mixed languages: Dataset and techniques," in *IEEE Conf. Proc.*, 2022, pp. 1692-1698.
[103] [103] I. Markov, N. Ljubešić, D. Fišer, and W. Daelemans, "Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection," in *EACL Proc.*, 2021, pp. 149-159.
[104] [104] M. Hamlett, G. Powell, Y. N. Silva, and D. Hall, "A labeled dataset for investigating

cyberbullying content patterns in Instagram," *Data Brief*, vol. 42, p. 108068, 2022.

[105] [105] C. Emmery et al., "Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity," *Lang. Resour. Eval.*, vol. 55, no. 3, pp. 597-633, Sep. 2021.

[106] [106] Y. Gong, G. Liu, Y. Xue, R. Li, and L. Meng, "A survey on dataset quality in machine learning," *Inf. Softw. Technol.*, vol. 162, p. 107268, Oct. 2023.

[107] [107] D. Van Bruwaene, Q. Huang, and D. Inkpen, "A multi-platform dataset for detecting cyberbullying in social media," *Lang. Resour. Eval.*, vol. 54, no. 4, pp. 851-874, Dec. 2020.

[108] [108] P. Röttger, D. Nozza, F. Bianchi, and D. Hovy, "Data-efficient strategies for expanding hate speech detection into under-resourced languages," *arXiv preprint*, 2022.

[109] [109] S. Subasree, N. K. Sakthivel, M. Shobana, and A. K. Tyagi, "Deep learning based improved generative adversarial network for addressing class imbalance classification problem in breast cancer dataset," *Int. J. Uncertainty Fuzziness Knowl.-Based Syst.*, vol. 31, no. 3, pp. 387-412, Jun. 2023.

[110] [110] N. V. Chereddy and B. K. Bolla, "Evaluating the utility of GAN generated synthetic tabular data for class balancing and low resource settings," in *Springer MIWAI Proc.*, 2023, pp. 48-59.

[111] [111] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1-16, Dec. 2019.

[112] [112] K. Kowsari et al., "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[113] [113] M. Subramanian et al., "A survey on hate speech detection and sentiment analysis using machine learning and deep learning models," *Electronics*, vol. 12, no. 6, p. 1313, 2023.

[114] [114] N. A. Azeez, S. O. Idiakose, C. J. Onyema, and C. Van Der Vyver, "Cyberbullying detection in social networks: Artificial intelligence approach," *J. Cyber Security Mobility*, vol. 10, no. 4, pp. 745-774, 2021.

[115] [115] R. Pawar and R. R. Raje, "Multilingual cyberbullying detection system," in *Proc. IEEE Int. Conf. Electro Inf. Technol.*, May 2019, pp. 40-44.

[116] [116] K. S. Alam, S. Bhowmik, and P. R. K. Prosun, "Cyberbullying detection: An ensemble based machine learning approach," in *Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mobile Netw. (ICICV)*, Feb. 2021, pp. 710-715.

[117] [117] M. Raj, S. Singh, K. Solanki, and R. Selvanambi, "An application to detect cyberbullying using machine learning and deep learning techniques," *SN Comput. Sci.*, vol. 3, no. 401, pp. 1-13, Sep. 2022.

[118] [118] R. Beniwal, S. Jha, S. Mehta, and R. Dhiman, "Cyberbullying detection using deep learning models in Bengali language," in *Proc. IEEE CONIT*, 2023, pp. 1-5.

[119] [119] C. Iwendi, G. Srivastava, S. Khan, and P. K. R. Maddikunta, "Cyberbullying detection solutions based on deep learning architectures," *Multimedia Syst.*, vol. 29, no. 4, pp. 1839-1852, Jun. 2023.

[120] [120] G. Njovangwa and G. Justo, "Automated detection of bilingual obfuscated abusive words on social media forums: A case of Swahili and English texts," *Tanzania J. Sci.*, vol. 47, no. 4, pp. 1352-1361, Oct. 2021.

[121] [121] D. Ottosson, "Cyberbullying detection on social platforms using large language models," Master's thesis, KTH Royal Institute of Technology, 2022.

[122] [122] B. Ogunleye and B. Dharmaraj, "The use of a large language model for cyberbullying detection," *Analytics*, vol. 2, no. 3, pp. 694-707, Sep. 2023.

[123] [123] K. L. Chiu, A. Collins, and R. Alexander, "Detecting hate speech with GPT-3," *arXiv preprint*, pp. 1-29, Mar. 2021.

[124] [124] S. Agrawal and A. Awekar# Comprehensive Review of Cyberbullying Detection Techniques and Datasets on Social Media Platforms

[125] [125] Maity, K., Saha, S., & Bhattacharyya, P. (2022). Cyberbullying detection in code-mixed languages: Dataset and techniques. *IEEE Conference Proceedings*, 1692-1698.

[126] [126] Venu, V. S., Shanmugasundaram, H., Seelam, M. R. R., et al. (2023). Detection of

cyberbullying on user tweets and Wikipedia text using machine learning. *IEEE Conference*, 327-332.

[127] [127] Bawane, V., Gaupale, V., Wakalkar, A., Rathod, K., & Shelke, S. (2023). Cyberbullying detection on social media. *International Journal of Innovations in Engineering and Science*, 8(7), 27-28.

[128] [128] Rachidi, R., Ouassil, M. A., Errami, M., et al. (2023). Social media's toxic comments detection using artificial intelligence techniques. *IEEE IRASET*, 1-6.

[129] [129] Albayari, R., & Abdallah, S. (2022). Instagram-based benchmark dataset for cyberbullying detection in Arabic text. *Data*, 7(7), 1-11.

[130] [130] Al-Hashedi, M., Soon, L. K., Goh, H. N., Lim, A. H. L., & Siew, E. G. (2023). Cyberbullying detection based on emotion. *IEEE Access*, 11, 53907-53918.

[131] [131] Mathur, S. A., Isarka, S., Dharmasivam, B., & Jaidhar, C. D. (2023). Analysis of tweets for cyberbullying detection. *IEEE ICSCCC*, 269-274.

[132] [132] Samalo, D., Martin, R., & Utama, D. N. (2023). Improved model for identifying the cyberbullying based on tweets of Twitter. *Informatica*, 47(4), 159-164.

[133] [133] Gamal, D., Alfonse, M., Jiménez-Zafra, S. M., & Aref, M. (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.

[134] [134] Bezrukov, A., Yashkina, V., Polishko, N., & Budilova, O. (2023). Multidiscursivity of cyberbullying as the epiphenomenon of social media communication in a screen society. *ASR: CMU Journal of Social Sciences and Humanities*, 10(2), 1-15.

[135] [135] Kumar, A., & Sachdeva, N. (2022). Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. *Multimedia Systems*, 28(6), 2027-2041.

[136] [136] Power, A., Keane, A., Nolan, B., & O'Neill, B. (2017). A lexical database for public textual cyberbullying detection. *Revista de Lenguas para Fines Específicos*, 23(2), 157-186.

[137] [137] Salazar, L. R., Garcia, N., Diego-Medrano, E., & Castillo, Y. (2020). Workplace cyberbullying and cross-cultural differences. *IGI Global*, 284-309.

[138] [138] Barlett, C. P., Seyfert, L. W., Simmers, M. M., et al. (2021). Cross-cultural similarities and differences in the theoretical predictors of cyberbullying perpetration. *Aggressive Behavior*, 47(1), 111-119.

[139] [139] Cammaerts, B., Anstead, N., Stupart, R., Smith, P. K., Görzig, A., & Robinson, S. (2018). Issues of cross-cultural variations in cyberbullying across Europe and beyond. *COST Action IS0801 Report*.

[140] [140] Smith, P. K., Görzig, A., & Robinson, S. (2019). Cyberbullying in schools: Cross-cultural issues. *Educational Research Review*, 27, 23-35.

[141] [141] Sheanoda, V., Bussey, K., & Jones, T. (2021). Sexuality, gender and culturally diverse interpretations of cyberbullying. *New Media and Society*, 26(1), 154-171.

[142] [142] Scott, J. E., & Barlett, C. P. (2023). Understanding cyber-racism perpetration within the broader context of cyberbullying theory: A theoretical integration. *Aggression and Violent Behavior*, 71, 101844.

[143] [143] Olweus, D., & Limber, S. P. (2018). Some problems with cyberbullying research. *Current Opinion in Psychology*, 19, 139-143.

[144] [144] Galarza, C. R., Pasquel, M. B., Cardenas, J. C., & Cedillo, P. (2022). Cyberbullying in the educational context. *AHFE International*, 68, 1-3.

[145] [145] Korzhov, H., & Yenin, M. (2022). Sociological dimensions of cyberbullying: Essence, consequences, and coping strategies. *Sociology: Theory, Methods, Marketing*, 4, 103-120.

[146] [146] Babu, M. C., & Pushpa, S. (2020). Imbalanced dataset analysis with neural network model. *Springer Proceedings*, 93-104.

[147] [147] Röttger, P., Nozza, D., Bianchi, F., & Hovy, D. (2022). Data-efficient strategies for expanding hate speech detection into under-resourced languages. *arXiv preprint*.

[148] [148] Van Bruwaene, D., Huang, Q., & Inkpen, D. (2020). A multi-platform dataset for detecting cyberbullying in social media.

Language Resources and Evaluation, 54(4), 851-874.

[149] [149] Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, 162, 107268.

[150] [150] Yi, P., & Zubiaga, A. (2023). Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media*, 36, 1-15.

[151] [151] Markov, I., Ljubešić, N., Fišer, D., & Daelemans, W. (2021). Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. *EACL Proceedings*, 149-159.

[152] [152] Hamlett, M., Powell, G., Silva, Y. N., & Hall, D. (2022). A labeled dataset for investigating cyberbullying content patterns in Instagram. *Data in Brief*, 42, 108068.

[153] [153] Al-Harigy, L. M., Al-Nuaim, H. A., Moradpoor, N., & Tan, Z. (2022). Building towards automated cyberbullying detection: A comparative analysis. *IEEE Access*, 10, 45512-45534.

[154] [154] Ahirwar, R., Ajay, M., Sathyabalan, N., & Lakshmi, K. (2022). Online harassment detection using machine learning. *IEEE ICICT*, 1222-1224.

[155] [155] Teng, T. H., & Varathan, K. D. (2023). Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches. *IEEE Access*, 11, 55533-55560.

[156] [156] Chew, S. X., Liew, J. S. Y., Fikry, W. A. L. W. I., & Ibrahim, N. F. (2022). Examining the generalizability of English cyberbullying detection models on Malay informal text using direct translation. *International Journal of Asian Language Processing*, 32(3), 2350005.

[157] [157] Kahate, S. A., & Raut, A. D. (2023). Design of a deep learning model for cyberbullying and cyberstalking attack mitigation via online social media analysis. *IEEE ICITIIT*, 1-7.

[158] [158] Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2021). A deep dive into multilingual hate speech classification. *Lecture Notes in Computer Science*, 12461, 423-439.

[159] [159] Yafooz, W. M. S., Al-Dhaqm, A., & Alsaeedi, A. (2023). Detecting kids cyberbullying using transfer learning approach: Transformer fine-tuning models. *Springer Proceedings*, 255-267.

[160] [160] Roy, P. K., & Mali, F. U. (2022). Cyberbullying detection using deep transfer learning. *Complex and Intelligent Systems*, 8(6), 5449-5467.

[161] [161] Aldhyani, T. H. H., Al-Adhaileh, M. H., & Alsubari, S. N. (2022). Cyberbullying identification system based on deep learning algorithms. *Electronics*, 11(20), 1-19.

[162] [162] Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., & Yeung, D. Y. (2019). Multilingual and multi-aspect hate speech analysis. *EMNLP Proceedings*, 4675-4684.

[163] [163] Kiran, A., & Kumar, S. S. (2023). A comparative analysis of GAN and VAE based synthetic data generators for high dimensional, imbalanced tabular data. *IEEE INOCON*, 1-6.

[164] [164] Kushwaha, A., & Pandey, R. S. (2023). Imbalanced dataset classification using fuzzy ARTMAP and computational intelligence techniques. *Indonesian Journal of Electrical Engineering and Computer Science*, 30(2), 909-918.

[165] [165] Subasree, S., Sakthivel, N. K., Shobana, M., & Tyagi, A. K. (2023). Deep learning based improved generative adversarial network for addressing class imbalance classification problem in breast cancer dataset. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 31(3), 387-412.

[166] [166] Chereddy, N. V., & Bolla, B. K. (2023). Evaluating the utility of GAN generated synthetic tabular data for class balancing and low resource settings. *Springer MIWAI Proceedings*, 48-59.

[167] [167] Robert, A. (2024). The challenges of detecting cyberbullying in a complex online world. Retrieved from Digital Defense Institute.

[168] [168] Jiang, A., & Zubiaga, A. (2024). Cross-lingual offensive language detection: A systematic review of datasets, transfer approaches and challenges. *ACM Computing Surveys*, 37(4), 1-35.

[169] [169] Viswanath, S., & Kumar, A. K. M. (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(11), 406-430.

[170] [170] Wen, X., & Li, W. (2023). Time series prediction based on LSTM-attention-LSTM model. *IEEE Access*, 11, 48322-48331.

[171] [171] Chandra, R., Goyal, S., & Gupta, R. (2021). Evaluation of deep learning models for multi-step ahead time series prediction. *IEEE Access*, 9, 83105-83123.

[172] [172] Abarna, S., Sheeba, J. I., & Devaneyan, S. P. (2023). A novel ensemble model for identification and classification of cyber harassment on social media platform. *Journal of Intelligent and Fuzzy Systems*, 45(1), 13-36.

[173] [173] Anwar, G. B., & Anwar, M. W. (2022). Textual cyberbullying detection using ensemble of machine learning models. *IEEE ICIT*, 1-7.

[174] [174] Zhao, W. X., Zhou, K., Li, J., et al. (2023). A survey of large language models. *arXiv preprint*, 1-124.

[175] [175] Nee, C. N., Samsudin, N., Chuan, H. M., et al. (2023). The digital defence against cyberbullying: A systematic review of tech-based approaches. *Cogent Education*, 10(1), 2288492.

[176] [176] Barlett, C. P., Dewitt, C. C., Maronna, B., & Johnson, K. (2018). Social media use as a tool to facilitate or reduce cyberbullying perpetration: A review focusing on anonymous and nonanonymous social media platforms. *Violence and Gender*, 5(3), 147-152.

[177] [177] Elsafoury, F., Katsigiannis, S., Pervez, Z., & Ramzan, N. (2021). When the timeline meets the pipeline: A survey on automated cyberbullying detection. *IEEE Access*, 9, 103541-103563.

[178] [178] Al-Harigy, L. M., Al-Nuaim, H. A., Moradpoor, N., & Tan, Z. (2022). Deep pre-trained contrastive self-supervised learning: A cyberbullying detection approach with augmented datasets. *IEEE CICN*, 16-22.

[179] [179] Guo, X., Anjum, U., & Zhan, J. (2022). Cyberbully detection using BERT with augmented texts. *IEEE Big Data*, 1246-1253.

[180] [180] Qiu, J., Hegde, N., Moh, M., & Moh, T. S. (2022). Investigating user information and social media features in cyberbullying detection. *IEEE Big Data*, 3063-3070.

[181] [181] Luo, Y., Zhang, X., Hua, J., & Shen, W. (2021). Multi-featured cyberbullying detection based on deep learning. *IEEE ICCSE*, 746-751.

[182] [182] Indumathi, V., & Megala, S. S. (2023). Enhanced multi-label classification model for bully text using supervised learning techniques. *Springer Proceedings*, 763-778.

[183] [183] Qiu, J. (2021). Multimodal detection of cyberbullying on Twitter. *Master's Thesis, San Jose State University*, 1-38.