AI-BASED APPROACHES FOR CYBERBULLYING AND TOXIC COMMENT DETECTION

Iqra Zaheer[1], Astha Singh[1], Dev Singh[2]

[1]B. Tech Student, Allenhouse Institute of Technology, Kanpur (India)
[2]B. Tech Student, Allenhouse Institute of Technology, Kanpur (India)
[3]Professor, Allenhouse Institute of Technology, Kanpur (India)

DOI:10.71182/aijmr.2512.0302.2002

ABSTRACT

The growth of digital communication technologies has made it easier to pose threats to one's psychological well-being and even one's safety due to cyberbullying and online toxicity. AIs have become one of the few technologies capable of mitigating the problem of harmful interactions and toxicity digitally via automated content moderation. The purpose of this review is to assess the development of AI-enabled systems in the automated moderation of cyberbullying and toxicity to digitally detect comments and to evaluate how the systems have adapted from using rules and algorithms to more advanced systems using deep learning and other new digital technologies.

This paper also reviews the criticisms and challenges surrounding benchmark datasets, evaluation research, text and discourse structures and even inconsistency and disregard for key AI ethics in cyber moderation, responsibility, algorithm bias, explanatory disaggregation, and even digital-privacy. We also examine new concepts of research in context integrating multimodal compositions for discourse, improvement of emotion, and computational efficiency in new forms of deep learning architectures. This advances digital moderation using low computational resources. This comprehensive review is intended to detail, critique, and analyze the breadth of research in cyberbullying and online toxicity.

Keywords: *Cyberbullying Detection, Toxic Content Classification, Natural Language Processing, Machine Learning, Deep Learning, Transformers, Multilingual AI*

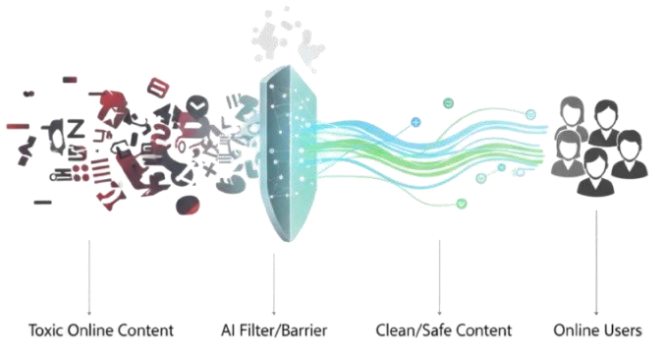
1. INTRODUCTION

1.1 Background and Motivation

The rapid evolution of digital communications technologies has changed the way people and organizations produce, obtain, and share information. On the other hand, such connectivity also amplifies potentially injurious online behaviors, including cyberbullying, hate speech, trolling, and targeted online harassment. Newly conducted surveys by the Pew Research Center indicate that about 37% of users have faced at least one form of digital harassment, therefore illustrating the extent and gravity of the situation. Victims often face longer-term psychological repercussions, which include anxiety, depression, reduced self-esteem, and social isolation.

Effectively detecting and moderating toxic online content remains a significant challenge due to the nuanced and context-dependent nature of human communication. Explicit intent is usually obscured in sarcastic expressions, cultural references, regional idioms or slang, and multilingual code-switching, such as Hinglish or Spanglish. Apart from this, implicit forms of aggression most often slip by traditional moderation strategies. These classical approaches—pattern-based filtering and manual review—lack linguistic sophistication to capture these subtleties, leading to high rates of misclassification and unscalable moderation workflows [1] [2]. Limitations like this have gained momentum toward more advanced AI-driven moderation systems that can learn contextual and semantic cues.

Fig. 1: Conceptual Representation of AI-Driven Content Filtering



1.2 Scope and Contributions

Given the explosion in reliance on automated moderation pipes, there is an imminent need for a

systematic understanding of currently available AI-based approaches for the detection of toxic content. This review synthesizes contemporary research developments in this area and provides insight into how we have progressed from classical machine learning models to state-of-the-art transformer-based architectures. Our contributions are threefold: (1) we analyze current methodological trends and benchmark performances across diverse linguistic and cultural contexts; (2) we identify key technological gaps, dataset limitations, and challenges to fairness, robustness, and generalization; and (3) we outline future research trajectories with a view to supporting the development of more reliable, equitable, and scalable AI-driven content moderation systems.

2. PROBLEM FORMULATION

Despite marked advances in automated content moderation, current systems still fall short in regularly detecting cyberbullying and toxic discourse under operational conditions. The limitations of these latter approaches can be summarized as follows:

i) Linguistic Limitations: Models that are trained primarily on high-resource languages have shown significantly diminished performance when presented with low-resource languages, dialectal variations, and code-switched communication [3]. This narrows their applicability in multilingual digital ecosystems.

ii) Contextual Ambiguity: Most systems currently lack the functionality that would allow them to capture subtle or implicit forms of toxicity, such as sarcasm, coded expressions, and culturally-bound references; this leads to inconsistent classification and less reliability [4].

iii) Operational Latency: Most moderation pipelines rely on user reporting or manual reviews, therefore introducing huge latency prior to toxic content being acted upon and allowing those interactions to continue causing harm.

iv) Computational Overheads: Modern transformer-based architectures involve massive computational and memory costs, making their real-time deployment and usage in resource-constrained settings infeasible [5].

These challenges together point to the fact that linguistic diversity, contextual complexity, and real-time demands on online platforms are not adequately supported by currently deployed toxic content detection systems.

The central problem that this study addresses, therefore, is the lack of a scalable, linguistically inclusive, and contextually robust framework for accurately and efficiently detecting toxic content in heterogeneous digital environments.

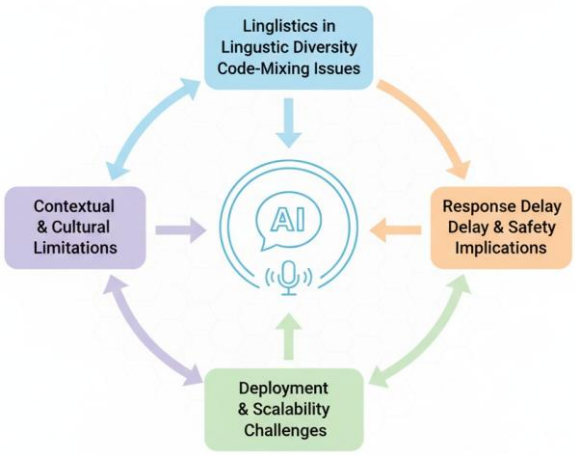
3. LITERATURE REVIEW

3.1 Traditional Machine Learning Approaches

One of the primary methods of automated toxicity detection in early works was the use of classical machine learning algorithms such as Naïve Bayes, Logistic Regression, Random Forests, and Support Vector Machines (SVMs). In these methods, the features highly depended on manual extraction from the data, e.g., TF-IDF vectors, bag-of-words models, sentiment lexicons, and syntactic n-grams.

For instance, landmark research by [6] illustrated that SVM and Logistic Regression classifiers could obtain close to state-of-the-art results (F1-score: 0.78–0.85) for the task of explicit hate speech detection in English social media texts. In the same vein, [1] created rule-based systems that were proficient in overtly toxic lexicon identification. However, these systems had little ability to recognize context-dependent or implicitly harmful expressions. Even though these conventional techniques brought about advantages in terms of speed and interpretability of the models [7], the success of these methods was limited due to their semantic understanding deficiency and the inability to capture intricate linguistic patterns [8].

Fig. 2: Multidimensional challenges in toxic content detection spanning linguistic, contextual, and computational domains



3.2 Deep Learning Architectures

The advent of deep learning catalyzed significant advancements through architectures capable of learning hierarchical feature representations. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory networks (LSTMs) demonstrated superior performance in capturing contextual dependencies and semantic nuances.

Comparative studies revealed that hybrid architectures combining CNNs for local feature extraction with LSTMs for sequential modeling achieved state-of-the-art performance (accuracy: 89–93%) across multiple toxicity detection benchmarks [12]. However, these models exhibited limitations including substantial data requirements, vulnerability to adversarial examples, and diminished performance on low-resource languages and code-mixed content.

3.3 Transformer-Based Models

As noted by [13], the introduction of the transformer model brought about changes in the field of natural language processing thanks to the development of

self-attention mechanisms that handle long-distance contextual dependencies. Boosting the Ensembling of Roberta transformer models [14], [15], [16], and their successors demonstrates performance gains on multilingual toxicity detection tasks.

The domain-adapted transformer [11] HateBERT shows high effectiveness in detecting hate speech in other specialized contexts. The most recent of these are LLM [17] [18], where large language models are used to detect hate speech in a specific context. Although transformer models demonstrate high performance, they are also very demanding in terms of computational power, energy, and large volumes of annotated data needed to fine-tune [5].

TABLE I: Performance Comparison of Cyberbullying and Toxicity Detection Models

Model / Method	Dataset Used	Key Observations	F1/Accuracy (%)
Logistic Regression, SVM	Hate Speech Dataset [6]	Performs well for explicit hate words but weak in contextual and sarcasm detection	~80
CNN, LSTM	Twitter Corpus	Captures deeper contextual cues and provides better semantic representation	84–86
CNN + GRU	OffensEval [10]	Handles sequential dependencies effectively	85
MBERT	TRAC-2 (Hinglish) [9]	Multilingual transformer achieving strong results on code-mixed text	90
HateBERT [11]	Reddit, Twitter	Domain-specific fine-tuning enhances contextual precision and recall	91
Hybrid LSTM + TF-IDF	HASOC	Balances interpretability with contextual learning	88
Hybrid ML + Transformer (Proposed)	HASOC, Kaggle	Shows highest multilingual adaptability and overall performance	92

3.4 Datasets and Evaluation Benchmarks

The development of comprehensive datasets has been instrumental in advancing toxicity detection research. Table II illustrates the benchmarks that are most often employed in current empirical work. These benchmarks contrast in breadth and depth of coverage and in varying patterns of annotation and labelling, which determine and alter the generalization and inter-domain transfer of the models.

TABLE II: Benchmark Datasets for Cyberbullying and Toxicity Detection

Dataset	Content Focus	Size
Kaggle Toxic Comment (2018)	Multi-label toxicity classification	159k samples
TRAC [9]	Aggression identification	15k samples
HASOC	Hate speech and offensive content	12k samples
OLID / OffensEval [10]	Offensive language detection	14k samples
ADHAR [20]	Hate speech detection	10k samples
Multimodal Hate Dataset [19]	Video-based hate speech	8k samples

4. DISCUSSION

The convergence of technology, linguistic complexity, and ethical considerations is evident in the review of AI-based approaches to detect cyberbullying and toxic comments, and this convergence is an important aspect of how systems perform. The discussion takes a holistic approach by combining the three elements to offer a complete view of the current capabilities and limitations of these systems without treating the three elements as separate.

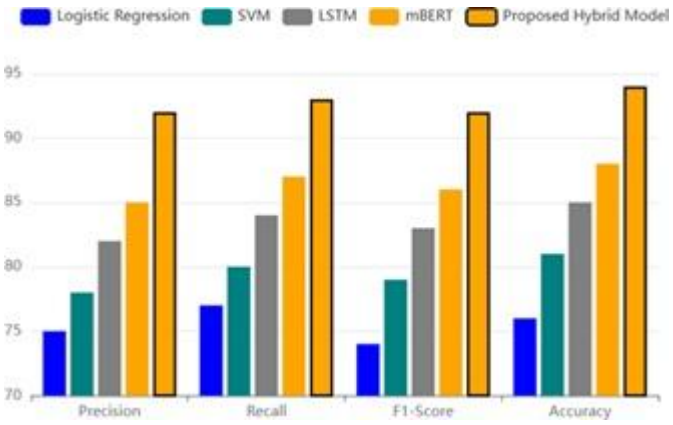
4.1 Synthesis of model evolution and empirical performance

The body of research reviewed provides an increasing trend from traditional, feature-based modelling techniques to Contextual and Transformer-based modelling, as a function of improved detection performance, due in part to their ability to capture longer-range logical and pragmatic relationships. In addition, due to the deployment of LLM(s), there has been significant development of the ability to provide enhanced contextualization and few-shot adaptation of the Transformer architecture. However, the literature indicates that, in many lower-resource environments, the higher level of performance justifies the higher quantity of labelled data sets, computational capacity, and engineering expertise to attain this level of performance. Consequently, improvements on benchmark metrics do not provide the same degree of assurance for reliable and manageable deployment of Transformer architectures.

4.2 Language variation, code-mixing, and dataset limitations

A consistent finding in the literature is that the presence of low-resource languages and code-mixed text (such as Hinglish) will have a significant negative impact on the ability to generalize across multiple domains. Although multilingual benchmarks and pretrained models such as XLM-R and datasets like AADHAR have provided important baselines, there are still gaps in performance due to various factors, including: (1) a lack of annotated examples of code-mixed constructions; (2) the presence of inconsistent orthography and informal registers in social media texts; Evidence indicates that improving performance in this area will require not only the development of better models but also targeted data collection for code-mixed and low-resource languages, culturally informed annotation schemes, and transfer methods that incorporate models of codeswitching behavior.

Fig. 3: Distribution of Dataset Categories and Linguistic Coverage in Toxicity Detection Research



4.3 Integrated view of Implicit and Contextual Toxicity

Implicit toxicity is also one of the hardest forms of toxicity, alongside sarcasm and humour that is specific to a culture, to detect because of the contextual nature [4] [1]. More often than not, models tend to mistake friendly banter for harassment, and so, there is a high degree of false positives occurring. This demonstrates the need for sophisticated models that possess contextual understanding, richer datasets for annotation, and also consider multimodal approaches [19].

4.4 Ethical and Computational Trade-offs

Studies have demonstrated that implicit toxicity, which refers to sarcasm, humor, and culturally dependent disparages, presents researchers with their greatest challenge for detection. Research findings indicate that these instances of implicit toxicity should not be considered isolated challenges but rather as components of a larger issue: "contextual toxicity." Contextual toxicity necessitates (1) a broader understanding of discourse-level context and user context, (2) an understanding of annotation through a cultural lens, and (3) a greater number of multimodal indicators (such as accompanying images and audio) when possible, to assist in determining intent. When research or technology fails to incorporate these contextual elements, it results in a high rate of false positives (incorrectly identifying harmless banter as toxic) and false negatives (missing subtle forms of harm). Consequently, the best approach to reduce instances of these types of error is to enhance the processes used to annotate contextual toxicity, enrich data related to users and conversations, and incorporate multiple forms of media into multimodal modelling.

4.5 Practical Implications

The review provides a series of recommendations for practitioners, including the need to implement contextualized models for high-stakes detection tasks in conjunction with model compression and human review, prioritising the collection of targeted datasets for code-mixed and low-resource languages before starting on a full-scale deployment; and placing a strong emphasis on regular continuous monitoring and evaluation for fairness in order to quickly identify any disproportionate error rates. These recommendations will allow you to convert proprietary gains made on the benchmarks to safer and more equitable operational systems in the real world.

5 TASK FORMULATION

Toxic comment detection is typically treated as a binary or multi-class classification problem. Let a comment $C = (w_1, w_2, \ldots, w_n)$ be a sequence of words from a vocabulary V. Each token w_i is embedded as a dense vector $x_i \in R^d$. The model learns a mapping function that predicts the toxicity probability of a given comment as:

$$f(C) = \sigma(W.\bar{x} + b) \quad (1)$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

σ denotes the activation function, and $f(C)$ gives the probability that the comment is toxic ($y = 1$) [14].

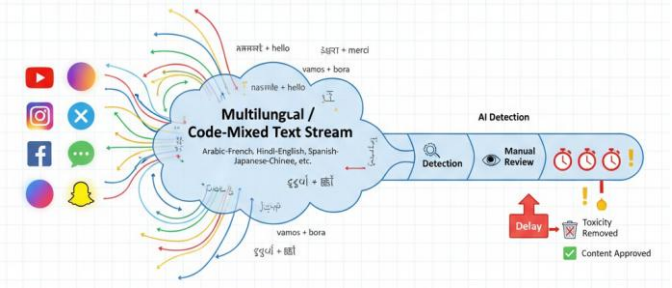
The model is trained to minimize the binary cross-entropy loss function defined as:

$$L = -\frac{I}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \quad (2)$$

where $yi \in \{0, 1\}$ represents the ground truth label for it i^{th} sample, and $\hat{y}_i$ is the predicted probability for the i^{th} sample.

This formulation is consistent in classical and modern models, including Logistic Regression, LSTM, and BERT, because the underlying objects remain the same, which convert textual inputs into vector representations (x_i) [17]. These representations are aggregated into a fixed-length vector, mapping it to a toxicity probability. The main distinction of these models can be how they compute or contextualize the embeddings. Classic models use a bag-of-words or TF-IDF features, whereas neural models derive contextual embeddings from data.

Fig. 4: Architecture of Traditional Machine Learning Pipeline for Toxic Content Classification, Comprising Feature Extraction, Selection, and Classification Components



6 FUTURE RESEARCH DIRECTIONS

6.1 Federated Learning for Privacy

Federated or distributed learning models enable local training on users' devices or in organizational settings while keeping raw data out of central servers. This distributed approach guarantees privacy and fulfils data protection laws like GDPR. Toxic comment detection through federated learning in cross-platform scenarios allows global models to iteratively improve without leaking sensitive user data [20]. Future works may investigate the combination of secure aggregation and differential privacy to further enhance security and mitigate risks of model inversion, data leakage and membership inference attacks.

6.2 Multimodal Toxicity Detection

Multiple modalities, including written content, visual content, audio content, and video content, are frequently used in modern online communication, such as voice notes, reels, and memes. To identify toxicity conveyed through tone, visual symbols, or sarcasm, future studies should focus on multimodal fusion architectures that combine linguistic, visual, and acoustic features. Deeper semantic alignment among text and imagery can be achieved by utilizing cross-modal attention mechanisms and vision-language models (such as CLIP and Flamingo), which will result in more exact and context-sensitive moderation systems [19].

6.3 Emotion and Context-Aware Models

Differentiating between genuine harassment, constructive criticism, and friendly banter requires an understanding of the conversational context and emotional tone [21]. Affective computing methods and contextual embeddings that dynamically modify predictions based on sentiment, conversation history, and user intent can be incorporated into future models. The interpretability and human-likeness in moderation systems can be enhanced by incorporating psychological signals and empathy-aware modelling, which reduces the likelihood of false positives in complex social interactions.

6.4 Explainable and Ethical AI

The need for transparent and morally sound AI is growing because computerized moderation systems have a greater impact on online conversation. Interpretability—offering clear explanations for classification results via attention visualization, counterfactual explanations, or feature attribution techniques—should be a key component of future frameworks [7]. Furthermore, equitable AI governance should limit algorithmic bias and unintentional censorship while encouraging accountability in content moderation across socioeconomic groups, cultural contexts, and languages.

6.5 Lightweight and Low-Resource Models

Large-scale transformers have remarkable performance, but their high memory footprint and computational cost prevent them from being used in low-resource settings like edge platforms or mobile devices [18]. To compress big models without causing appreciable performance degradation, future research should investigate efficient transformer variants (such as DistilBERT, ALBERT, and TinyBERT) along with knowledge distillation techniques. Toxicological detection models can be made more scalable, sustainable, and available for real-time, portable moderation in diverse global locations through research on quantization, pruning, and parameter-efficient fine-tuning (PEFT) methods.

6.6 Large Language Models (LLMs) for Toxicity Moderation

Large Language Models (LLMs) such as GPT-4, LLaMA, and Mistral represent a new direction for toxicity detection, and these models have shown strong reasoning capabilities, emerging contextual understanding, and multilingual generalization. Future work may investigate instruction tuning, reinforcement learning from human feedback-RLHF- and alignment strategies to improve LLM moderation reliability. However, issues such as hallucinations, high inference cost, safety inconsistencies, and sensitivity to adversarial prompting remain an unsolved problem that requires serious consideration [15].

6.7 Open Challenges in Toxicity detection

Even after making such progress, some challenges remain:

- **Data Imbalance:** Toxic examples are often rare making models biased towards non-toxic labels.
- **Real-time deployment:** Ensuring low latency and high accuracy in fast moving social platforms remain difficult.
- **Robustness:** Models struggle against adversarial attack, paraphrasing and code switching.

These challenges gives and important direction to future research and benchmarking [16].

7 CONCLUSION

While AI models for detecting toxicity are improving, there are still areas where AI models struggle, such as identifying implicit toxicity, dealing with multilingual or code-mixed content and being fair to all user groups. Transformers show superior performance when compared to traditional methods. However high resource requirements along with low generalization capability of these models present challenges to scalability efforts while ensuring equity issues in the context of community moderation systems. In order to successful enable and facilitate continued advancement, it is necessary to develop further definitions of multimodal understanding, Explainability, along with incorporating privacy-preserving solutions through Federated Learning and integration of lightweight Transformers approaches to facilitate the continued development of reliable, fair, contextually aware moderation systems that will assist in creating safe and inclusive Digital Spaces for all users.

REFERENCES

1) Waseem, Z., & Hovy, D. (2017). *Context is key: The case of hate speech detection*. In Proceedings of the SocialNLP Workshop. https://aclanthology.org/W17-1105/
2) Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). *The risk of racial bias in hate speech detection*. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). https://aclanthology.org/P19-1163/
3) Raza, S., & Chatrath, V. (2024). HarmonyNet: Navigating hate speech detection. *Natural Language Processing Journal*. https://www.sciencedirect.com/science/article/pii/S2949719124000463
4) Prabhu, R., et al. (2025). A comprehensive framework for multi-modal hate speech detection. *Nature Scientific Reports*. https://www.nature.com/articles/s41598-025-94069-z
5) Kar, P., et al. (2023). Sentimental analysis hate speech detection using hybrid DGRNN. *Engineering Applications of Artificial Intelligence*. https://www.sciencedirect.com/science/article/pii/S0952197623013271
6) Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media (ICWSM). https://ojs.aaai.org/index.php/ICWSM/article/view/14955
7) Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). https://dl.acm.org/doi/10.1145/2939672.2939778
8) Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4). https://dl.acm.org/doi/10.1145/3232676
9) Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2020). Benchmarking aggression identification (TRAC-2). https://aclanthology.org/2020.trac-1.1/
10) Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Predicting offensive posts. In Proceedings of NAACL-HLT. https://aclanthology.org/N19-1144/
11) Zampieri, M., et al. (2021). HateBERT: Retraining BERT for abusive language. https://aclanthology.org/2020.alw-1.12/
12) Rajput, A., et al. (2022). Hybrid deep learning framework for offensive text. https://arxiv.org/abs/2204.07965
13) Vaswani, A., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS). https://papers.nips.cc/paper/7181-attention-is-all-you-need
14) Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT. https://aclanthology.org/N19-1423/
15) Liu, Y., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. https://arxiv.org/abs/1907.11692
16) Conneau, A., et al. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of ACL. https://aclanthology.org/2020.acl-main.747/
17) Choudhary, M., et al. (2025). LLM-GPT2 fine-tuning approaches. https://www.sciencedirect.com/science/article/pii/S1877050925016461
18) Zhou, Y., Gupta, T., & Khan, S. (2025). LLM-Hate: Large language models for hate speech. https://ieeexplore.ieee.org/document/10752193
19) Perez, J. P., Muzzi, D., & Cámara, J. (2020). Multimodal classification of hate speech in videos. In Proceedings of WACV. https://ieeexplore.ieee.org/document/9093437
20) Charfi, A., et al. (2024). Hate speech detection with ADHAR dataset. https://www.frontiersin.org/articles/10.3389/frai.2024.1391472
21) Xie, H. S., et al. (2023). A review of hate speech detection. https://aisel.aisnet.org/digit2023/15/