A Survey of Abusive Language Detection on Online Platforms: Policy Analysis and Neural Network Solutions

Ahmadjamy Kohistani
Department of Computer Science and Engineering
South Asian University
New Delhi, India
Ahmadjamykohistani@gmail.com

Shachi Sharma
Department of Computer Science and Engineering
South Asian University
New Delhi, India
shachi@sau.int

Muhammad Abulaish
Department of Computer Science and Engineering
South Asian University
New Delhi, India
abulaish@sau.ac.in

***Abstract*—The internet connects billions of users through online social media platforms. While enabling global communication and commerce, these platforms face increasing challenges from abusive content, which poses serious risks to user well-being and the integrity of online interactions. Detecting such abuse is complicated by its subjective nature and the inconsistent content moderation policies across platforms. This paper presents a comprehensive comparative analysis of moderation policies from four major platforms, Facebook, Google, X (formerly Twitter) and Amazon. Thereafter, it proposes a generic framework for a policy-aware automatic abusive language detection system. A prototype implementation using BERT and neural networks is developed, and its performance is evaluated on a sample dataset to assess the feasibility of such a system. The paper also reviews key challenges in designing policy-aware, generalizable abusive language detection methods. This work aims to support future research by offering insights into developing more effective, fair, and context-sensitive detection models that align with platform-specific policies.**

***Index Terms*—Social Network Analysis, Abusive Language Detection, Hate Speech, Offensive Language, Neural Network.**

I. Introduction

Nowadays online platforms serve as the primary medium for global communication, content creation, information dissemination, and social engagement [1], [2]. This is evident from the continuous growth in the number of users on social networks. According to Strata[1], more than 5 billion people have been using social media world-wide in 2024 and this number is expected to exceed 6 billion by year 2028. These users generate vast volumes of data in the form of posts, tweets, comments, messages, replies etc. As user-generated content continues to grow, ensuring respectful and constructive interactions has become increasingly critical. Recent studies show that the use of abusive and inappropriate language as well as expressions are increasingly becoming common [3]. Therefore, a considerable research effort has been devoted to moderating online content and identifying abusive language to safeguard users and uphold the integrity of these platforms [4]. Effectively detecting and removing harmful content not only enhances the user experience but also contributes significantly to user retention and the sustained growth of social networks.

In the context of online social media platforms, abusive language refers to user-generated text that is harmful, offensive, or disruptive to others and violates community guidelines or widely accepted standards of respectful communication. It can take various forms, including offensive speech, cyberbullying, hate speech, aggression, rumors, and toxic comments [4]–[7]. Often, such content carries mixed sentiments or subtle intentions, making it difficult to detect through simple keyword-based methods. Therefore, automated and intelligent solutions are essential for identifying abusive content, as manual moderation is both time-consuming and inconsistent due to the vast volume and evolving nature of language. Consequently, online platforms increasingly rely on automated systems to detect abusive language effectively and at scale. In this regard, techniques ranging from traditional Machine Learning (ML) to advanced Deep Learning (DL) methods particularly those incorporating Natural Language Processing (NLP) models such as Bidirectional Encoder Representations from Transformers (BERT) have been successfully employed [8]. While several research efforts have focused on abusive language detection, many remain limited in scope, addressing specific types of abuse or targeting particular platforms. There is still a lack of a generalized, adaptable framework that can operate across platforms with similar content moderation policies. This gap poses a significant challenge in developing robust moderation systems.

Designing an effective detection framework requires the careful selection of appropriate tools, high-quality training datasets, and context-aware algorithms capable of recognizing nuanced and implicit forms of abusive language. The objective of this paper is to facilitate researchers by providing an in-depth overview of moderation policies adopted by major online platforms with respect to abusive language, reviewing

[1] https://www.statista.com/topics/1164/social-networks/

TABLE I: Timeline of Abusive Language Detection Techniques

Period	**Dominant Methods**	**Milestones**
Before 2015	Lexicons, Feature Engineering + Machine Learning	TF-IDF and SVM for hate speech classification
2016–2018	Deep Learning (CNN, LSTM, GRU)	Introduction of attention mechanisms; Shared tasks like SemEval (HatEval, HASOC); hybrid DL
2018–2020	Pretrained Transformers (BERT)	BERT fine-tuning; HurtBERT, OffensEval models improve context sensitivity
2019–2021	Specialized Transformers	HateBERT retrained on Reddit; community embeddings improve domain generalization; focus on thread-level context.
2022–2023	Multilingual Transformers, Fairness-aware Models	ArabicBERT, BanglaHateBERT; studies on racial bias and dialect fairness; datasets expanded across languages.
2023 onwards	Implicit Abuse Detection, Adversarial Robustness	Focus on obfuscated and euphemistic abuse; detection of subtle language cues; cross-platform policy alignment emerges.

existing detection approaches, including those based on neural networks such as BERT and its derivatives, and proposing a generalized solution for automatic abusive language detection. In summary, the key contributions of this paper can be outlined as follows:

- A comparative analysis of major content moderation policies.
- A new architecture for integrating policy constraints with machine learning.
- A prototype experiment validating feasibility of the proposed architecture.

The rest of the paper is organized as follows. Section II provides a discussion on the related work. An overview of policies of online platforms and automatic abusive language detection is outlined in section III. The policy-aware abusive language detection framework along with policy analysis methodology is presented in section IV. The results of empirical validation of this framework are presented in section V. The challenges in designing effective automatic abusive language detection methods are discussed in section VI and those of policy-aware abusive language detection are contained in section VII. The last section contains concluding remarks.

II. Related Work

A. History and Progress

Abusive language detection has received increasing attention in recent years, particularly in the NLP and computational linguistics communities. However, the area has a history of more than a decade. Starting with manual lexicons and rule-based systems, the research proceeded towards using TF-IDF, bag-of-words, character n-grams, part-of-speech tags, and sentiment lexicons as features for machine learning classifiers like SVM, Naïve Bayes, and Logistic Regression. A notable work worth mentioning is by Waseem and Hovy who applied these features to detect racism and sexism on Twitter [9]. Besides, multilingual datasets, such as Arabic and code-mixed English–Hindi, were introduced to demonstrate that capability of lexicon-based ML beyond English [10], [11]. Later, the DL methods like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTMs), and Gated Recurrent Units (GRUs) increasingly replaced feature engineering by learning c text representations directly. The CNNs improved precision by capturing local patterns of abuse (e.g., insult styles) efficiently, while LSTMs and GRUs offered better recall by modeling longer contextual dependencies in text sequences. Hybrid models, combining CNN with BiLSTM/BiGRU, excelled on short text like tweets and YouTube comments, often outperforming single-architecture models [12].

With the invent of BERT (Devlin et al., 2018) and similar transformers, the research on abusive language detection was also undergone significant change. Fine-tuned BERT on tasks like SemEval OffensEval and OLID consistently surpassed earlier DL and ML baselines [13]. Specialized models such as HurtBERT combined lexical features with BERT, capturing nuanced signals for domain-general and cross-domain detection [14], HateBERT retrained on Reddit communities banned for harassment, outperformed base BERT on several English abuse-detection benchmarks and showed better cross-dataset portability [15]. Many performance studies established the BERT's ability to handle obfuscated or encoded insults and toxic language effectively [16]. A comprehensive survey spanning on 29 research papers related to transformer-based architectures (with variants like RoBERTa, ELECTRA, ArabicBERT, and mBERT/XLM-R) from 2020-24 established the effective of BERT in abusive language detection in multiple languages including Arabic, Spanish, Hindi, and Russian [17]. A chronological development in the area of abusive language detection is provided in Table I.

B. Key Literature

Numerous studies have addressed the challenges of identifying various forms of harmful content on social-media platforms. A core issue across these works is the difficulty in defining and categorizing abusive language consistently due to differences in platform-specific moderation policies and socio-cultural contexts. The popular supervised learning approaches attempt to classify abusive content using binary (abusive versus non-abusive) or ternary (offensive, abusive, or neutral) classification models. However, these methods are often designed within the constraints of specific policy frameworks, which vary across platforms. As a result, the same content may be labeled differently depending on the platform's moderation rules and community guidelines. This inconsistency presents a significant challenge in developing cross-platform or generalized abusive language detection systems.

Ross *et al.* [18] explored hate speech in the context of the European refugee crisis by building a German-language

TABLE II: Comparison of Abusive Language and Related Policy Clauses Across Four Major Online Platforms (Facebook, X, Google and Amazon)

Policy category	Facebook	X (formerly Twitter)	Google	Amazon
Violence	Mentioned	Mentioned	Mentioned	Partially Mentioned
Protecting the crimes	Mentioned	Mentioned	Mentioned	Broadly mentioned under illegal
Illegal goods	Mentioned	Mentioned	Mentioned	Broadly mentioned under illegal
Dangerous people	Mentioned	Mentioned	Maps, Gmail, Meet	Broadly mentioned under illegal
Self-harm	Mentioned	Mentioned	Mentioned	Broadly mentioned under illegal
Sexual Abuse	Mentioned	Mentioned	Not Mentioned	Not mentioned
Animal abuse	Mentioned	Mentioned (sensitive media policy)	Earth, Drive, Meet	Broadly mentioned under illegal
Graphic content	Mentioned	Mentioned	Mentioned	Mentioned
Hate speech	Mentioned	Mentioned (Hateful Conduct)	Mentioned	Partially Mentioned
Pornography	Mentioned	Mentioned	Maps, Gmail, Meet	Under absence
Child abuse	Mentioned	Mentioned	Broadly mentioned	Not Mentioned

corpus. Waseem [19] used a dataset collected via CrowdFlower and analyzed annotations by amateur labelers to study gendered and racial abuse on X (also called Twitter) platform. Van Hee *et al.* [20] attempted to classify cyberbullying in social media, while Wang [21] focused on the detection of racist tweets. Nobata *et al.* [22] developed a binary classification system to detect abusive language in Yahoo! comments. Hosseini *et al.* [23] demonstrated how minor textual modifications could bypass toxic content detectors, highlighting the vulnerabilities of existing systems. Papegnies *et al.* [24] investigated insulter message detection and the role of conversational network modeling in improving context understanding. In a recent work, Khan *et al.* have shown that the Bi-LSTM with attention model, utilizing custom Word2Vec embedding provides better performance in Urdu text [25]. Vidgen *et al.* [26] provided a comprehensive overview of the challenges in abusive content detection, including limitations in current datasets, such as systematic biases against particular groups or forms of abuse. Waqas *et al.* [27] conducted a scientometric analysis of Internet hate speech research to map its development over time. Chetty *et al.* [28] examined the spread of hostile, inflammatory, and extremist content through digital platforms and online social networks. Fortuna and Nunes [29] conducted an extensive literature review covering the evolution of hate speech detection methods, dataset availability, and ethical concerns. Bensalem *et al.* have provided a survey of Arbic language datasets for identification of toxic content [30]. On similar footings, a dataset for Persian language has been developed recently for offensive content identification [31]. LSTM based deep sequential model has also been proposed in context of Urdu language [32].

Despite the valuable insights offered by aforementioned studies, most approaches are tailored to specific use cases or datasets and fail to provide a scalable, cross-platform solution. The lack of standardized definitions and labeling practices across platforms further complicates the development of generalizable models. Further, the abusive language detection models and policy frameworks are treated separately. Therefore, there is a clear need for a unified framework that incorporates both policy analysis and robust deep learning techniques to detect abusive language in a consistent and adaptable manner.

III. Policies of Online Platforms and Automatic Detection of Abusive Language

Each online social media platform defines and categorizes abusive content differently, reflecting its unique community guidelines, values and operational focus. In this section, we examine the popular platforms Facebook, X, Google and Amazon in terms of their respective definitions of abusive content and the policies they have in place to prevent it. The terms of service and privacy policies of online platforms outline the rules and guidelines that users must follow when accessing the services provided. These policies function as a contractual agreement between the platform and its users, establishing a framework to ensure user safety, content integrity, and legal compliance. Violations of these policies may result in consequences such as content removal, account suspension, or even legal action, depending on the severity of the offense [33], [34].

The comparison of content moderation policies across Facebook, X, Google, and Amazon is presented in Table II. Eleven categories of abusive language—namely hate speech, sexual harassment, graphic violence, glorification of crime, child abuse, and animal cruelty, among others—have been identified. The presence and treatment of these categories are examined in each platform's policies. Google policies (covering services like Gmail, Maps, and Meet) explicitly prohibit hate speech, glorification of crime, self-harm, illegal goods, and child abuse. However, certain categories, such as sexual harassment, may not be uniformly recognized or enforced across all Google services. Similarly, while Facebook and Instagram (both owned by Meta) share overlapping community standards, their interpretations of abusive behavior are not entirely identical. Further, Facebook emphasizes "harmful content" with a particular focus on real-world violence and often ties enforcement actions to geopolitical contexts. In contrast, X adopts a broader "hateful conduct" policy that incorporates greater flexibility by allowing exceptions in contexts such as satire or newsworthy speech [22], [26], [34].

As shown in Table II, each platform defines and addresses abusive language in its own way, complicating the development of universal automated detection systems. Given the immense volume of user-generated content, manual moderation is inadequate. As a result, platforms increasingly rely on

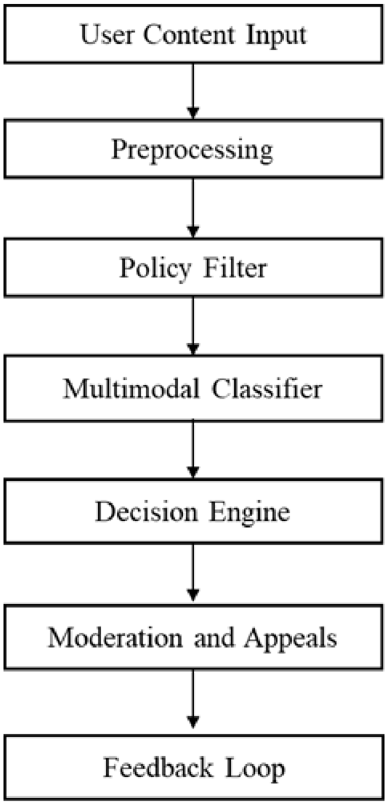

Fig. 1: Policy-aware automated abusive language detection framework.

automated tools powered by machine learning algorithms to detect abusive language. However, the absence of standardized definitions across platforms necessitates that these tools be highly adaptable, context-sensitive, and capable of aligning with the unique moderation frameworks of each platform.

In the next section, an integrated framework combining policies of social platform with machine learning is proposed. This framework can be applied to any platform generically.

IV. Policy-Aware Abusive Language Detection Framework

The conceptual view of policy-aware abusive language detection framework is proposed in Fig. 1. The process starts with the **User Content Input**, which includes text, images or audio submitted on the social-media platform. The **Reprocessing** module then prepares raw data from the input through normalization and feature extraction tailored for multi-modal inputs. Next, the **Policy Filter** applies platform-specific rules and guidelines, ensuring that detection aligns with standards of the underlying online platform. Thereafter, the **Multimodal Classifier** analyzes the processed content to identify abusive behavior by leveraging advanced machine learning models and **Decision Engine** interprets classifier outputs to determine whether content should be flagged, blocked or allowed. Finally, flagged content proceeds to **Moderation & Appeals**, combining automated and human review, to handle disputes and maintain fairness. The critical last step **Feedback Loop** enables continuous model and policy updates based on moderation outcomes and user input, fostering adaptability to evolving abusive behaviors and platform standards.

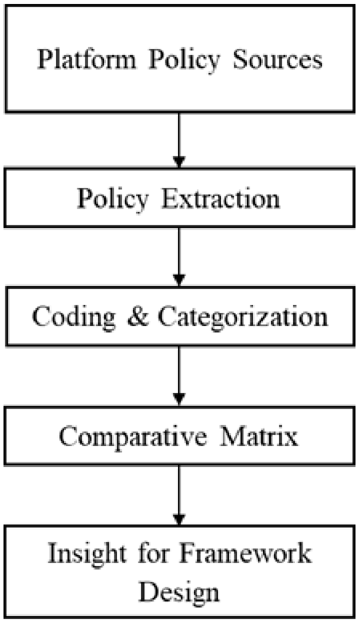

Fig. 2: Flowchart of policy analysis process.

An important step in the proposed framework is to design a mechanism to represent the policy of a platform which is outlined in the following subsection.

A. *Policy Analysis*

To systematically analyze abusive language policies of online platforms, a structured approach is usually adopted as shown in Fig. 2.The process starts with policy data collection. First, the official policy documents are gathered from various sources like platforms and communities. Next in policy categorization and comparison step, the part in the policy documents related to the abusive language are extracted manually and stored in a repository. These policy elements are categorized along key dimensions like definitions of abusive language, types of prohibited content (hate speech, harassment, threats), enforcement mechanisms (warnings, suspensions, removals) and options for appeals and user redressal. Discrepancies (if any) are removed [19]. Such coded policies are then compared to identify overlapping definitions, divergent enforcement strategies and notable gaps across platforms [20]. In the last step of validation, the coded data and findings are cross-verified with secondary sources including academic literature on content moderation frameworks and recent industry reports [18], [21]. This three step provide a structured evaluation of how different platforms conceptualize, regulate and enforce policies related to abusive language.

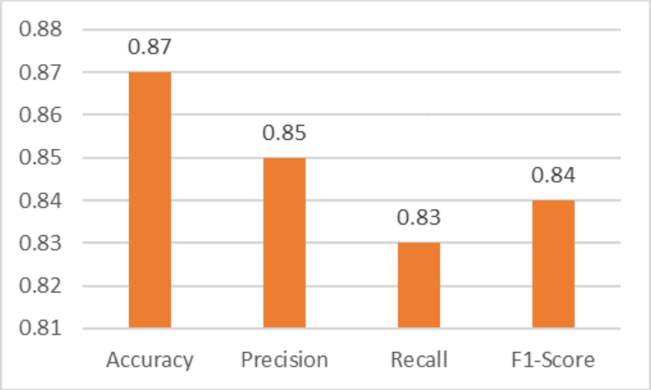

Fig. 3: Performance of abusive language detection system prototype on the HateEval dataset.

V. Experimental Validation

In this section, we present the details of a prototype of policy-aware abusive language detection system following the steps of Fig. 1. The prototype is validated on a publicly available dataset. The prototype architecture consists of the following:

- **Data Input:** Social media posts were collected from X using publicly available datasets (HateEval 2019) [2].
- **Preprocessing:** Includes text normalization, tokenization and translation of emojis into text descriptors to capture emotional and contextual nuances.
- **Abuse Detection Module:** Devlin et al. [35] proposed fine-tuned BERT-based and trained to detect offensive language in social media posts and the same was utilized.
- **Policy Filtering Layer** The classifier's output was filtered according to unified policy categories derived from our comparative analysis of platform content moderation guidelines.
- **Output:** Posts flagged as abusive are categorized and logged for further review.

The 80% of the data was used for training and remaining 20% for testing. The performance was evaluated by computing Accuracy, Precision, Recall, and F1-Score. The results are contained in Fig. 3.

The prototype system demonstrates strong performance, achieving an overall accuracy of 0.87, indicating effective classification of tweets as abusive or non-abusive. A precision of 0.85 suggests that the majority of flagged posts are genuinely abusive, thereby minimizing false positives. The recall of 0.83 reflects the model's capability to detect a substantial portion of abusive content, though some instances may remain undetected. The balanced F1-score of 0.84 highlights an effective trade-off between precision and recall, underscoring the model's robustness. These results validate the effectiveness of integrating a BERT-based classifier with a policy-aware filtering layer. Furthermore, the system's performance is consistent with recent benchmarks reported in the literature, reinforcing its suitability for real-world deployment. Continued refinement—through expanded datasets, inclusion of multimodal inputs, and dynamic policy updates—could further enhance detection accuracy and contextual understanding.

VI. Challenges in Automatic Detection of Abusive Language

While numerous Natural Language Processing (NLP) techniques have been applied to abusive content detection, much of the existing research tends to address isolated facets of the problem. A significant gap remains between academic advancements and the complex, real-world challenges faced by online platforms. Bridging this gap requires holistic approaches that not only detect abusive language but also incorporate contextual, cultural, and linguistic nuances. In a comprehensive review, Vidgen *et al.* [36] analyzed existing datasets and detection methods, highlighting several key sub-problems that collectively define the broader task of automated abusive language detection. These sub-problems are discussed in detail in the following subsections.

A. Identification of Abusive Language and Hate Speech

Hate speech detection remains one of the most critical and persistent challenges in the field of abusive language analysis. Davidson *et al*. [**?**] manually annotated a dataset of 24,000 tweets, categorizing them into hate speech, offensive language, and neutral content. Building on this, Basile *et al.* [37] introduced HateEval, a multilingual benchmark designed to support hate speech detection across different languages and cultural contexts. Facebook contributed to this research area by releasing a meme dataset containing 10,000 labeled samples for classifying images as hateful or non-hateful [38]. Founta *et al.* [39], [40] compiled a large-scale Twitter dataset comprising over 100,000 tweets, which were classified into categories such as hate speech, abusive content, and spam. Wiegand *et al.* [41] proposed a hierarchical modeling approach, breaking down hate speech detection into more granular subcategories to improve classification accuracy. Similarly, Glavaš *et al.* [42] investigated hate speech and aggression across domains using cross-lingual embeddings, enabling better generalization across languages. Ranasinghe *et al.* [43] applied multilingual deep learning models to enhance performance in cross-lingual scenarios, while Founta et al. [40] employed neural network architectures to develop robust classification systems. Waseem *et al.* [44] further contributed by distinguishing between individual-targeted and group-targeted abusive language—an important step toward improving contextual understanding in automated systems.

B. Aggression and Offensive Language Detection

The detection of offensive and aggressive language has been extensively explored across a variety of languages and online platforms. Several benchmark competitions, including TRAC [45], OffensEval [46], GermEval [41] and HASOC [47], have contributed significantly to this area by releasing multilingual datasets in languages such as English, Hindi,

[2] http://bit.ly/3TReG5T

Bengali, Arabic, German, Danish, Greek, and Turkish. Among the various classification methods employed, deep learning approaches, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have consistently demonstrated superior performance in identifying offensive content across these diverse datasets.

C. Poisonous Comment Detection

The detection of toxic or harmful comments has received increasing attention in response to the growing prevalence of hostile interactions on social media platforms. The issue gained widespread recognition following a Kaggle competition based on the Jigsaw dataset. In this context, Juuti *et al.* [48] introduced a detection approach that integrates transformer-based architectures with data augmentation techniques to improve classification performance. A key challenge addressed in their work is the identification of subtle forms of toxicity that are often embedded within otherwise neutral language, an area where traditional classifiers frequently struggle.

D. Identification of Malicious Accounts

The detection of malicious accounts consisting of bots, fake profiles, spammers and phishing entities is a critical component in maintaining the integrity of online platforms. These accounts are frequently exploited to disseminate abusive content, manipulate public discourse or engage in fraudulent behavior. Studies such as [49], [50] have investigated detection strategies based on behavioral patterns, network topology and account-level metadata. The convergence of malicious account detection and abusive language analysis underscores the importance of developing integrated frameworks that simultaneously evaluate both content and user-level signals for more comprehensive moderation.

To tackle the aforementioned challenges, neural network-based classification techniques have been widely adopted. Traditional models such as Support Vector Machines (SVMs), along with deep learning approaches like CNNs and RNNs have shown effectiveness in early abusive language detection tasks. More recently, transformer-based architectures—such as BERT, RoBERTa, and ALBERT [51]—have emerged as state-of-the-art solutions due to their advanced contextual language modeling capabilities. In multilingual settings, models like mBERT and XLM-RoBERTa [35] have outperformed earlier methods by offering robust cross-lingual generalization. Furthermore, in multimodal contexts involving both textual and visual elements (e.g., memes), visual-linguistic models such as ViLBERT [35], VLP [52], and UNITER [53] have demonstrated strong performance in detecting abusive content. These advanced methodologies provide the foundation for developing robust, cross-platform detection frameworks that can effectively address the complex and multifaceted nature of abusive language in online environments.

VII. Challenges in Policy-Aware Abusive Language Detection

While we have discussed the advantages of policy-aware generic abusive language detection, there are some challenges there that require to be addressed before this framework can be effectively adopted in online social-media platforms.

A. Contextual Limitations of Generalized Detection

Abusive language is inherently performative and highly context-dependent [54]. For instance, reclaimed slurs may be considered non-abusive within in-group contexts but perceived as harmful in other settings. However, platform policies often fail to account for such nuances, resulting in instances of over-enforcement.

B. Platform Biases

Most online platforms operate as commercial entities driven by financial incentives and, at times, influenced by political considerations. Consequently, user safety and fairness may be subordinated to corporate interests. This dynamic can lead to selective or inconsistent enforcement of content moderation policies, often shaped by geopolitical or economic factors. For example, while Facebook permanently suspended former U.S. President Donald Trump's account following the incitement of violence in Washington, D.C., comparable actions in less geopolitically prominent regions, such as Afghanistan or India, have not always elicited similar responses. Such inconsistencies reveal a potentially biased approach to moderation, undermining user trust and raising concerns about the global fairness and applicability of abuse policies.

C. Misalignment Between Challenges and Solutions

A significant disconnect exists between the complex challenges encountered by online platforms and the research solutions developed within the academic community. While platforms grapple with multifaceted and dynamic forms of abuse, academic research often addresses narrowly scoped problems using limited or constrained datasets. This misalignment hinders the practical applicability of research findings in real-world contexts. To bridge this gap, greater transparency and clearer communication from platform providers are crucial, enabling researchers to develop solutions that are both contextually relevant and practically impactful.

D. Lack of Harmonized Policies Across Platforms

While social platforms typically offer similar core functionalities, such as content sharing and messaging, their abuse detection policies differ significantly in both structure and terminology. The lack of standardized definitions and thresholds for abusive language presents a major obstacle to developing universal detection systems. Greater alignment through harmonized policies and a shared framework for defining abusive behavior would facilitate the creation of more comprehensive, interoperable, and transferable detection solutions across platforms.

E. Multimodal Content Complexity

Contemporary online content is increasingly multimodal integrating text, images, audio, video, emojis, and other formats. This diversity presents significant challenges for abuse detection, as harmful content may be conveyed through image captions, vocal tone, visual memes, or the contextual use of emojis. As a result, effective detection systems must be equipped to process and interpret multiple modalities to accurately identify and assess abusive behavior.

F. Post-Publication Filtering Instead of Pre-Publication Moderation

Abusive content is typically identified only after it has been published, by which point it may have already caused harm. Transitioning toward pre-publication filtering where content is screened prior to being made publicly visible, offers the potential to curb the immediate dissemination of harmful material. Although this approach introduces challenges related to processing latency and freedom of expression, the implementation of an intermediate moderation layer could substantially strengthen proactive abuse prevention mechanisms.

G. Lack of Generalized Datasets

Abusive language appears in diverse forms and is highly context-dependent, posing significant challenges to the creation and availability of generalized datasets. Existing datasets are often constrained to particular platforms, languages, or categories of abuse, limiting their effectiveness for training broadly applicable models. Additionally, the sensitive nature of abusive content, coupled with concerns around user privacy, further restricts access to high-quality, diverse datasets essential for comprehensive research.

While the existing research on abusibe language detection has yielded valuable contributions in terms of technical methodologies and dataset development [22], [29], [34], a critical gap persists between these advancements and the complex, context-sensitive nature of abuse as reflected in real-world platform policies. Many studies remain limited by narrowly defined abuse categories and constrained datasets, resulting in models that often lack generalizability across platforms, languages and cultural settings [43]. This exposes a fundamental tension between the pursuit of scalable, automated solutions and the nuanced reality that abusive language is embedded within social, cultural and contextual frameworks [24], [44]. This challenge is further compounded by the divergence in platform policies, as identified in our comparative analysis presented in section III. The lack of standardized definitions and enforcement mechanisms complicates the portability of detection models as well as raises ethical concerns regarding fairness, consistency and bias in automated moderation [26], [27]. For instance, a term deemed abusive in one cultural context may be reclaimed or non-abusive in another, illustrating the limitations of applying a universal detection standard [44]. Addressing these issues requires a paradigm shift toward context-aware models that integrate sociolinguistic and cultural variables. Drawing insights from interdisciplinary domains such as critical discourse analysis and sociolinguistics [24], future research should focus on developing adaptive frameworks that are customizable to reflect platform-specific policies and community norms, rather than imposing rigid, one-size-fits-all thresholds [26]. This aligns with contemporary critiques that highlight the performative and contingent nature of linguistic harm [24]. Further, incorporating user feedback and leveraging community moderation mechanisms presents a promising direction for creating dynamic systems that evolve over time. Such participatory approaches can enhance the transparency, fairness, and accountability of automated moderation tools [26], [42]. By embracing this complexity and moving beyond purely technical paradigms, the field can foster the development of more ethical, effective and socially responsive abusive language detection systems.

VIII. Conclusion

In this study, we have conducted a comprehensive investigation into the abusive language detection policies of several major online platforms, aiming to understand both their areas of convergence and divergence. The analysis reveals significant variability in how platforms define, categorize and enforce policies related to abusive language. A prototype of policy-aware automatic abusive language detection system is also implemented using BERT and its performance has been evaluated. An extensive review of the literature highlighting challenges in the area of automatic abusive language detection, specifically in relation to policy-aware abusive language detection is carried out. One of the most intriguing research problems in this ever-expanding field appears to be the design and implementation of a context-aware adaptive and dynamic abusive language detection system.

References

[1] A. Acharya, J. Manweiler, S. Sharma, and N. Banerjee, "Presence based open contact center leveraging social networks," in *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*, 2013, pp. 990–1003.

[2] M. Fazil, A. K. Sah, and M. Abulaish, "Deepsbd: A deep neural network model with attention mechanism for socialbot detection," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4211–4223, 2021.

[3] Federal Bureau of Investigation, "2015 hate crime statistics," https://ucr.fbi.gov/hate-crime/, 2015, accessed: Jul. 19, 2025.

[4] E. W. Pamungkas, V. Basile, and V. Patti, "Towards multidomain and multilingual abusive language detection: a survey," *Personal and Ubiquitous Computing*, vol. 27, no. 1, pp. 17–43, 2023.

[5] M. Abulaish, A. Saraswat, and M. Fazil, "A multi-task learning framework using graph attention network for user stance and rumor veracity prediction," in *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Kusadasi, Turkey*. ACM, November 6-9, 2023, pp. 1–5.

[6] A. Haque and M. Abulaish, "An emotion-enriched and psycholinguistics features-based approach for rumor detection on online social media," in *Proceedings of the 11th International Workshop on SocialNLP, Co-located with IJCNLP-AACL, Bali*, Nov 1-4, 2023, pp. 28–37.

[7] ——, "A graph-based approach leveraging posts and reactions for detecting rumors on online social media," in *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation (PACLIC), Manila*, Oct 20-22, 2022, pp. 1–12.

[8] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "Hatebert: Retraining bert for abusive language detection in english," in *arXiv preprint arXiv:2010.12472*, 2020.

[9] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL Student Research Workshop*, 2016, pp. 88–93.

[10] C. Park, S. Kim, K. Park, and K. Park, "K-haters: A hate speech detection corpus in korean with target-specific ratings," *arXiv preprint arXiv:2310.15439*, 2023.

[11] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760.

[12] E. W. Pamungkas and V. Patti, "Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon," in *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop*, 2019, pp. 363–370.

[13] B. Alrashidi, A. Jamal, I. Khan, and A. Alkhathlan, "A review on abusive content automatic detection: approaches, challenges and opportunities," *PeerJ Computer Science*, vol. 8, p. e1142, 2022.

[14] A. Koufakou, E. W. Pamungkas, V. Basile, and V. Patti, "Hurtbert: Incorporating lexical features with bert for the detection of abusive language," in *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Association for Computational Linguistics, 2020, pp. 34–43.

[15] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "Hatebert: Retraining bert for abusive language detection in english," *arXiv preprint arXiv:2010.12472*, 2020.

[16] N. Baratalipour, C. Y. Suen, and O. Ormandjieva, "Abusive language detection using bert pre-trained embedding," in *International Conference on Pattern Recognition and Artificial Intelligence*. Springer, 2020, pp. 695–701.

[17] J. A. Diaz-Garcia and J. P. Carvalho, "A literature review of textual cyber abuse detection using cutting-edge natural language processing techniques: Language models and large language models," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 15, no. 3, p. e70029, 2025.

[18] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "Measuring the reliability of hate speech annotations: The case of the european refugee crisis," *arXiv preprint arXiv:1701.08118*, 2017.

[19] Z. Talat, "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter," in *Proc. 1st Workshop on NLP and Computational Social Science*, Nov. 2016, pp. 138–142.

[20] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste, "Detection and fine-grained classification of cyberbullying events," in *Proc. Int. Conf. Recent Advances in Natural Language Processing*, Sep. 2015, pp. 672–680.

[21] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2013, pp. 1621–1622.

[22] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, pp. 145–153.

[23] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Deceiving google's perspective api built for detecting toxic comments," *arXiv preprint arXiv:1702.08138*, 2017.

[24] D. Jurgens, E. Chandrasekharan, and L. Hemphill, "A just and comprehensive strategy for using nlp to address online abuse," *arXiv preprint arXiv:1906.01738*, 2019.

[25] A. Khan, A. Ahmed, S. Jan, M. Bilal, and M. F. Zuhairi, "Abusive language detection in urdu text: Leveraging deep learning and attention mechanism," *IEEE Access*, vol. 12, pp. 37 418–37 431, 2024.

[26] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, and H. Margetts, "Challenges and frontiers in abusive content detection," in *Proc. 3rd Workshop on Abusive Language Online*. ACL, Aug. 2019.

[27] A. Waqas, J. Salminen, S. G. Jung, H. Almerekhi, and B. J. Jansen, "Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate," *PLoS One*, vol. 14, no. 9, p. e0222194, Sep. 2019.

[28] N. Chetty and S. Alathur, "Hate speech review in the context of online social networks," *Aggress. Violent Behav.*, vol. 40, pp. 108–118, 2018.

[29] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–30, 2018.

[30] I. Bensalem, P. Rosso, and H. Zitouni, "Toxic language detection: A systematic review of arabic datasets," *Expert Systems*, vol. 41, no. 8, p. e13551, 2024.

[31] E. Kebriaei, A. Homayouni, R. Faraji, A. Razavi, A. Shakery, H. Faili, and Y. Yaghoobzadeh, "Persian offensive language detection," *Machine Learning*, vol. 113, no. 7, pp. 4359–4379, 2024.

[32] A. Ullah, K. U. Khan, A. Khan, S. T. Bakhsh, A. U. Rahman, S. Akbar, and B. Saqia, "Threatening language detection from urdu data with deep sequential model," *Plos one*, vol. 19, no. 6, p. e0290915, 2024.

[33] E. Papegnies, V. Labatut, R. Dufour, and G. Linares, "Conversational networks for automatic online moderation," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 1, pp. 38–55, Mar. 2019.

[34] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. 5th Int. Workshop on Natural Language Processing for Social Media*, Apr. 2017, pp. 1–10.

[35] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguistics: Human Language Technology*, Jun. 2019, pp. 4171–4186.

[36] B. Vidgen and L. Derczynski, "Directions in abusive language training data, a systematic review: Garbage in, garbage out," *PLoS One*, vol. 15, no. 12, p. e0243300, Dec. 2020.

[37] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 11, no. 1, May 2017, pp. 512–515.

[38] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo *et al.*, "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in *Proc. 13th Int. Workshop on Semantic Evaluation*, Jun. 2019, pp. 54–63.

[39] M. Juuti, T. Gröndahl, A. Flanagan, and N. Asokan, "A little goes a long way: Improving toxic language classification despite data scarcity," *arXiv preprint arXiv:2009.12344*, 2020.

[40] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak *et al.*, "Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020)," *arXiv preprint arXiv:2006.07235*, 2020.

[41] T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification with cross-lingual embeddings," *arXiv preprint arXiv:2010.05324*, 2020.

[42] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini *et al.*, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 12, no. 1, Jun. 2018.

[43] G. Glavaš, M. Karan, and I. Vulić, "Xhate-999: Analyzing and detecting abusive language across domains and languages," 2020.

[44] Z. Waseem, T. Davidson, D. Warmsley, and I. Weber, "Understanding abuse: A typology of abusive language detection subtasks," *arXiv preprint arXiv:1705.09899*, 2017.

[45] S. Modha, P. Majumder, and T. Mandl, "Filtering aggression from the multilingual social media feed," in *Proc. 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, Aug. 2018, pp. 199–207.

[46] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 2611–2624.

[47] M. Wiegand, M. Siegel, and J. Ruppenhofer, "Overview of the germeval 2018 shared task on the identification of offensive language," in *Proc. GermEval Workshop 2018*, 2018.

[48] S. T. Aroyehun and A. Gelbukh, "Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling," in *Proc. 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, Aug. 2018, pp. 90–97.

[49] M. Fire, R. Goldschmidt, and Y. Elovici, "Online social networks: threats and solutions," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 4, pp. 2019–2036, 2014.

[50] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, "Dark of the social networks," *J. Netw. Comput. Appl.*, 2016.

[51] A. Almaatouq, E. Shmueli, M. Nouh, A. Alabdulkareem, V. K. Singh, M. Alsaleh *et al.*, "If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts," *Int. J. Inf. Secur.*, vol. 15, no. 5, pp. 475–491, 2016.

[52] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[53] L. H. Li, M. Yatskar, D. Yin, C. J. Hsieh, and K. W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[54] J. Butler, *Excitable Speech: A Politics of the Performative*. Routledge, 2021.