International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 02 | Feb 2026 www.irjet.net p-ISSN: 2395-0072

A REVIEW OF COMPARATIVE ANALYSIS OF CONTENT FILTERING USING LLMS Vs. TRADITIONAL NLP CLASSIFIERS

Sandeep Vishwakarma[1], Mrs. Arifa Khan[2]

[1]*Master of Technology, Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India*

Abstract - *Content filtering plays a critical role in ensuring safe and policy-compliant digital communication across social media platforms, online forums, and enterprise systems. Traditionally, content moderation has relied on rule-based systems and classical machine learning classifiers such as Naïve Bayes, Support Vector Machines, and Logistic Regression, which depend heavily on handcrafted features like n-grams and TF-IDF representations. The recent emergence of Large Language Models (LLMs), built on transformer architectures and pretrained on massive corpora, has introduced new paradigms for semantic understanding, contextual reasoning, and zero-shot or few-shot classification. This review presents a structured comparative analysis of LLM-based filtering approaches and traditional NLP classifiers. It synthesizes findings from prior studies to evaluate performance metrics, computational efficiency, generalization capability, interpretability, and robustness against adversarial content. The study further discusses trade-offs between scalability and accuracy, highlighting practical deployment considerations. The review identifies research gaps and outlines future directions for hybrid and resource-efficient moderation frameworks.*

Key Words: Content Filtering, Large Language Models, Traditional NLP Classifiers, Text Classification, Transformer Architecture, Content Moderation

1. INTRODUCTION

Content filtering has emerged as a foundational component of modern digital ecosystems, where vast volumes of user-generated text are produced continuously across platforms. From social media networks to enterprise communication systems, automated mechanisms are required to detect spam, misinformation, hate speech, and other forms of harmful or policy-violating content. Historically, filtering relied on rule-based heuristics and traditional machine learning classifiers; however, the rapid evolution of Natural Language Processing (NLP), particularly the advent of transformer-based Large Language Models (LLMs), has significantly altered the methodological landscape (Manning, Raghavan and Schütze, 2008; Vaswani et al., 2017). This section introduces the conceptual background, identifies the core research problem, and clarifies the scope and contributions of the present review.

1.1 Background and Motivation

1.1.1 Importance of Content Filtering in Modern Applications

Content filtering is central to maintaining platform integrity, user safety, and regulatory compliance. Social media platforms employ automated moderation to detect hate speech, harassment, extremist propaganda, and misinformation (Schmidt and Wiegand, 2017). Email systems utilize spam filtering to protect users from phishing and malicious campaigns, often relying on probabilistic classifiers such as Naïve Bayes (Metsis, Androutsopoulos and Paliouras, 2006). In addition, enterprise and educational platforms deploy filtering mechanisms to prevent the dissemination of inappropriate or harmful content.

The scale and velocity of online communication make manual moderation infeasible, necessitating automated systems capable of high precision and recall. Furthermore, the increasing sophistication of adversarial content—such as obfuscated hate speech or context-dependent misinformation—demands models that move beyond surface-level lexical patterns toward deeper semantic understanding (Zhang, Robinson and Tepper, 2018).

1.1.2 Evolution from Traditional NLP to Large Language Models

Early NLP-based filtering systems relied on handcrafted rules and keyword matching, which lacked adaptability and contextual awareness. The introduction of machine learning classifiers such as Support Vector Machines (SVM), Logistic Regression, and Naïve Bayes improved generalization by learning statistical patterns from labeled corpora (Joachims, 1998; Manning, Raghavan and Schütze, 2008). These approaches typically used feature representations such as Bag-of-Words (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF).

Subsequently, distributed word representations (e.g., Word2Vec, GloVe) enabled semantic encoding of text (Mikolov et al., 2013; Pennington, Socher and Manning, 2014). Deep learning architectures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), further enhanced contextual modeling (Kim, 2014). The transformer architecture introduced self-attention

mechanisms that captured long-range dependencies more effectively (Vaswani et al., 2017). Pretrained transformer models such as BERT demonstrated substantial improvements in text classification tasks through fine-tuning (Devlin et al., 2019). More recently, generative LLMs have shown zero-shot and few-shot classification capabilities, reshaping the paradigm of content moderation (Brown et al., 2020).

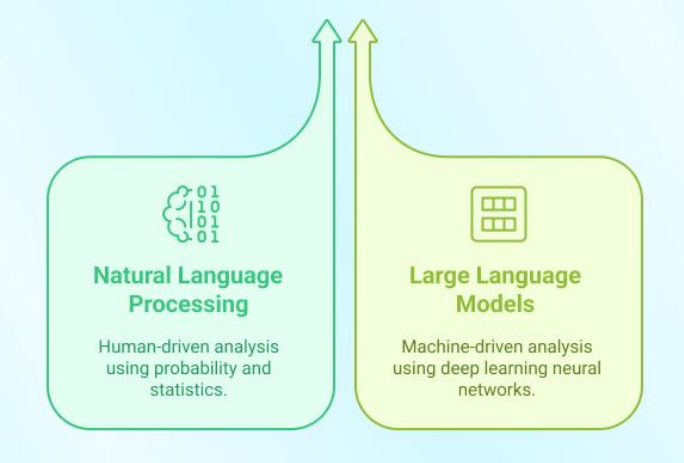

Figure-1: NLP Vs LLM

1.2 Scope and Contributions of This Paper

This review systematically examines and compares traditional NLP classifiers and LLM-based approaches in the domain of content filtering. It synthesizes prior research spanning rule-based systems, classical machine learning models, deep neural architectures, and transformer-based LLMs. The review focuses on supervised and zero-/few-shot classification paradigms applied to text moderation tasks. However, it does not extensively cover multimodal filtering (e.g., image or video moderation) or non-textual signal integration.

The primary contributions of this paper are threefold. First, it proposes a structured taxonomy of content filtering methods, categorizing them into rule-based, classical machine learning, deep learning, and large language model paradigms. Second, it provides a comparative synthesis of empirical findings across benchmark datasets, highlighting differences in accuracy, generalization, interpretability, and computational requirements. Third, it identifies research gaps related to cross-domain robustness, bias mitigation, and hybrid architectures, thereby outlining potential future research trajectories. By consolidating fragmented literature into a coherent analytical framework, this review aims to support both academic inquiry and practical system design.

2. FUNDAMENTAL CONCEPTS

Content filtering within Natural Language Processing (NLP) is grounded in classification theory, representation learning, and evaluation science. Understanding its conceptual foundations is essential for comparing traditional machine learning classifiers with modern Large Language Models (LLMs). This section defines the taxonomy of filtering approaches, outlines evaluation metrics, and explains the technical evolution from classical algorithms to transformer-based architectures.

2.1 Content Filtering in Natural Language Processing

Content filtering refers to automated methods for identifying, categorizing, or blocking textual data according to predefined criteria such as spam, hate speech, misinformation, or policy violations. It is fundamentally a text classification problem but may extend to anomaly detection or semantic similarity tasks depending on the application (Manning, Raghavan and Schütze, 2008).

2.1.1 Taxonomy of Content Filtering Approaches

Content filtering approaches can be broadly categorized into rule-based, supervised learning, and unsupervised techniques.

Rule-based systems rely on handcrafted lexicons, regular expressions, and manually defined heuristics. These systems are transparent and computationally efficient but lack adaptability and contextual awareness (Schmidt and Wiegand, 2017).

Supervised learning methods utilize labeled datasets to train classifiers that learn decision boundaries from feature representations. Algorithms such as Naïve Bayes and Support Vector Machines became standard due to their robustness and scalability (Joachims, 1998; Metsis, Androutsopoulos and Paliouras, 2006). Supervised models generally outperform rule-based approaches when sufficient labeled data are available.

Unsupervised techniques include clustering and anomaly detection methods that identify unusual or suspicious patterns without labeled training data. These approaches are particularly useful in detecting emerging or previously unseen harmful content but may lack precise categorical interpretation (Aggarwal and Zhai, 2012).

2.1.2 Metrics and Evaluation Criteria

The effectiveness of content filtering systems is assessed using standard classification metrics. Precision measures the proportion of correctly identified positive instances among all predicted positives, while recall measures the proportion of actual positives correctly identified. The F1-score, the harmonic mean of precision and recall, balances false positives and false negatives (Sokolova and Lapalme, 2009).

In content moderation contexts, false positives may unjustly censor benign content, whereas false negatives may allow harmful material to pass undetected. Therefore, performance evaluation must consider domain-specific trade-offs. In large-scale deployment, additional criteria such as latency, scalability, and robustness to adversarial manipulation are also critical (Zhang, Robinson and Tepper, 2018).

2.2 Traditional NLP Classifiers

Traditional NLP classifiers rely on statistical learning theory and structured feature engineering. Their effectiveness depends heavily on input representation and the reparability of classes in feature space.

2.2.1 Classical Machine Learning Algorithms

Naïve Bayes (NB) is a probabilistic classifier based on Bayes' theorem and conditional independence assumptions. It has been widely applied in spam detection due to its efficiency and strong baseline performance (Metsis, Androutsopoulos and Paliouras, 2006).

Support Vector Machines (SVM) aim to maximize the margin between classes in high-dimensional feature space, offering strong generalization performance in text categorization tasks (Joachims, 1998).

Decision Trees and Random Forests construct hierarchical decision rules; Random Forests improve robustness by aggregating multiple trees to reduce variance (Breiman, 2001).

Logistic Regression models class probabilities using a linear decision boundary and is particularly effective for linearly separable text data (Manning, Raghavan and Schütze, 2008).

These algorithms are computationally efficient and interpretable but often require extensive feature engineering.

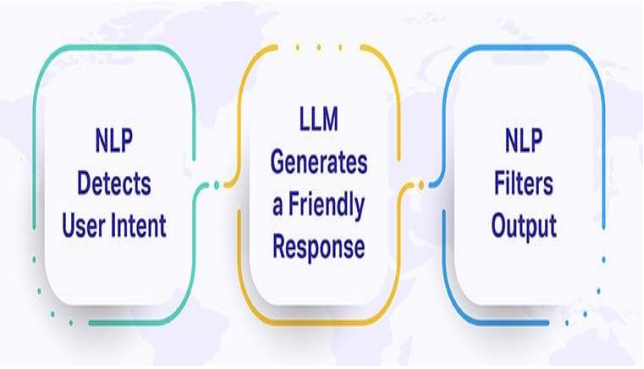

Figure-2: Classical Machine Learning

2.2.2 Feature Representations

Traditional classifiers depend on explicit textual representations. Bag-of-Words (BoW) encodes text as word frequency vectors without capturing word order (Harris, 1954). TF-IDF improves this representation by weighting terms according to their importance within a corpus (Salton and Buckley, 1988).

To incorporate semantic relationships, distributed word embeddings such as Word2Vec and GloVe were introduced. Word2Vec uses neural networks to learn vector representations capturing syntactic and semantic similarities (Mikolov et al., 2013), while GloVe integrates global co-occurrence statistics (Pennington, Socher and Manning, 2014). Although these embeddings improved contextual modeling, they remain static and do not adapt to sentence-level context.

2.3 Large Language Models (LLMs)

Large Language Models represent a paradigm shift in NLP by leveraging deep neural architectures trained on massive corpora. They reduce reliance on handcrafted features and enable contextualized understanding of language.

2.3.1 Evolution from RNN/LSTM to Transformers

Early deep learning models for NLP included Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which captured sequential dependencies but struggled with long-range context and parallelization (Hochreiter and Schmidhuber, 1997).

The introduction of the Transformer architecture, based on self-attention mechanisms, addressed these limitations by enabling parallel computation and improved modeling of global dependencies (Vaswani et al., 2017). Pretrained models such as BERT demonstrated that bidirectional contextual representations significantly enhance text classification performance (Devlin et al., 2019). Generative models like GPT further extended capabilities to few-shot and zero-shot learning scenarios (Brown et al., 2020). Successive architectures such as RoBERTa optimized pretraining strategies for improved downstream performance (Liu et al., 2019).

2.3.2 Capabilities Relevant to Content Filtering

LLMs provide several advantages for content filtering. First, they generate contextualized embeddings, meaning the representation of a word depends on its surrounding context, improving detection of subtle or implicit harmful content. Second, they support transfer learning, allowing models pretrained on large corpora to be fine-tuned on smaller domain-specific datasets (Devlin et al., 2019). Third, generative LLMs demonstrate zero-shot and few-shot

classification, enabling filtering in low-resource or rapidly evolving domains (Brown et al., 2020).

However, LLMs introduce challenges, including high computational cost, potential bias inherited from training data, and reduced interpretability compared to traditional linear models (Bommasani et al., 2021). These characteristics necessitate systematic comparative evaluation in content moderation contexts.

3. LITERATURE REVIEW

The evolution of content filtering methodologies reflects broader developments in Natural Language Processing (NLP). Rather than presenting studies chronologically, this review synthesizes literature thematically—contrasting methodological paradigms, empirical findings, and unresolved research gaps. The discussion progresses from rule-based and classical machine learning approaches to deep neural architectures and large-scale transformer models, culminating in comparative and critical insights.

3.1 Traditional Content Filtering Methods

3.1.1 Rule-Based and Early Machine Learning Approaches

Early content filtering systems relied on rule-based mechanisms such as regular expressions, blacklist lexicons, and handcrafted heuristics. These systems were widely adopted in spam detection and keyword-based moderation due to their transparency and low computational overhead (Manning, Raghavan and Schütze, 2008). However, they were brittle, easily circumvented through lexical obfuscation, and unable to capture contextual nuance (Schmidt and Wiegand, 2017).

The shift toward supervised machine learning introduced probabilistic and margin-based classifiers. Naïve Bayes became prominent in spam filtering due to its simplicity and strong empirical performance on high-dimensional sparse text data (Metsis, Androutsopoulos and Paliouras, 2006). Support Vector Machines (SVM) demonstrated improved generalization in abusive language detection and text categorization tasks (Joachims, 1998). These approaches relied heavily on feature engineering, particularly n-grams and TF-IDF representations.

Comparative studies indicate that while n-gram features perform well for surface-level lexical discrimination, they fail to capture semantic similarity and contextual relationships (Wang and Manning, 2012). The introduction of distributed embeddings such as Word2Vec improved semantic modeling but remained limited by static word representations (Mikolov et al., 2013). Traditional classifiers are computationally efficient and interpretable; however, they depend on labeled data and struggle with domain adaptation.

3.2 Deep Learning and Feature-Learning Approaches

3.2.1 Neural Architectures for Text Classification

Deep learning introduced automated feature learning, reducing reliance on manual engineering. Convolutional Neural Networks (CNNs) demonstrated strong performance in sentence-level classification by capturing local n-gram features through convolutional filters (Kim, 2014). Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, modeled sequential dependencies and improved context sensitivity (Hochreiter and Schmidhuber, 1997).

Empirical evaluations on hate speech and sentiment benchmarks showed CNN and LSTM models outperforming traditional machine learning baselines in F1-score and recall (Zhang, Zhao and LeCun, 2015). Hybrid architectures integrating CNN layers with attention mechanisms further enhanced contextual awareness by assigning dynamic importance to relevant tokens.

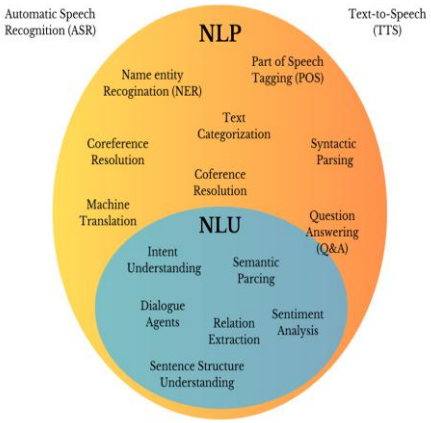

Figure-3: Neural Architectures

3.2.2 Comparative Performance and Limitations

While neural models improved classification accuracy, they introduced increased computational complexity and reduced interpretability. Studies comparing SVMs and CNNs observed consistent performance gains for neural models but at higher training cost and data requirements (Zhang, Zhao and LeCun, 2015). Moreover, deep models remained sensitive to adversarial rephrasing and domain shifts, highlighting limitations in robustness.

3.3 Transformer-Based Models and Large Language Models

3.3.1 Fine-Tuned Transformer Models

The introduction of the Transformer architecture marked a paradigm shift by enabling parallelized self-attention

mechanisms capable of modeling long-range dependencies (Vaswani et al., 2017). Pretrained transformer models such as BERT significantly improved text classification outcomes through bidirectional contextual embeddings and fine-tuning strategies (Devlin et al., 2019). RoBERTa optimized pretraining procedures, yielding further improvements across benchmark datasets (Liu et al., 2019).

In content moderation tasks, fine-tuned BERT models consistently outperformed CNN and SVM baselines in hate speech and toxic comment classification benchmarks, demonstrating superior recall and semantic discrimination (Mozafari, Farahbakhsh and Crespi, 2019).

3.3.2 Zero-Shot and Few-Shot LLM-Based Filtering

Generative Large Language Models (LLMs), including GPT-3 and GPT-4, introduced zero-shot and few-shot classification capabilities (Brown et al., 2020). These models leverage large-scale pretraining to perform classification tasks without task-specific fine-tuning. Literature indicates that zero-shot LLMs can achieve competitive accuracy compared to supervised baselines, particularly in low-resource or cross-domain scenarios (Bommasani et al., 2021).

However, empirical findings also reveal sensitivity to prompt formulation and increased computational costs. Unlike traditional classifiers, LLM-based systems may introduce latent biases inherited from pretraining corpora, affecting fairness in moderation outcomes.

3.4 Comparative Studies in Literature

Several studies directly compare traditional classifiers, deep neural networks, and transformer-based models for filtering tasks. Empirical results consistently demonstrate that transformer-based models outperform classical algorithms in accuracy and F1-score across benchmark datasets (Devlin et al., 2019; Mozafari, Farahbakhsh and Crespi, 2019). Quantitatively, improvements often range between 3–10% in macro-F1 compared to SVM or logistic regression baselines.

Qualitative analyses indicate that transformers better capture implicit hate speech and contextual offensiveness, whereas traditional models rely heavily on explicit lexical cues (Schmidt and Wiegand, 2017). Despite performance gains, literature identifies gaps in cross-domain generalizability. Models trained on one platform frequently underperform when deployed in another domain due to linguistic variation and distributional shift.

Bias and fairness concerns are also recurrent themes. Studies show that both traditional and transformer-based models may disproportionately flag content from certain demographic groups, reflecting dataset imbalance and representational bias (Bommasani et al., 2021). Comparative literature therefore emphasizes not only accuracy but also ethical and deployment considerations.

3.5 Challenges Identified in Literature

3.5.1 Data and Robustness Challenges

Data imbalance remains a critical issue, particularly in hate speech detection where harmful instances represent a minority class (Schmidt and Wiegand, 2017). Models often achieve high overall accuracy while underperforming on minority categories. Additionally, adversarial inputs—such as misspellings or coded language—reduce classifier robustness (Zhang, Robinson and Tepper, 2018).

3.5.2 Computational and Ethical Trade-Offs

Transformer-based models and LLMs demand significant computational resources for training and inference, raising scalability and energy consumption concerns (Bommasani et al., 2021). In contrast, traditional classifiers offer efficiency but may sacrifice contextual precision.

Ethical concerns include algorithmic bias, over-censorship, and lack of explainability. Interpretability is particularly challenging in LLM-based systems due to their scale and complexity. The literature increasingly emphasizes the need for transparent, fair, and accountable moderation systems.

4. COMPARATIVE ANALYSIS

A systematic comparative analysis of traditional NLP classifiers and Large Language Models (LLMs) for content filtering requires multidimensional evaluation. Performance must be assessed not only in terms of predictive accuracy but also computational feasibility, interpretability, robustness, and deployment constraints. This section synthesizes comparative evidence from existing literature across these dimensions.

4.1 Evaluation Criteria

4.1.1 Classification Performance Metrics

The primary quantitative indicators for evaluating content filtering systems include accuracy, precision, recall, and F1-score. Accuracy measures overall correctness, but in imbalanced moderation datasets—where harmful content forms a minority—precision and recall provide more meaningful insight (Sokolova and Lapalme, 2009). Precision quantifies the proportion of correctly predicted positive instances, while recall reflects the system's ability to detect all relevant harmful instances. The F1-score balances these two measures and is widely used in hate speech and spam detection benchmarks (Schmidt and Wiegand, 2017).

Beyond classification metrics, operational systems must consider latency and scalability. Latency refers to inference time per instance, which is critical for real-time moderation

pipelines. Traditional classifiers such as Logistic Regression and SVM generally offer lower inference latency compared to transformer-based models (Manning, Raghavan and Schütze, 2008). Scalability evaluates the system's ability to handle large volumes of streaming data, where model size and computational throughput become decisive factors.

4.1.2 Interpretability and Explain ability

Interpretability remains a significant comparative dimension. Linear models and decision trees provide transparent decision boundaries and feature importance scores, facilitating explain ability (Breiman, 2001). In contrast, deep neural networks and LLMs operate as high-dimensional non-linear systems, making interpretability more complex. Although attention visualization and post-hoc explanation techniques exist, they do not fully resolve transparency concerns (Bommasani et al., 2021). For regulatory compliance and fairness auditing, interpretability may outweigh marginal gains in predictive performance.

4.2 Benchmark Datasets and Tasks

4.2.1 Common Datasets in Content Filtering

Comparative studies frequently employ benchmark datasets such as hate speech corpora, toxic comment datasets, and spam email collections. For example, the Davidson et al. hate speech dataset and similar annotated corpora are commonly used for abusive language detection (Schmidt and Wiegand, 2017). Spam filtering research often relies on publicly available corpora containing labeled legitimate and phishing emails (Metsis, Androutsopoulos and Paliouras, 2006).

Transformer-based studies typically evaluate models on multi-domain benchmarks to assess generalization (Devlin et al., 2019). However, dataset heterogeneity in annotation schemes and class definitions complicates direct comparison across studies.

4.2.2 Task Formulations

Content filtering tasks may be framed as binary classification (harmful vs. non-harmful), multi-label classification (e.g., toxicity, hate, harassment simultaneously), or hierarchical classification, where content is categorized into nested taxonomies. Traditional classifiers perform effectively in binary settings with well-defined features, whereas LLMs demonstrate advantages in multi-label and context-dependent tasks due to their contextual embeddings (Liu et al., 2019).

4.3 Performance Comparison

4.3.1 Traditional Classifiers vs. Transformer-Based Models

Empirical comparisons consistently indicate that transformer-based models such as BERT outperform traditional classifiers on benchmark datasets in terms of macro-F1 and recall (Devlin et al., 2019; Mozafari, Farahbakhsh and Crespi, 2019). Improvements are particularly evident in detecting implicit or context-dependent harmful language. Traditional SVM and Logistic Regression models remain competitive in lexically explicit tasks, especially when computational efficiency is prioritized (Wang and Manning, 2012).

While classical algorithms often serve as strong baselines, they rely heavily on surface-level lexical features. Transformer-based models, through contextual embeddings, demonstrate superior semantic discrimination and reduced dependence on manual feature engineering.

4.3.2 Zero-Shot/Few-Shot LLMs vs. Supervised Traditional Models

Large generative models such as GPT-3 exhibit strong zero-shot and few-shot classification capabilities (Brown et al., 2020). Comparative evidence suggests that zero-shot LLM performance can approach supervised traditional classifiers in certain domains, especially when labeled data are scarce (Bommasani et al., 2021). However, fully fine-tuned transformer models generally surpass zero-shot approaches in stable high-resource environments.

Thus, the comparative advantage depends on data availability: traditional supervised models remain viable in resource-constrained computational environments, whereas LLMs provide flexibility in low-labeled or rapidly evolving contexts.

4.4 Computational and Deployment Considerations

4.4.1 Model Size and Resource Requirements

Traditional classifiers typically require minimal memory and processing resources, enabling deployment on edge devices or real-time moderation systems (Manning, Raghavan and Schütze, 2008). In contrast, transformer-based LLMs contain millions or billions of parameters, leading to high memory usage and increased inference cost (Bommasani et al., 2021).

Model compression techniques, including knowledge distillation and quantization, attempt to mitigate these constraints; however, trade-offs between efficiency and performance remain evident.

4.4.2 Real-Time Filtering Feasibility

Real-time filtering systems demand low-latency inference. Classical machine learning models can process high-throughput streams with minimal delay. Transformer-based systems, unless optimized or deployed with hardware acceleration, may introduce latency unsuitable for high-frequency moderation scenarios. Therefore, deployment feasibility often depends on infrastructure capacity and cost tolerance.

4.5 Robustness and Generalization

4.5.1 Sensitivity to Noise and Domain Shift

Traditional classifiers relying on lexical features are sensitive to spelling variations, slang, and adversarial manipulation (Zhang, Robinson and Tepper, 2018). Transformer-based models demonstrate improved robustness to paraphrasing and contextual variation due to semantic modeling capabilities (Devlin et al., 2019). Nevertheless, both approaches exhibit performance degradation under domain shift, particularly when applied to unseen platforms or linguistic communities.

4.5.2 Handling Unseen or Emerging Content

LLMs possess a comparative advantage in handling unseen categories through transfer learning and zero-shot generalization (Brown et al., 2020). Their large-scale pretraining allows broader semantic coverage, making them more adaptable to emerging harmful patterns. However, bias inherited from training corpora may compromise fairness and reliability (Bommasani et al., 2021).

Overall, robustness and generalization remain active research areas, with hybrid architectures increasingly proposed to balance efficiency, adaptability, and fairness.

5. DISCUSSION

The comparative analysis of traditional NLP classifiers and Large Language Models (LLMs) for content filtering reveals both methodological advancements and persistent challenges. While transformer-based systems demonstrate measurable improvements in contextual understanding and generalization, classical approaches retain relevance in resource-constrained and high-throughput environments. This section synthesizes the principal findings, outlines practical implications, and critically evaluates the limitations of existing literature.

5.1 Key Findings

5.1.1 Performance and Generalization Insights

The literature consistently indicates that transformer-based models, particularly BERT and its variants, outperform traditional classifiers such as SVM and Logistic Regression on benchmark moderation datasets in terms of macro-F1 and recall (Devlin et al., 2019; Mozafari, Farahbakhsh and Crespi, 2019). These gains are largely attributed to contextualized embeddings and bidirectional attention mechanisms, which enable deeper semantic interpretation compared to n-gram-based features (Vaswani et al., 2017).

However, classical machine learning models remain competitive when tasks rely on explicit lexical cues and when computational efficiency is a priority (Wang and Manning, 2012). Zero-shot and few-shot LLMs introduce flexibility in low-resource settings, though their performance can vary depending on prompt design and domain alignment (Brown et al., 2020). Thus, the key finding is not a universal superiority of LLMs, but a context-dependent performance advantage shaped by task complexity, data availability, and deployment constraints.

5.2 Practical Implications

5.2.1 Deployment Strategies and System Design

For real-world moderation systems—such as social media filtering, enterprise email spam detection, and online community management—the choice of model architecture must balance accuracy with scalability and latency. Traditional classifiers are well-suited for large-scale, real-time filtering due to their low inference cost and interpretability (Manning, Raghavan and Schütze, 2008). They are particularly effective in structured environments where harmful patterns are lexically explicit and stable over time.

Conversely, transformer-based and LLM-driven systems are advantageous in dynamic environments characterized by evolving slang, implicit hate speech, or multilingual contexts. Their transfer learning capabilities allow adaptation to emerging content types without extensive feature engineering (Devlin et al., 2019). Hybrid frameworks—combining lightweight traditional models for initial screening with LLM-based secondary review—have been proposed to optimize efficiency and semantic depth.

From a governance perspective, explainability and fairness auditing must be integrated into system design, particularly in regulatory environments requiring transparency and accountability (Bommasani et al., 2021). Practitioners should therefore adopt evaluation pipelines that include bias analysis, robustness testing, and domain adaptation validation.

6. CONCLUSION

This review systematically compared traditional NLP classifiers and Large Language Models (LLMs) for content filtering, highlighting their methodological distinctions, empirical performance, and deployment trade-offs. Evidence from benchmark studies indicates that transformer-based

models, particularly fine-tuned architectures such as BERT and RoBERTa, consistently outperform classical algorithms in terms of contextual understanding, recall, and macro-F1 scores, especially in detecting implicit or nuanced harmful content. However, traditional classifiers—including Naïve Bayes, SVM, and Logistic Regression—remain computationally efficient, interpretable, and suitable for large-scale real-time filtering environments. Zero-shot and few-shot LLMs offer flexibility in low-resource or rapidly evolving domains but introduce variability in performance and higher inference costs. The comparative findings suggest that no single approach is universally optimal; rather, model selection should align with application-specific constraints, including data availability, latency tolerance, fairness requirements, and infrastructure capacity. Hybrid and resource-aware moderation architectures represent a promising direction for balancing semantic depth with operational feasibility in future content filtering systems.

7. LIMITATIONS

Despite comprehensive synthesis, this review is constrained by several limitations. First, it relies primarily on benchmark-based studies, which may not fully reflect real-world moderation complexity, multilingual diversity, or platform-specific linguistic dynamics. Second, variations in dataset annotation standards and evaluation protocols restrict direct cross-study comparability. Third, quantitative metrics such as accuracy and F1-score often overshadow operational considerations, including energy consumption, infrastructure cost, and environmental impact. Additionally, while bias and fairness concerns are acknowledged, existing literature provides limited standardized frameworks for systematic bias evaluation across models. Finally, the rapidly evolving nature of LLM architectures means that comparative findings may become outdated as newer models emerge. Continuous empirical reassessment is therefore essential to maintain relevance and methodological rigor.

REFERENCES

1. Aggarwal, C.C. and Zhai, C. (2012) Mining Text Data. New York: Springer.
2. Bommasani, R. et al. (2021) 'On the opportunities and risks of foundation models', arXiv preprint arXiv:2108.07258.
3. Breiman, L. (2001) 'Random forests', Machine Learning, 45(1), pp. 5–32.
4. Brown, T.B. et al. (2020) 'Language models are few-shot learners', Advances in Neural Information Processing Systems, 33, pp. 1877–1901.
5. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', Proceedings of NAACL-HLT, pp. 4171–4186.
6. Harris, Z.S. (1954) 'Distributional structure', Word, 10(2–3), pp. 146–162.
7. Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', Neural Computation, 9(8), pp. 1735–1780.
8. Joachims, T. (1998) 'Text categorization with support vector machines: Learning with many relevant features', Proceedings of ECML, pp. 137–142.
9. Kim, Y. (2014) 'Convolutional neural networks for sentence classification', Proceedings of EMNLP, pp. 1746–1751.
10. Liu, Y. et al. (2019) 'RoBERTa: A robustly optimized BERT pretraining approach', arXiv preprint arXiv:1907.11692.
11. Manning, C.D., Raghavan, P. and Schütze, H. (2008) Introduction to Information Retrieval. Cambridge: Cambridge University Press.
12. Metsis, V., Androutsopoulos, I. and Paliouras, G. (2006) 'Spam filtering with Naive Bayes – Which Naive Bayes?', CEAS Conference on Email and Anti-Spam.
13. Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) 'Efficient estimation of word representations in vector space', arXiv preprint arXiv:1301.3781.
14. Mozafari, M., Farahbakhsh, R. and Crespi, N. (2019) 'A BERT-based transfer learning approach for hate speech detection in online social media', Complex Networks & Their Applications, pp. 928–940.
15. Pennington, J., Socher, R. and Manning, C.D. (2014) 'GloVe: Global vectors for word representation', Proceedings of EMNLP, pp. 1532–1543.
16. Salton, G. and Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval', Information Processing & Management, 24(5), pp. 513–523.
17. Schmidt, A. and Wiegand, M. (2017) 'A survey on hate speech detection using natural language processing', Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp. 1–10.
18. Sokolova, M. and Lapalme, G. (2009) 'A systematic analysis of performance measures for classification tasks', Information Processing & Management, 45(4), pp. 427–437.

19. Vaswani, A. et al. (2017) 'Attention is all you need', Advances in Neural Information Processing Systems, 30.

20. Wang, S. and Manning, C.D. (2012) 'Baselines and bigrams: Simple, good sentiment and topic classification', Proceedings of ACL, pp. 90–94.

21. Zhang, X., Zhao, J. and LeCun, Y. (2015) 'Character-level convolutional networks for text classification', Advances in Neural Information Processing Systems, 28.

22. Zhang, Z., Robinson, D. and Tepper, J. (2018) 'Detecting hate speech on Twitter using a convolution-GRU based deep neural network', European Semantic Web Conference, pp. 745–760.

23. Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017) 'Enriching word vectors with subword information', Transactions of the Association for Computational Linguistics, 5, pp. 135–146.

24. Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017) 'Automated hate speech detection and the problem of offensive language', Proceedings of ICWSM, 11(1), pp. 512–515.

25. Dixon, L. et al. (2018) 'Measuring and mitigating unintended bias in text classification', Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 67–73.

26. Dos Santos, C.N. and Gatti, M. (2014) 'Deep convolutional neural networks for sentiment analysis of short texts', Proceedings of COLING, pp. 69–78.

27. ElSherief, M., Nilizadeh, S., Nguyen, D. and Belding, E. (2018) 'Peer to peer hate: Hate speech instigators and their targets', Proceedings of ICWSM, 12(1), pp. 52–61.

28. Fortuna, P. and Nunes, S. (2018) 'A survey on automatic detection of hate speech in text', ACM Computing Surveys, 51(4), pp. 1–30.

29. Gehman, S. et al. (2020) 'RealToxicityPrompts: Evaluating neural toxic degeneration in language models', Findings of EMNLP, pp. 3356–3369.

30. Howard, J. and Ruder, S. (2018) 'Universal language model fine-tuning for text classification', Proceedings of ACL, pp. 328–339.

31. Jurafsky, D. and Martin, J.H. (2023) Speech and Language Processing. 3rd edn (draft). Stanford University.

32. Kowsari, K. et al. (2019) 'Text classification algorithms: A survey', Information, 10(4), p. 150.

33. Kumar, R., Ojha, A.K., Malmasi, S. and Zampieri, M. (2018) 'Benchmarking aggression identification in social media', Proceedings of TRAC Workshop, pp. 1–11.

34. Lample, G. and Conneau, A. (2019) 'Cross-lingual language model pretraining', Advances in Neural Information Processing Systems, 32.

35. Li, J., Sun, A., Han, J. and Li, C. (2018) 'A survey on deep learning for named entity recognition', IEEE Transactions on Knowledge and Data Engineering, 34(1), pp. 50–70.

36. Malmasi, S. and Zampieri, M. (2018) 'Challenges in discriminating profanity from hate speech', Journal of Experimental & Theoretical Artificial Intelligence, 30(2), pp. 187–202.

37. Mathew, B. et al. (2019) 'Spread of hate speech in online social media', Proceedings of WebSci, pp. 173–182.

38. Pavlopoulos, J., Malakasiotis, P. and Androutsopoulos, I. (2017) 'Deep learning for user comment moderation', Proceedings of ACL, pp. 25–35.

39. Ruder, S. (2019) 'Neural transfer learning for natural language processing', PhD thesis, National University of Ireland.

40. Sun, C., Qiu, X., Xu, Y. and Huang, X. (2019) 'How to fine-tune BERT for text classification?', Proceedings of CCL, pp. 194–206.

41. Vidgen, B. and Derczynski, L. (2020) 'Directions in abusive language training data: Garbage in, garbage out', PLOS ONE, 15(12), e0243300.

42. Vidgen, B. et al. (2021) 'Learning from the worst: Dynamically generated datasets to improve online hate detection', Proceedings of ACL, pp. 1667–1682.

43. Waseem, Z. and Hovy, D. (2016) 'Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter', Proceedings of NAACL Student Research Workshop, pp. 88–93.

44. Weidinger, L. et al. (2021) 'Ethical and social risks of harm from language models', arXiv preprint arXiv:2112.04359.

45. Wolf, T. et al. (2020) 'Transformers: State-of-the-art natural language processing', Proceedings of EMNLP: System Demonstrations, pp. 38–45.

46. Zampieri, M. et al. (2019) 'Predicting the type and target of offensive posts in social media', Proceedings of NAACL-HLT, pp. 1415–1420.